Workshop #20

Cognitive Bias Issues in the Forensic Analysis of Pattern and Impression Evidence and in Medicolegal Evaluations

AAFS 67th Annual Scientific Meeting
Orlando, FL
Tuesday, February 17, 2015

Chair: **Andrew Sulner, MSFS, JD**
Forensic Document
Examinations, LLC
New York, NY

Co-Chair: **Barry C. Scheck, JD**
The Innocence Project
New York, NY

Pre-Registration Required — $200 w/registration; $250 workshop only

#20 Cognitive Bias Issues in the Forensic Analysis of Pattern and Impression Evidence and in Medicolegal Evaluations

Tuesday, February 17, 2015 **8:30 a.m. - 5:00 p.m.** **7.0 CE Hours**

Educational Objectives: After attending this presentation, the attendees will be acquainted with the different types of bias that can influence the outcome of forensic investigations. Attendees will learn about classic psychological research studies and real-life case histories demonstrating the effects of bias upon interpretations of pattern and impression evidence and upon medicolegal evaluations and assessments, especially in Shaken Baby Syndrome/Abusive Head Trauma (SBS/AHT) cases. Attendees will also discover how bias can improperly sway the perceptual and cognitive judgments of forensic examiners and produce faulty conclusions, even in the absence of malicious intent.

Impact on the Forensic Science Community: This presentation will impact the forensic science community by clearly demonstrating how various types of bias can adversely impact evaluations of evidence and decision-making in all forensic disciplines. Understanding the sources of bias and learning how to limit or minimize their influence is essential for improving the reliability and accuracy of decisions made by forensic experts and avoiding miscarriages of criminal and civil justice. All forensic scientists and laboratory directors must be keenly aware of the potential for bias and the types of internal procedures and protocols that can and should be implemented to minimize the impact of bias in forensic investigations and casework.

Chair:
Andrew Sulner, MSFS, JD
Forensic Document Examinations, LLC
New York, NY

Co-Chair:
Barry C. Scheck, JD
The Innocence Project
New York, NY

Faculty:
Keith A. Findley, JD
University of Wisconsin Law School
Madison, WI

Lucy B. Rorke-Adams, MD
The Children's Hospital of Philadelphia
Department of Pathology
Philadelphia, PA

Saul Kassin, PhD
John Jay College of Criminal Justice
New York, NY

Donald E. Shelton, JD, PhD
Saline, MI

Glenn M. Langenburg, PhD
Saint Paul, MN

Dan S. Simon, LLB, MBA, SJD
University of Southern California
Gould School of Law
Department of Psychology
Los Angeles, CA

Daniel A. Martell, PhD
Park Dietz & Associates
Newport Beach, CA

William C. Thompson, PhD, JD
University of California
Dept of Criminology
Law & Society
Irvine, CA

Daniel C. Murrie, PhD
Charlottesville, VA

Deborah Tuerkheimer, JD
Northwestern University School of Law
Chicago, IL

Michael Risinger, JD
One Newark Center
Newark, NJ

Pre-Registration Required — $200 w/registration; $250 workshop only

#20 Cognitive Bias Issues in the Forensic Analysis of Pattern and Impression Evidence and in Medicolegal Evaluations (continued)

Program Description: A multidisciplinary faculty of distinguished psychologists, lawyers, forensic scientists, and others will provide attendees with a clear picture and concrete examples of how and why bias affects the outcome of forensic investigations. Attendees will learn about the various experimental research studies that reveal the susceptibility of investigations to the prospect of psychological error due to cognitive and motivational factors, thereby increasing the risk of miscarriages of criminal and civil justice. Attendees will learn about practices that should be avoided and followed in order to minimize potential biasing influences. Examples from actual forensic casework in both criminal and civil cases will be used to illustrate the impact of bias on the outcome of forensic examinations and the manner in which such opinions are reported or expressed in court. Attendees will also learn about how proffered expert opinion evidence tainted by bias can be challenged or impeached at trial and how trial judges may rule on the admissibility of such evidence in the face of challenges predicated on examiner (cognitive) bias.

Program:

8:30 a.m. - 8:35 a.m.	Introductory Remarks by AAFS President Daniel A. Martell, PhD *Daniel A. Martell, PhD*
8:35 a.m. - 9:05 a.m.	Bias Control: The National Commission on Forensic Science, The National Institute of Standards and Technology (NIST) and the Draft Guidance on Cognitive Bias Effects From the Forensic Science Regulator for England and Wales *Michael Risinger, JD*
9:05 a.m. - 9:40 a.m.	Bias Effects in Forensic Handwriting Investigations and Expert Testimony: An Insider's View *Andrew Sulner, MSFS, JD*
9:40 a.m. - 10:15 a.m.	Bias in Forensic Science Evidence: A Judicial Perspective *Donald E. Shelton, JD, PhD*
10:15 a.m. - 10:30 a.m.	Break
10:30 a.m. - 11:15 a.m.	Cognitive and Motivational Causes of Investigative Error *Dan S. Simon, LLB, MBA, SJD*
11:15 a.m. - 12:00 p.m.	The Forensic Confirmation Bias: Problems in Human Nature and Solutions *Saul Kassin, PhD*
12:00 p.m. - 1:00 p.m.	Lunch
1:00 p.m. - 1:30 p.m.	Recent Research Addressing Cognitive Bias in Forensic Evaluations and Psychological Assessments *Daniel C. Murrie, PhD*
1:30 p.m. - 2:00 p.m.	Bias Effects in Forensic Science: A Perspective From a Caseworking Forensic Scientist Who Uses Sequential Unmasking Techniques *Glenn M. Langenburg, PhD*

Pre-Registration Required — $200 w/registration; $250 workshop only

#20 Cognitive Bias Issues in the Forensic Analysis of Pattern and Impression Evidence and in Medicolegal Evaluations (continued)

Program cont.

2:00 p.m.	-	2:45 p.m.	Contextual Bias and Domain-Relevance: Lessons From Weapons of Mass Destruction (WMD) Forensics *William C. Thompson, PhD, JD*
2:45 p.m.	-	3:00 p.m.	Break
3:00 p.m.	-	3:30 p.m.	Cognitive Bias Issues in Evaluating Shaken Baby Syndrome/Abusive Head Trauma (SBS/AHT) *Deborah Tuerkneimer, JD*
3:30 p.m.	-	4:00 p.m.	The Bias of the Gold Standard *Lucy B. Rorke-Adams, MD*
4:00 p.m.	-	4:30 p.m.	A Critical Look at Cognitive Bias Issues in Expert Testimony About Non-Accidental Head Injury *Keith A. Findley, JD*
4:30 p.m.	-	5:00 p.m.	Cognitive Bias Issues in Shaken Baby Syndrome/Abusive Head Trauma (SBS/AHT) Cases: A Litigation Perspective *Barry C. Scheck, JD*

Targeted Audience: All Disciplines

Knowledge Level Required: Basic (little or no knowledge of subject presented)

Expected Handout Length: 600 Pages

DAN MARTELL

Dan Martell is the President of the American Academy of Forensic Sciences. He is a forensic neuropsychologist with over 25 years of experience in both criminal and civil litigation, and has consulted on hundreds of forensic cases in over 30 states and internationally, specializing in issues of mental disorder, brain damage, neuroscience, and criminal behavior. He serves on the faculty of the Semel Institute for Neuroscience and Human Behavior at the David Geffen School of Medicine at UCLA, and has authored over 100 scientific journal articles and research presentations at national conferences.

Dr. Martell is Board-Certified in Forensic Psychology by the American Board of Professional Psychology; a Fellow of the American Academy of Forensic Psychology; a Fellow of the National Academy of Neuropsychology; and a Fellow of the American Academy of Forensic Sciences. He is also a consultant for the United Nations International Criminal Court in The Hague.

Dr. Martell graduated cum laude with a degree in psychology from Washington and Jefferson College and earned his Master's and Ph.D. degrees in Clinical Psychology at the University of Virginia. During his graduate training at Virginia, he studied at the School of Law and the Institute of Law, Psychiatry, and Public Policy, developing his interests in clinical psychology and the Law. He completed his clinical internship and a post-doctoral fellowship in forensic psychology at Bellevue Hospital and New York University Medical Center in New York City, where he specialized in forensic neuropsychology.

D. MICHAEL RISINGER

Michael Risinger is the John J. Gibbons Professor of Law at Seton Hall University School of Law. He holds a B.A. from Yale University, and a J.D. from Harvard Law School. He is a past chair of the Association of American Law Schools Section on Civil Procedure, the past chair of the AALS Section on Evidence, a life member of the American Law Institute, and was for 25 years a member of the New Jersey Supreme Court Committee on Evidence, which was responsible for the current version of the New Jersey Rules of Evidence. He is the author of two chapters in Faigman, Kaye, Cheng and Mnookin, MODERN SCIENTIFIC EVIDENCE ("Handwriting Identification" and "A Proposed Taxonomy of Expertise"). He is also the author of articles on a diverse range of subjects, including many articles on expert evidence issues. He is in addition Associate Director of the Last Resort Exoneration Project at Seton Hall, which is devoted to freeing the convicted innocent of New Jersey.

Professor Risinger was recently appointed a member of the Human Factors Subcommittee of the National Commission on Forensic Science (NCFS), which was created by the Department of Justice to make recommendations to the Attorney General that will serve to enhance forensic science in the United States.

Forensic Science Regulator

Overseeing Quality

Draft guidance: Cognitive bias effects relevant to forensic science examinations

August 2014

This is a consultation draft and therefore should not be regarded or used as a standard. This draft is issued to allow comments from interested parties; all comments will be given consideration prior to publication. Comments should be sent to FSRConsultation2@homeoffice.gsi.gov.uk and should be submitted by 31st October 2014. This mailbox is not for general correspondence and is not routinely monitored so no acknowledgement will normally be sent.

THIS DRAFT IS NOT CURRENT BEYOND 31st OCTOBER 2014.

CONTENTS

1. INTRODUCTION

A key requirement of the Forensic Science Regulator's Codes of Practice and Conduct for forensic science providers and practitioners (the Codes) is that they "Act with honesty, integrity, objectivity and impartiality..." (p9 bullet point 2).

However many fields of forensic science include subjective assessment and comparison stages that are potentially susceptible to unconscious personal bias (cognitive contamination), which in turn could undermine the objectivity and impartiality of the forensic process. The focus of this appendix to the Codes is on providing general guidance on cognitive bias relevant to forensic examinations with the aim of alerting readers on how to recognise it and therefore help safeguard against biasing effects, through adherence to good practice. This document also provides examples of good practice for specific subject areas listed in sections 7 to 12. This document sets out the policy to ensure the format and content of all annexes issued by the Regulator are consistent.

2. EFFECTIVE DATE

This is a draft issue of this document for consultation.

3. SCOPE

These guidelines are limited to the consideration of cognitive bias within processes associated with forensic science examinations at scenes and within the laboratory only and therefore do not cover the wider aspects of the criminal justice system (CJS) such as court processes including activities of the judiciary/legal profession.

4. MODIFICATIONS

This is a draft issue of this document.

5. TERMS AND DEFINITIONS

Anchoring or focalism: The tendency to rely too heavily on one piece of information when making decisions.

Blinding: Shielding the forensic examiner from information about the case that is not required in order to conduct the examination.

Cognitive bias: a pattern of deviation in judgement whereby inferences about other people and situations may be drawn in an illogical fashion.

Confirmation bias: The tendency to test hypotheses by looking for confirming evidence rather than potentially conflicting evidence.

Contextual bias: The tendency for a consideration to be influenced by background information.

Debias: The reduction or elimination of the impact of bias in decision making and problem solving.

Expectation bias: also known as experimenter's bias, is where the expectation of what you will find affects what you do actually find.

Photogrammetry: The art science and technology of obtaining reliable information about physical objects through the processes of recording measuring and interpreting photographic images.

Psychological contamination: Exposure to other information which is irrelevant to their assessment but introduces unconscious bias into their findings.

Reconstructive effects: The tendency when people rely on memory, to fill in gaps on recall with what they believe should have happened.

Role effects: The tendency for individuals to identify themselves as part of a team with common goals which may introduce subconscious bias.

6. AN EXPLANATION AND BRIEF OVERVIEW OF COGNITIVE BIAS

6.1 Overview

Cognition is the mental process of knowing, including awareness, perception, reasoning and judgement[1], and is distinct from emotion and volition[2]. Cognitive bias may be defined as a pattern of deviation in judgement whereby inferences about other people and situations may be drawn in an illogical fashion[3]. We all tend to display bias in judgements that we make in everyday life, indeed this is a natural element of the human psyche: Jumping to a conclusion, tunnel vision, only seeing what we want to see, being influenced by the views of others, are all behaviours we recognise in ourselves and others. However whilst such biases may be commonplace and part of human nature, it is essential to guard against these in forensic science, where many processes require subjective evaluations and interpretations. The consequences of cognitive bias may be far-reaching: decisions by the investigator to follow a particular line of enquiry, the CPS to prosecute or not, and decisions in the CJS as to guilt or innocence of an individual upon which may rest their liberty or even their life in some jurisdictions, frequently depends on the reliability of the forensic evidence and the conclusions drawn from its interpretation.

Cognitive bias has been identified as a potential issue within criminal justice systems since the 1970s[4,5,6], and in more recent years some high profile cases

---

[1] The American Heritage® Science Dictionary Copyright © 2005

[2] The Concise Oxford Dictionary, 18th edition

[3] Haselton, M. G., Nettle, D., & Andrews, P. W. (2005). *The evolution of cognitive bias.* In D. M. Buss (Ed.), The Handbook of Evolutionary Psychology: Hoboken, NJ, US: John Wiley & Sons Inc. pp. 724–746

[4] Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124–1131. http://dx.doi.org/10.1126/science.185.4157.1124

[5] Charlton, D., Fraser-Mackenzie, P.A.F. & Dror I.E. (2010). Emotional experiences and motivating factors associated with fingerprint analysis. Journal of Forensic Sciences, 55, p385-393

including false positive fingerprint identifications[7,8] have brought the issue into sharp relief. This has been reinforced by an assessment of forensic science published in 2009 by the US National Academy of Sciences in which a diverse range of forensic disciplines within the USA were identified to have wide-ranging issues including lack of validation, standardisation, reliability, accuracy and potential for bias[9].

6.2 Categories of cognitive bias

There are a number of categories of cognitive bias, including those described briefly below; some are very similar and can sometimes apply in combination in real life situations. Further information on different sources of bias in forensic science is provided in a paper by Dror[10].

Expectation bias, also known as experimenter's bias, is where the expectation of what you will find affects what you do actually find i.e. where there is scope for ambiguity, people only see what they expect to see. For example, an experimenter may disbelieve or downgrade the significance of findings that conflict with their original expectations, whilst believing and certifying material that supports preexisting expectations. This is also closely related to observer expectancy effects in which a researcher unconsciously manipulates an experiment or data interpretation in order to find a result consistent with expectations.

Confirmation bias is closely related to expectation bias, whereby people test hypotheses by looking for confirming evidence rather than potentially conflicting evidence[11,12]. For example, in the evaluation of DNA mixtures, if the reference sample is compared before the crime profile has been interpreted, confirmation bias would result if the analyst then looked only for features supporting the inclusion of the reference profile within the mixture. Some verification processes have potential for confirmation bias if the verifier has knowledge of the original examiner's findings before reaching their own conclusions. They may also be influenced by the experience or status of the previous examiner where these are known to them (so-called conformity effects, and institutional bias).

---

[6] Dror, I.E., Peron, A.E., Hind, S.-L. & Charlton, D. (2005), When emotions get the better of us: The effect of contextual top-down processing on matching fingerprints. Applied Cognitive Psychology, 19, p799-809.

[7] Office of the Inspector General (2006). A review of the FBI's handling of the Brandon Mayfield case. Office of the Inspector General, Oversight & Review Division, US Department of Justice.

[8] Campbell, A. (2011). The fingerprint inquiry report. Available at: http://www.thefingerprintinquiryscotland.org.uk/inquiry/3127-2.html

[9] NAS. (2009). Strengthening forensic science in the United States: A path forward. Washington, DC: National Academy of Sciences, National Academies Press.

[10] Dror, I.E. (2009) How can Francis Bacon help forensic science? The four idols of human biases. Jurimetrics, 50, p93-110

[11] Balcetis, E., Dunning, D. (2006) See What You Want to See: Motivational Influences on Visual Perception, Journal of Personality & Social Psychology, Vol.91, No.4, p612-625

[12] Sanitioso, R., Kunda, Z., Fong, G.T., 1990. Motivated Recruitment of Autobiographical Memories, Journal of Personality & Social Psychology, 59 p229-241

Examples such as a request to "Quickly check this match" demonstrate the potential for confirmation bias in verification processes.

Anchoring effects or focalism is closely related to both the above and occurs when an individual relies too heavily on an initial piece of information when making subsequent judgements, which are then interpreted based around the anchor. For example investigators may fix too readily on a specific subject early on in an investigation and look to explain the circumstances around that person, whilst subsequently ignoring simpler alternative explanations of what may have happened, or who else may have committed the crime.

Contextual bias is where someone has other information aside from that being considered which influences (either consciously or unconsciously) the outcome of the consideration. Psychological research has demonstrated that perception is responsive to both the individual's psychological and cognitive state along with the environment in which they are operating. For example, a scientist working within a police laboratory could be influenced by knowing that detectives believe they have a strong suspect, or that the suspect has already confessed to having committed the crime. Provision of information not required by the scientist to undertake their evaluation and that potentially influences this type of biasing has been termed 'psychological contamination' or 'cognitive contamination'[13], as opposed to the more widely understood issue within forensic science of 'physical contamination'[14].

Role effects are where scientists identify themselves within adversarial judicial systems as part of either the prosecution or defence teams, and this may introduce subconscious bias which can influence decisions especially where some ambiguity exists. In fibre examinations when potential contact between two textile items is under consideration but no matching fibres are found, cognitive bias may be seen from a scientist acting on behalf of the prosecution, and interpreting the findings as neutral rather than considering whether the absence of matching fibres might support the view that the contact had not occurred. Role effects are differentiated from a similar effect called motivational bias, which is often considered separately to cognitive biases. Motivational bias occurs where, for example, motivational influence on decision making results in information consistent with a favoured conclusion tending to be subject to a lower level of scrutiny than information which may support a less favoured outcome[15,16]. An extreme example of this is where an individual wants one side

---

[13] Dror, I.E. (2013) Practical solutions to cognitive and human factor challenges in forensic science. Forensic Science Policy & management 4 p1-9.

[14] Kassin, S.M. et al (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. Journal of Applied Research in Memory and Cognition. 2, p42-52

[15] Pyszczynski, T., Greenberg, J., 1987 Toward an Integration of Cognitive & Motivational Perspectives on Social Inference: A biased Hypothesis-testing Model, Advances in Experimental Social Psychology, vol 20 p297-340.

[16] Dawson, E., Gilovich, T., Regan, D. T., 2002 Motivated Reasoning and Performance on the Wason Selection Task, *Personality & Social Psychology Bulletin*, 28 p1379-1387

to win and when in doubt will always make a conscious decision in one direction i.e. to routinely inculpate (or conversely exculpate) suspects; examples of such misconduct have been well documented[17].

Reconstructive effects[18] can occur when people rely on memory rather than taking contemporaneous notes: people tend to subsequently fill in gaps with what they believe should have happened and so may be influenced by protocol requirements when recalling events some time later from memory.

6.3 Academic research into cognitive bias in forensic science

Academic research into cognitive bias in forensic science, conducted through both experimentation and identification of examples from past cases, has indicated that effectively any technique or process which includes subjective assessment and comparison is potentially susceptible to bias. A particularly useful overview of this topic has been published recently by Kassin et al[19]. Other research papers have describe studies on bias in DNA mixture interpretation[20], fingerprint comparison[21,22], handwriting comparison[23], fire investigation[24], forensic odontology[25], bullet comparisons[26], hair comparison[27], and forensic anthropology[28]. The extent of the issue in real life has yet to be fully evaluated, however it is likely to be highly variable depending on the type of forensic analysis being conducted and the extent of safeguards built into the

---

[17] Giannelli P.C. (2010) Independent crime laboratories: the problem of motivational and cognitive bias: Utah Law Review 2, p247-256

[18] Risinger, D.M. et al (2002) The Daubert/Kumho Implications of Observer Effects in Forensic Science: Hidden Problems of Expectation and Suggestion Author(s): California Law Review, Vol. 90, No. 1, pp. 1-56

[19] Kassin, S.M. et al (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. Journal of Applied Research in Memory and Cognition. 2, p42-52

[20] Dror, I. & Hampikian, G. (2011). Subjectivity and bias in forensic DNA mixture interpretation. Sci. Justice 51 p204-208

[21] Dror, I. et al (2006 check) Contextual Information Renders Experts Vulnerable to Making Erroneous Identifications: Forensic Science International 156 74-78

[22] Dror, I.E & Charlton, D. (2006) Why experts make errors, J. Forensic Identif. 56 600–616.

[23] Found, B. & Ganas, F. (2013) The management of domain irrelevant context information in forensic handwriting examination casework, Sci. Justice 53 p154–158.

[24] Bieber, P. (2012) Measuring the impact of cognitive bias in fire investigation. International symposium on fire investigation. Sci. Technol. (2012) p3–15.

[25] Page, M. et al (2012), Context effects and observer bias—implications for forensic odontology, J. Forensic Sci. 57 p108–112.

[26] Kerstholt, J., Eikelboom, A., Dijkman, T., Stoel, R., Hermsen, R., van Leuven, B., Does suggestive information cause a confirmation bias in bullet comparisons? (2010) *Forensic Science International* **198** 138–142

[27] Miller, L. (1987) Procedural Bias in Forensic Science Examinations of Human Hair, Law and Human Behaviour 11(2) p157-163

[28] S. Nakhaeizadeh, et al., Cognitive bias in forensic anthropology: Visual assessment of skeletal remains is susceptible to confirmation bias, Sci. Justice (2013), http://dx.doi.org/10.1016/j.scijus.2013.11.003

processes within which organisations or individuals are working. From a global perspective, it will also depend on the overarching quality requirements and expectations of the particular justice system within which the outcomes are delivered.

6.4 Bias countermeasures (also known as "Debiasing techniques")

Blinding precautions

Providing the forensic examiner only with information about the case that is required in order to conduct an effective examination is the most powerful means of safeguarding against the introduction of contextual bias. Such information could be for example a statement from the victim, and for this reason direct contact with the investigating officer should be avoided prior to assessment. That said, it should be borne in mind that the information required may vary from case to case, and it is hard to perform case assessment and interpretation effectively without having access to background information. For example, targeting effectively for "touch" DNA may require information from witness statements.

Most forensic science providers would be able to control the flow of information to analysts, however some forensic science practitioners are in sole practice and the instructing agency needs to have role and therefore a working knowledge. In such situations, the practitioner may need to ensure the officer in the case is well aware of appropriate information, images and disclosure through the investigation.

Good practice in forensic science requires that independent checking of critical findings is undertaken (Codes 15.3.2). Independent checking that minimizes the risk of cognitive bias would entail assessment without knowing the outcome of the initial analysis, or even where practicable the identity of the original examiner in order to avoid confirmation bias.

Structured approach

Application of a structured approach to performing a comparison and arriving at a decision using an essentially "linear" process can effectively reduce or eliminate the influence of the target (i.e. information pertaining to suspect) from the conclusions drawn. A good example of a general methodology for undertaking comparisons is "Analysis, Comparison Evaluation and Verification" (ACE-V). It is the most commonly accepted approach to fingerprint comparison in the UK and USA. The sequence of working is: i) an examiner analyses a mark: ii) the examiner then compares the mark to a known print: iii) having compared the images, the examiner evaluates what they have seen and reaches a decision iv) the results are then subject to verification by one additional examiner or more. Although most literature sets out the ACE-V process as a sequential process it is in fact not linear in application to fingerprint comparisons – the Analysis phase can be revisited in a well-structured way

during the comparison phase. However the evaluation is a separate stage as described.

Another framework that has been applied to give structure to the evaluation of scientific findings is the Case Assessment and Interpretation (CAI) model[29,30]: this helps scientists design effective, efficient, and robust case-examination strategies. The CAI model is founded on Bayesian[31] thinking and provides clarity on the role of forensic scientists within the criminal justice process. It also encourages consistency of approach, and helps direct research effort. In common with ACE-V it describes an approach in which examination and analysis of scene-related material is undertaken prior to assessment. However whilst ACE-V often entails some re-iteration of the assessment process, CAI is essentially a linear approach and both provide a practical means of safeguarding against confirmation bias. Further information on the CAI-type approach is given in section 7.

Method development

As the potential for cognitive bias arises at different stages in the examination process, method development ought to look at risks or perceived risks in the method and apply the most practicable control strategy. It ought to be borne in mind that simply because there is a risk of an event, it doesn't mean it automatically manifests itself affecting critical judgment.

Having a complete picture is often vital for constructing and testing relevant hypothesis and propositions. However if knowing about certain aspects are assessed to work against the objective process in a particular method (i.e. assessment recommends a blinding method is used), then the methodology right down to design and content of paperwork as well as interaction with the officers in the case might be considered. If the whole case file is handed over to an analyst with all the extraneous detail, then even if there is no perceptible bias there is the perception that it could have occurred and may be open to challenge in court.

Awareness, training and competence assessment

It is not sufficient to simply have well defined evaluation procedures in place as outlined above: practitioners need to be aware of the risks and issues arising from cognitive bias, and to receive substantial training in how to overcome these in their respective roles. Similarly those involved in method development require training regarding the risks and issues so that they are best equipped to design out cognitive bias from processes as far as is practicable.

---

[29] Cook, R. et al (1998a) A model for case assessment and interpretation. Science and Justice 38: 151-156.

[30] Association of Forensic Science Providers. (2009). Standards for the formulation of evaluative forensic science expert opinion. Sci. Justice 49, 161–164.10.1016/j.scijus.2009.11.004

[31] The use or application of Bayes' Theorem, a mathematical formula that can be applied to update probabilities of issues in the light of new evidence.

Given that susceptibility to psychological and cognitive influences varies between individuals, there may be merit in assessing these susceptibilities as part of the recruitment or selection procedures for new staff, such as the recruitment testing procedure for fingerprint examiners developed by Dror et al[32]. Competence in applying evaluative processes should be formally assessed prior to commencing casework and thereafter on a regular basis. This may be achieved through a proficiency testing programme, utilizing mocked up casework samples for which the expected outcomes of testing and evaluation are known. Whilst blind trials are effectively the gold standard in providing the most reliable indicator of real-life performance, in reality they can be very time-consuming and challenging to set up, especially in avoiding alerting the person being assessed that it is a trial rather than another piece of casework. Good practice adopted by many laboratories is to undertake a mixed programme of both declared and undeclared trials, with the proficiency of all individuals tested on a regular basis.

Avoidance of reconstructive effects

The taking of contemporaneous notes or technical records is another stipulation in the Codes (section 15.2.3) Adherence with this requirement wherever it is practicable to do so at and at all stages in the collection and processing of forensic evidence provides the best safeguard against potential reconstructive effects.

Avoidance of role effects

Role effects whereby scientists are subconsciously influenced by acting on behalf of the defence or prosecution are difficult to demonstrably eliminate given the adversarial nature of the CJS within the UK, and which are potentially compounded by the pressures of a commercial market in which a supplier/customer relationship for the delivery of forensic science is the norm. These pressures apply whether an FSP is providing contracted services to the prosecuting side or to the defence, or in the case of police laboratories in providing services to an internal customer.

However a wider customer is being served here i.e. the CJS, not just the defence or prosecution sides paying for the services: the Regulator's Codes of Conduct for forensic science stipulate that practitioners shall:

a. Have an overriding duty to the court and to the administration of justice, and,
b. Act with honesty integrity and impartiality.

This is reinforced in section 7.2 of the Regulator's Codes of practice, in which conflicts of interest, perceived or otherwise, and threats to impartiality of a practitioner are identified, including the following:

a. Being the sole reviewer of their critical findings.

---

[32] Charlton, D., Fraser-Mackenzie, P.A.F. & Dror I.E. (2010). Emotional experiences and motivating factors associated with fingerprint analysis. Journal of Forensic Sciences, 55, p385-393

b. Being over-familiar with or trusting another person instead of relying on objective evidence.
c. Having organisational and management structures that could be perceived to reward, encourage or support bias, where for example a culture of performance measurement and time pressures could potentially pressurize examiners into biasing decisions.

Whilst point c) may be erring towards misconduct rather than being a cognitive phenomenon, the overriding issue with all these points is the effect of subconscious influences on impartiality. Furthermore, compliance with the ISO 17025 quality standard which is an integral requirement of the Codes stipulates that personnel undertaking the analyses shall be free from any undue commercial, financial and other pressures which might influence their technical judgement. In other words, organisational systems and safeguards are required to ensure scientists are insulated from potential biasing pressures.

The Criminal Procedure Rules state in part 33.2 that (1) An expert must help the court to achieve the overriding objective by giving objective, unbiased opinion on matters within his expertise; (2) This duty overrides any obligation to the person from whom he receives instructions or by whom he is paid; (3) This duty includes an obligation to inform all parties and the court if the expert's opinion changes from that contained in a report served as evidence or given in a statement. Every expert report must contain a statement that the expert understands his duty to the court, and has complied and will continue to comply with that duty.

Adoption of a structured approach such as the CAI principles as described in 4.3.1.2 and expanded further in section 6 below, in which consideration of both prosecution and defence hypotheses, can help ensure evidence is evaluated and presented in a more balanced manner, regardless of defence or prosecution role. This requires that:

a. Experience is brought to bear by a person who has all the information regarding the case in formulating a coherent strategy that underpins the rationale for analytical submissions;
b. Analysis is undertaken only with relevant facts disclosed to the analyst; and,
c. The results of the analysis are reviewed and interpreted from the perspective of the whole case, and should accept the conclusions drawn by the analyst.

7. A GENERIC PROCESS TO MANAGE COGNITIVE BIAS FOR A RANGE OF FORENSIC EVIDENCE TYPES

7.1 The role of the investigating officer or instructing authority

Appropriate flow of information is very important in all cases, one limiting factor in the assistance forensic science can give to the investigation is pertinent information not being passed on. Contextual or case information can be made available for the leading examiner for case building purpose, the lead can then ensure analysts receive information appropriate for that stage, while still

ensuring proper case assessment can be made and the most appropriate techniques are used.

However, when instructing experts in sole practice, a greater onus is placed on the investigating officer (or instructing authority) to manage the flow of information. The expert is still likely to need the contextual or case information, but this may be required to be held back until certain analytical stages are complete.

However, anybody instructing experts should always think hard about including comments such as the 'suspect admitted to the crime', 'we already have a DNA match', or even in the question asked '...can you identify whether suspect A (the stabber) is carrying anything and, if he is, what that item is...' Being exposed to such information doesn't automatically result in a biased decision, but it can influence and should be guarded against.[33]

The investigating officers or instructing authority should deals with the following in their forensic strategy:

a. information flow based upon the nature of the evidence type, the phase of the analysis and the capability of the forensic science provider.
    i. Is the provider able to apply any debiasing techniques themselves i.e. a larger provider will probably control the flow of information to the analyst?
    ii. Is this a smaller provider or niche specialism where the lead examiner is the sole examiner? If this is the case then agree with them beforehand how the initial, and sometimes follow up, communications might be best handled.

7.2 The role of the scientist in the analysis or initial evaluation stage

The analyst should know through their training that they must stay separate from the rest of the investigation and accept the fact that they should undertake the analysis "blind", and not to seek other information beyond what is required, in order to protect their impartiality. If potentially biasing information is inadvertently disclosed to them, for example that someone is in custody or has confessed, the lead scientist should be informed that this has happened.

7.3 The role of a forensic expert

The role of the forensic science expert is to evaluate scientific findings and the results of analytical tests in the context of the relevant case circumstances. An expert opinion should meet the criteria that it is balanced, robust, logical and transparent[34]:

---

[33] In R v Rogers [2013] EWCA Crim 2406 the Court of Appeal (Criminal Division) rejected the argument the admission of a police officer's identification of the accused from photographs after being informed that there was a DNA match rendered the trial unfair or conviction unsafe.

[34] Cook, R. et al (1998a) A model for case assessment and interpretation. Science and Justice 38: 151-156.

a. Balanced – the expert has considered both the prosecution and defence views in their evaluation
b. Robust – it is based on data that are available for inspection and discussion
c. Logical – in the approach taken to the evaluation
d. Transparent - another suitably qualified scientist could follow all the steps and decisions taken[35].

If all of the above criteria are met, then any difference of opinion between experts could be limited to a well-defined part of the opinion rather than being a general disagreement, as well as identifying the reasons for each of the opinions. This is most helpful to the court in identifying the areas of dispute between scientists.

7.4 Process Outline

A very brief outline of forensic process within the laboratory is as follows:

a. Define requirement
b. Develop examination strategy
c. Agree examination strategy with client
d. Carry out forensic examinations and analyses
e. Review quality and content of examination results
f. Compare the results with the reference samples and marks
g. Evaluate and interpret the scientific findings and analytical tests
h. Verification by second expert
i. Communicate the scientific findings and analytical tests

During this process it is the responsibility of the expert to record, retain and reveal their work. This requires that they:

a. Record all information received
b. Record details of interpretation

Risks of cognitive bias

If it is not practical to mitigate or control the main forms of cognitive bias then the following may occur:

a. An incorrect conclusion may be made.
b. A critical check might be inadvertently administrative or cursory

The evidence may be challenged.

The risks associated with relying on the scientific findings and analytical results as a way of assigning a weight of evidence are that:

It can be difficult to consider alternative hypotheses since knowledge of the actual outcome provides a source of confirmation bias.

---

[35] Association of Forensic Science Providers. (2009). Standards for the formulation of evaluative forensic science expert opinion. Sci. Justice 49, 161–164.10.1016/j.scijus.2009.11.004

The limitations of the examination and tests performed can be overlooked when evaluating the findings.

Risk management in all disciplines usually starts with an assessment, and a process map detailing the critical control points as required in the Codes (19.4.2.) for building in contamination controls during method is development may be useful for this purpose. This practice should identify the stages where individuals being knowledge rich is not ideal and stages where being knowledge poor is damaging. This approach can inform the examination strategy as well as communication strategy. As the officer in the case may have a role, such a visual tool might be included in officer awareness training or supplied as service information.

7.5 Mitigation strategies to reduce the risk of cognitive bias:

The expert goes through a formal process of pre-assessing the expected probabilities for an exhaustive range of possible outcomes, in as many or as few categories as is sensible for the examination, recording their opinions.

Each category in the exhaustive list of outcomes is considered firstly under the assumption that the prosecution hypothesis is true, and secondly under the assumption that the defence hypothesis is true.

These are used to provide an expected outcome which may be either qualitative or quantitative with the latter expressed as a Likelihood Ratio (LR).

The background data and experience used for assessing the expected outcomes are documented and any gaps identified.

A second expert carries out the same process independently, without viewing the decisions made by first examiner and the experts jointly agree the expected outcomes.

Posterior probabilities are not provided for evaluation of findings[36].

7.6 Recommended good practice

Define requirement[37]:

a. Identify whether the scientist's role in the case is investigative (intelligence) or evaluative (judicial).
b. Seek clarity on which tests are required, the purpose and how this fits into the hierarchy of sub-source (e.g. touch DNA), source, activity and offence level propositions[38,39].

---

[36] The posterior probability is the conditional probability assigned after the scientific evidence has been taken into account; so considers the probability of the hypothesis *given* the evidence. This is an example of the prosecutors fallacy or transposed conditional. The scientist should provide the probability of the evidence *given* the hypothesis.

[37] Cook, R. et al (1998b). A hierarchy of propositions: Deciding which level to address in casework. Science and Justice 38:231-239.

Develop examination strategy:

a. Formulate relevant prosecution and defence alternatives based on the case circumstances and information provided.
b. Consider any agreed assumptions that are used in formulating these alternatives.
c. Use assessment of possible outcomes to determine which tests are most informative and discriminating.
d. Use this pre-assessment to assign a weight to an exhaustive list of possible outcomes, giving the expected outcome for each, expressed as a Likelihood Ratio (LR) where these are quantitative.

This approach provides clarity on the alternatives being considered, and the pre-assessment of weight for all outcomes avoids the potential bias of using the observed results to assign weight of evidence.

Carry out forensic examinations and analyses

Review quality and content of examination results: decisions on the suitability of the results and marks for later comparison are made at this stage, to avoid post-comparison rationalisation of opinion on quality.

Compare the results with the reference samples and marks: quality and suitability of the questioned result has already been assessed so this is not influenced by the reference result.

Evaluate and interpret the scientific findings and analytical tests

Verification by second expert: independent review at this stage in advance of communicating the result to the client.

Communicate the scientific findings and analytical tests.

Interpret the scientific findings and analytical tests:

a. Confirmation bias is mitigated by using the LR or qualitative expectation which has already been assigned to each outcome, before the examinations and tests have been performed.
b. Pre-assessment enables the scientist to explain how the weight of evidence has been assigned.
c. Provide details of the assumptions that have been made.
d. Give the basis of the expert opinion and specify the propositions considered, with reasoning for these, based on the case context.
e. Include any limitation of the opinion.
f. Describe the range of other opinions.

---

[38] Jackson, G. et al (2006) The nature of forensic science opinion--a possible framework to guide thinking and practice in investigations and in court proceedings. Science & justice□ : Journal of the Forensic Science Society 46, 33–44.

[39] RSS Practitioner Guide No 4: Case Assessment and Interpretation of Expert Evidence, Graham Jackson, Colin Aitken, Paul Roberts.

8. GOOD PRACTICE GUIDELINES - SCENES OF CRIME

The police response to a reported crime requires many factors to be taken into consideration and for priorities to be balanced accordingly. Preserving the scene, securing evidence, speed of response including making most effective use of the "Golden Hour", proportionate use of resources based on the seriousness of the crime: all are potentially conflicting in their requirements, and all are overridden by the most pressing priority of all, the preservation of life.

Within this context and from the outset of the investigation, the investigative team seeks to answer many questions that will assist in making sense of the incident under investigation. Frequently the answers to these questions can be provided by material which is obvious and readily to hand, but there will also be gaps. The latter may be filled by gathering of further information or material, identified during the course of the investigative decision-making process, and which may be present at the scene of crime, at other related sites or from other sources[40].

8.2 Scene of crime process

Serious crime

In major or serious crime investigations, forensic science resources are called upon by the Crime Scene Manager to attend the scene based on the specific needs of a case, especially where other evidence to detect the case is not readily available, and these resources are in proportion to the seriousness of the crime. Prior to entering the secured and controlled scene the examiners (e.g. Crime Scene Examiners, forensic scientists) are briefed regarding the scenario being evaluated and the questions that need to be answered. However, the emphasis here is on ensuring that relevant expertise is deployed with the capacity to look at the case and the inquiry to determine what value may be added and what inferences may be drawn from the collection and analysis of physical evidence[41].

Volume crime

The process for volume crime is markedly different to serious crime, due primarily to significant financial constraints impacting on time, personnel and other resources available. Therefore these processes deployed are about maximizing the benefits from these limited resources as a whole rather than for each crime that is reported. The process constitutes the following steps:

On notification of a crime, the police call handler has to make a decision based on information received, and guided by force policy regarding response to volume crime incidents, on whether or not to dispatch a police officer to attend.

---

[40] National Centre for Policing Excellence (2006) Murder investigation manual

[41] Tilley, N. & Townsley, M. (2009) Forensic science in UK policing: strategies, tactics and effectiveness. Published in Handbook of Forensic Science eds J. Fraser & R. Williams p359-379

If a police officer is dispatched to attend the scene they may collect physical evidence themselves or will determine whether a crime scene examiner is to be called to examine the scene for any physical evidence.

If an examiner attends the scene, they may be briefed regarding the offence and what might be most usefully looked for, in advance of their searching for and recovering physical evidence from the scene.

Recovered evidence is packaged labelled and transported back to police facilities, after which a decision is made on what if any evidence is subsequently processed[35].

Crime scene activities and risk of bias

Whilst some crime scene studies have been published by criminology specialists[42,43], cognitive bias at scenes of crime has been less comprehensively evaluated than other areas of forensic activity. Nevertheless its potential impact may be significant: for example, it could result in failure to secure the required evidence if a crime scene investigation is closed prematurely resulting in crucial evidence being lost; it could mislead an investigation by investigators focusing too early and incorrectly on a false lead, so that other evidence is potentially overlooked; or if undertaken incorrectly activities could result in "psychological contamination" of evidence downstream in the forensic analysis and interpretation processes.

Both volume and serious crime scene activities may be prone to errors and bias. For volume crime, given the severe time constraints, there is little scope to undertake anything more than a basic examination and recovery of evidence: focus is likely to be concentrated on the aspects of the case which are known from past experience to be most likely to yield fruitful results, e.g. fingerprints and DNA collection at the point of entry in a house burglary or vehicle theft, and on items which may have been handled or discarded at the scene, which the victim may be able to assist in identifying. Conversely, in major crime, context may be more of an issue with a risk that forensic strategies are written with a pre-conceived 'story' in mind.

Opportunities for cognitive bias can be usefully considered within the context of activities related to the crime scene, which can be categorised are as follows, as applied to serious crimes unless otherwise stated and is adapted from a conference presentation[44]:

---

[42] Lingwood, J., Smith, L.L., & Bond, J.W. (in press) 'Amateur vs professional: Does the recovery of forensic evidence differ depending on who assesses the crime scene?' International Journal of Police Science and Management

[43] Adderley, R., Smith, L.L., Bond, J.W., & Smith, M. (2012) 'Physiological measurement of crime scene investigator stress' International Journal of Police Science and Management 14 (2): 166-176.

[44] Fraser, J. (2013) Crime scene examination –final frontier or forgotten function? Paper presented at Forensic Horizons 2013: supporting research and development & delivering best practice for the justice system

Gathering of information prior to scene attendance

Prior to scene attendance information is gathered from any available source regarding the incident to be investigated. This may include witness or victim accounts as to what is alleged to have happened and by their nature these may be consciously or unconsciously biased. With volume crime, decisions on whether or not to attend the scene may be based on this potentially biased information and could therefore affect whether the crime is even investigated at all.

Controlling the forensic process at scenes

This entails creating inner and outer cordons to secure the scene, and establishing a common approach pathway. The cognitive processes entail determining locations and boundaries of the scene and the entry/exit points of the offender, based on observations, information received and inferences. Whilst there may be scope for bias to affect these decisions for example the past experiences of an individual on which they may base their decisions are subjective may not be reflective of typical scenes. However other factors may be more relevant, and have more impact in real life such as convenience: for example establishing the boundary by taping from lamppost to lamppost is commonplace simply because they are already there.

Creating a record of the scene

This includes image capture and writing notes and statements. The cognitive processes include selection of equipment, plus decisions on which images to capture, and entails assessment of the current case needs plus some anticipation of future needs. Depending on Force requirements, these may allow wide variation in how findings are documented and are therefore open to subjectivity. Depending on how the written record is crafted, there is a risk that contextual or confirmation bias may be introduced downstream in the investigative process. A gross example is "item X was recovered from suspect Y, a known repeat offender".

Undertaking forensic examinations at scenes

This requires an understanding of the investigative needs of the case, plus to observe, discover and recover evidence to meet both these present needs and those anticipated for the future. If guidance for these decision-making processes is not explicitly documented then actions taken at this stage are largely reliant on the examiners intuition and tacit knowledge, which in turn are susceptible to bias.

Packaging, storing, labelling and transporting recovered items

These actions are largely procedural rather than cognitive. However there is still scope for introduction of psychological contamination if inappropriate information is included on the labelling of recovered items, as described in section 6.2.1.3.

8.3 Bias Countermeasures and good practice

It is impossible to undertake certain tasks effectively without being provided with context within which to operate, and this is certainly true with scenes of crime

investigations, where some briefing regarding the alleged crime and circumstances are an essential starting point for the examiner's activities. Examiners must be safeguarded against the risks of contextual and other biases through their training and through adherence to formal documented evidence-based guidance. Of necessity such guidance may be more prescriptive in volume crime where scenarios under investigation are relatively consistent scene to scene and are amenable to application of highly directive, standardised and efficient approaches. For example an examiner is better able to make a balanced and informed decision on which parts of a scene to sample for touch DNA analysis if they are armed with knowledge of Force-wide success rates from the substrates available, rather than relying on their own subjective experience of outcomes from just a few of their own cases. However it is also essential that volume crime investigators are trained not to "switch off": given their extensive experience of volume crime scenes, they are better placed than anyone else to identify anything slightly out of the ordinary and therefore potentially indicative of an alternative explanation to that posited by the victim which may be biased or even completely false, e.g. identify evidence that a "burglary" has been staged in order to make a false claim on insurance.

Serious crime investigations of necessity require much more latitude in terms of approach by examiners, although fact-based guidance regarding approaches at their disposal is just as important as in volume crime. Regardless of this latitude of approach it must be demonstrably systematic and it is essential that examiners fully and contemporaneously document information regarding their examination. The latter provides transparency to the process, and is of particular value in:

a. subsequently reviewing the case internally to identify whether issues may have been introduced due to bias, and
b. facilitating review by the defence[45].

Communication of the examiners findings to others through written reports rather than verbal updates, whilst slower, is preferable as the former provides less risk of introducing bias into the transfer of information.

The activities of examiners are guided at the outset by briefing regarding the scenario being evaluated and the questions that need to be answered (6.1.1). Some may be readily answered by material that is easily available but there will also be gaps that cannot be filled[46]. Under these circumstances good practice has been identified of building hypotheses which can help bridge the knowledge gap and indicate where further material may be gathered[47].

The key points when building hypotheses have been identified in this guidance as follows:

---

[45] Butt, L. (2013) The forensic confirmation bias: Problems, perspectives, and proposed solutions – Commentary by a forensic examiner. Journal of Applied Research in Memory and Cognition 2 p59–60

[46] National Centre for Policing Excellence (2006) Murder investigation manual

[47] ACPO (2005) Practice Advice on Core Investigative Doctrine

a. Ensuring a thorough understanding of the relevance and reliability of all material gathered;
b. Ensuring that the investigative and evidential test has been applied to all the material gathered in the investigation;
c. Ensuring there is sufficient knowledge of the subject matter to interpret the material correctly;
d. Defining a clear objective for the hypothesis;
e. Developing hypotheses that 'best fit' with the known material;
f. Consulting colleagues and experts to formulate hypotheses;
g. Ensuring sufficient resources are available to develop or test the hypotheses;
h. Ensuring that hypotheses-building is proportionate to the seriousness of the offence.

This guidance emphasises that these assumptions must be developed objectively and that investigators should be aware of the dangers of making assumptions or believing that assumptions made by others are fact. It further states that where assumptions are used to develop hypotheses this should be made explicit.

In some circumstances where collection and analysis of physical evidence is complex spanning several different evidence types, a co-ordination and integration role is required to be undertaken by experienced forensic practitioners, termed crime scene coordinators, or 'Byford Scientists'. These liaise with senior investigating officers in overseeing the collection of physical evidence and ensuring that the disparate strands of forensic analysis are brought together and appropriate inferences are drawn[48]. This role was introduced after an HMIC inquiry into failings in the Yorkshire Ripper Inquiry[49] due to important leads not being followed up, and false ones being persisted with i.e. classic anchoring effects. It is also important that those undertaking this integration role are also aware of, and thereby safeguard against the fact that these activities are also fraught with potential bias and it may be appropriate under certain circumstances for the coordinators to act as gatekeepers for contextual information and only impart to practitioners information required to fulfill their tasks[50].

---

[48] Tilley, N. & Townsley, M. (2009) Forensic science in UK policing: strategies, tactics and effectiveness. Published in Handbook of Forensic Science eds J. Fraser & R. Williams p359-379

[49] Byford, L. (1982) Report by Sir Lawrence Byford into the police handling of the Yorkshire Ripper case. London: Home Office (Released in June 2006, under the Freedom of Information Act)

[50] Charman, S. (2013) The forensic confirmation bias: A problem of evidence integration, not just evidence evaluation. Journal of Applied Research in Memory and Cognition 2 (2013) 56–58

9. DNA MIXTURES GOOD PRACTICE GUIDANCE

9.1 Outline of the Forensic Process Involving DNA Mixture Interpretation

The generic forensic process that encompasses the interpretation and reporting of DNA profiling results, including complex DNA results, can be briefly described as follows and in figure 1:

a. Items are received along with case information and questions to be addressed by the scientific work.
b. The case information, supplied by the law enforcement customer, is used to direct the DNA recovery and analysis strategy, ideally within a framework of appropriate propositions.
c. If non-complex DNA results are obtained that match a suspect, an appropriate random match probability or Likelihood Ratio (LR) estimate is assigned.
d. If complex mixed DNA results are obtained that can be numerically evaluated the probability of the mixed result is calculated under appropriate prosecution and defence hypotheses and a LR is assigned.
e. If complex DNA results are obtained that do not lend themselves to statistical evaluation, in some circumstances, a qualitative assessment is made and an opinion about the significance of the DNA results can be put forward.
f. Findings are checked by a competent colleague/peer.
g. A statement or report is issued.
h. The scientist may be called to court to give oral testimony.

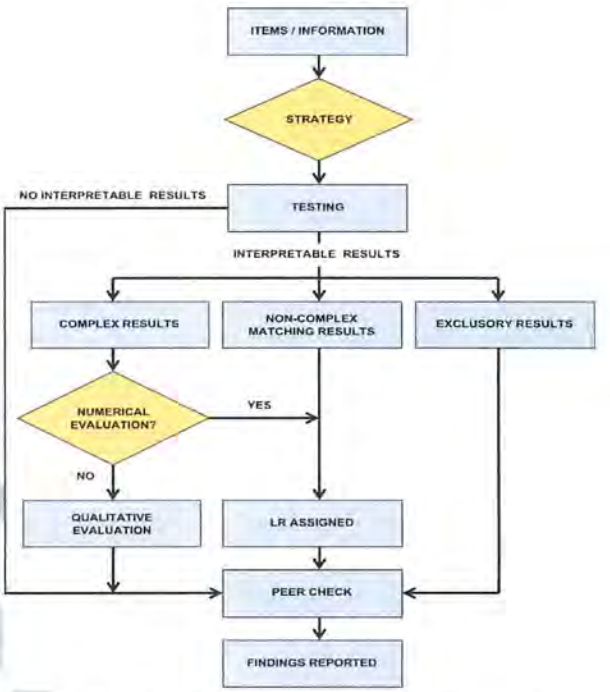


Figure 1: Outline of the Forensic Process Involving DNA Mixture Interpretation

9.2 The Risk of Cognitive Bias in DNA Mixture Interpretation

General Considerations

Just like other areas of science, the interpretation of DNA profiles can potentially be affected by some form of unconscious and unintended bias[51]. This can occur at points in the interpretation process where scientists are free to make decisions or put forward opinions that are formed outside of the mechanical application of a set of rules. Such opinions and decisions can be described as being subjective, since they arise from the individual's mental capabilities, relevant experiences, depth of knowledge and skill as well as any cognitive influences impacting on them at the time both manifest and unapprehended. Usually decisions are made and opinions are formed in the context of the information the scientist has been given about the case.

The interpretation of complex DNA mixtures requires care and skill and often includes a degree of qualitative and subjective decision-making. Indeed, regardless of any case-specific contextual information, practitioners may have a higher expectation of observing DNA profile matches simply because samples were submitted for analysis by police investigators.

General Conditions Impacting on the Level of Cognitive Bias Risk

Within DNA mixture interpretation there is a spectrum of bias risk that is shaped by multiple factors including the following:

a. Risks are low when results are clear and unambiguous and greater when results are complex, of poor quality and there is an increased reliance on subjective opinion.
b. Risks are lower when there is a methodical approach with defined standards built on principles that have been tested and validated, and greater when the approach is un-researched, ad hoc and personal to the operator.
c. Risks are lower when operators and checkers are well trained, experienced and continuously meet acceptable standards of competence; they are greater when operators and checkers are inexperienced, unmonitored and left to adopt their own approach.
d. Risks are lower when interpretation is checked by a competent peer who conducts a separate interpretation fully independent and without influence from the reporting scientist. Risks are higher when checking is less rigorous and/or conducted collaboratively.

[51] Dror, I. & Hampikian, G. (2011). Subjectivity and bias in forensic DNA mixture interpretation. Sci. Justice 51 p204-208.

Risk Source	Low risk	High risk
Result Quality	Results are clear and unambiguous	Results are complex, of poor quality and there is an increased reliance on subjective opinion.
Interpretation Approach	There is a methodical approach with defined standards built on principles that have been tested and validated	The approach is un-researched, ad hoc and personal to the operator.
Operator Competence	Operators are well trained, experienced and continuously meet acceptable standards of competence	Operators are inexperienced, unmonitored and left to adopt their own approach.
Checking	Full independent reinterpretation	Checking is conducted collaboratively

Table 1. Summary of Conditions Impacting on the Risk of Cognitive Bias

Advancing Technology

DNA testing technology continues to develop apace. In addition to the routine application of enhanced sensitivity techniques, today's new multiplexes frequently achieve results from low quantities of DNA (low template samples). The incidence of complex mixtures and of low template profiles exhibiting stochastic effects is increasing and so the conditions in which subjective opinion tends to be relied upon are more commonly encountered. As a consequence, there is an increasing risk of cognitive contamination affecting DNA evidence.

Contemporaneous Case and Reference Sample Interpretation

A substantial part of the risk relating to DNA mixture interpretation arises if the case sample is interpreted alongside the reference sample, or if the case sample interpretation is revised after examination of the reference sample. For example, during the interpretation of a two-person mixture (when the interpretation is not conditioned on the presence of an undisputed DNA source) knowledge of the reference sample may result in confirmation bias in the genotype combinations that are included or excluded as being possible, based on allele quantities.

Use of Qualitative and/or Subjective Approaches

Significant risk is also associated with the use of qualitative and subjective evaluation approaches that have increased considerably since the recent publication of the judgment in R v Dlugosz *et al* (R v Dlugosz, R v Pickering and R v MDS [2013] EWCA Crim 2). The Dlugosz judgment has been taken as a broad license to allow the qualitative evaluation of complex results and subjective expressions of evidential weight when a statistical approach is either difficult or considered inappropriate. Such non-statistical assessments can only

be conducted by comparing a reference sample directly with the complex result from the case sample and drawing conclusions based on the presence of alleles in common between case sample and reference sample, the absence of particular alleles and inferences from allele quantities. The Dlugosz judgment does specify safeguards that relate to whether or not such an evaluation can be considered admissible as evidence and how the evidence should be presented. The safeguards require that the expert is experienced, that the extent of their experience is explained for the consideration of the jury and that caveats relating to the limitations of the findings are clearly explained. Whilst the safeguards might seem reasonable they are dependent on the following underlying assumptions that might be considered dubitable in some circumstances:

a. That general familiarity with complex DNA mixtures and numerical evaluation methods is wholly relevant to the use of what is essentially a new and un-researched evaluative practice; and
b. Such experience enables the practitioner to form safe, reliable opinions relating to sources of DNA within complex mixtures.

To provide assurance in the use of methods that rely on the accuracy of such assumptions, it would assist if clear standards were developed relating to the circumstances in which such an approach is valid and when it is not. Also testing the performance of individual practitioners against developed standards would reduce the risk of inaccurate estimates of evidential strength having an impact in criminal trials. Current application of qualitative methods appears to be largely *ad hoc* without specifically designed controls. If effective quality, training and competency measures are in place, the impacts of cognitive contamination can be minimized.

Potential Oversights in DNA Interpretation Induced By Cognitive Bias

Unconscious cognitive bias has the potential to manifest itself as a skewed evaluation, partly because its influence can increase the likelihood of oversights during the DNA interpretation process. Some possible oversights are described below; most are applicable regardless of whether a numerical or qualitative approach is applied and, with most, the risk is either reduced or eliminated if an assessment is made without knowledge of the reference sample result. Examples include:

a. Restricted assumptions about numbers of contributors.
b. Automatic assumptions that a part of a mixture has originated from one individual.
c. Underestimating the significance of non-matching peaks when they can be considered sub-threshold or designated as artifacts.
d. Underestimating the uncertainty introduced by stochastic effects.
e. Overestimating the significance of unconfirmed matching peaks.
f. Underestimating the significance of unconfirmed non-matching peaks.
g. Taking account of matching alleles where their presence is uncertain due to masking by other components of the mixture.
h. Double counting peaks as homozygous that do not clearly represent a double contribution when the subject is homozygous.

i. Over emphasizing the absence of non-matching alleles when it is not clear if contributors are fully represented.

Further Flaws Potentially Induced by Cognitive Bias

The following points describe some further flaws that may be induced or exacerbated by cognitive bias. Most of these are afforded some latitude by the way in which disclosure tends to be approached by defendants and their representatives. The rules of disclosure within the legal system of England and Wales require no prior disclosure of the defendant's account. This often means that the DNA scientist is required to make their own, uninformed suppositions about appropriate defence hypotheses when deciding on analysis strategy and conducting their evaluation:

a. Greater focus on strategies for DNA recovery and testing that are likely prove a case rather than disprove a case.
b. Choice of propositions that maximize the strength of evidence against the suspect.
c. Observations that support the defence case are less rigorously considered or evaluated and are not given their true weight, particularly relating to the absence of evidence.
d. Failure to express alternative explanations.
e. Reluctance to express doubt particularly during oral evidence at court.

9.3 Case Examples Where Cognitive Bias May Have Contributed to Error

In this section, the identity of specific cases or the practitioners involved are not disclosed; rather, anonymised issues are described in several real cases that may have been caused or exacerbated by unintended cognitive bias. The examples are from cases in which the authors of this guidance had direct experience; all were reported in 2013. They stem from inaccurate evaluations or misleading descriptions of complex DNA mixtures, all biased in favour of the prosecution's case. It is, of course, not possible to be certain to what extent the issues were influenced by cognitive bias or some other source of inaccuracy but they illustrate the difficulties that relate to non-numerical evaluation of complex DNA results. As such, they are helpful in identifying procedural steps and controls that are likely to be effective to both limit cognitive bias and/or demonstrate that it has not occurred.

Qualitative evaluation shown to be at odds with numerical evaluation

A complex mixed DNA result from a case sample contained alleles in common with profiles in all four reference samples that were compared in the case. Most of the alleles in the case sample profile matched Subject X. No statistical analysis was conducted initially but, based on the reporting scientist's experience, s/he gave the opinion the result provided "at least moderate support" for the assertion that some of the DNA on the swabs came from Subject X. The results were later interpreted with the aid of LikeLTD[52], recently

[52] There are several relatively recently developed software programs that are available to providers and are designed to aid the numerical evaluation of some types of complex DNA profiles including complex mixtures.

developed software that is capable of numerical evaluation of some types of complex DNA mixture. The use of this software produced a LR of 4 indicating that, based on commonly accepted verbal descriptors, the strength of support should more fairly have been described as "weak".

Implying the absence of alleles is due to masking by a major component

One case relates to a duplicated, standard sensitivity test on vaginal swabs containing a trace of semen. A full, major component profile was obtained matching the complainant, together with a number of low-level minor component bands that were all present in the defendant's profile. Six duplicated bands in the minor component all matched the defendant and a further five unduplicated bands also matched the defendant. The unduplicated bands were described as unconfirmed. No other, non-matching, minor component bands were visible in either duplicate test and the ratio of the major component to the minor would not have allowed the identification of minor component alleles that were masked by the major component. Comparison of one duplicate result with the other showed that significant stochastic variation, including allelic drop-out, was a reality within these samples. It was not possible to tell whether or not there was full representation of the DNA source(s) within the minor component across the duplicates or to use peak quantities to determine whether there was more than a singular contribution from a specific minor component allele. In the presence of the jury, the scientist was invited to add up the number of alleles in the mixed profile that matched with the suspect's profile. The response was that there were six confirmed bands, five unconfirmed bands, seven that were shared with the major component profile and one further because the suspect was homozygous at one position. The scientist concluded that there were nineteen out of a possible twenty alleles matching the suspect within the mixed profile. There was no attempt to explain that the possible presence of minor component alleles in positions where the minor component would have been invisible was completely neutral to prosecution and defence hypotheses. There was a significant risk that this description of the evidence would be misleading to the jury in favour of the prosecution's case. There may be issues here relating to the approach to quality at the parent laboratory, in particular with the monitoring of competence and/or the support and training provided to reporting officers in the specialist field of low template mixture interpretation. Where there is a lack of understanding of evidence the potential for cognitive contamination is increased.

Ignoring the possibility that a sub-threshold peak is an intrinsic allele

This example relates to a major/minor mixed result from a standard sensitivity test in which a statistical evaluation of eight low level alleles in the minor component was reported. The low level alleles could only have been from the suspect if several of his alleles were not visible due to allelic drop-out. A sub-

---

The following have been used in criminal trials in the UK: *LikeLTD*, developed by David Balding, Professor of Statistical Genetics at University College London. *STRmix*, developed by forensic experts at ESR Ltd in New Zealand (J. Bright and J. Buckleton) and at Forensic Science South Australia (D. Taylor). *TrueAllele®*, developed by Mark Perlin of Cybergenetics in the USA.

threshold peak, distinct from background and with acceptable allelic morphology was present in one of two duplicates and did not match an allele in the suspect's profile. The presence of this peak was presumably considered a spurious occurrence (drop-in or artefact) and was not taken into consideration for the purpose of the statistical evaluation; its presence was not otherwise mentioned in the scientist's report. Although this peak did not satisfy the criteria to be included as a confirmed component of the profile, further testing may have clarified the presence of the peak and if not, a more appropriate statistical approach could have been taken. Failing to take account of the peak or to attempt to replicate it through further work may have been a consequence of cognitive bias.

Assuming all DNA bands in a low level profile are from the same person

This assumption is often made but not always explicitly stated and, based on the quality of the profile and nature of the mixture, there are varying extents to which it can be justified. In low-level profiles it is important for the scientist to consider whether or not it is appropriate to use the result for comparison purposes and to consider the possible number of contributors prior to comparing to any reference sample. When mixed DNA profiles are interpreted alongside reference sample(s) without any prior assessment of their suitability for comparison, the risk of cognitive bias increases substantially.

Only addressing the prosecution's case when a suspect cannot be excluded

This relates to cases in which the complexity of the DNA result is such that it cannot provide evidence of inclusion but is only suitable to exclude individuals as a possible contributing source. The assertion that an individual cannot be excluded as a possible contributor to such a mixture is often reported without the qualification that there are many other individuals with different profiles who similarly could not be excluded. Only expressing an inability to exclude the presence of the defendants DNA from a case sample invites an interpretation by jurors that favours the prosecution's case more than is justified.

9.4 Mitigation strategies currently deployed in the UK and overseas

Below are examples of mitigation strategies that are variously used in current practice. All are experience-based examples of good practice in appropriate circumstances and should be applied as described:

Prior-interpretation of case sample result before reference result is revealed. Formally noting the following from the DNA result, prior to comparison with the reference profile:

a. suitability to include or exclude;
b. assessment of number of contributors;
c. level of representation of contributors;
d. potential for stochastic effects;
e. identification of likely/unlikely genotype combinations that might explain the mixture.

This is a critical step and is recommended for DNA profile interpretation in all circumstances.

Full checking via repeat interpretation by an experienced and competent colleague including prior-interpretation of case sample result before reference result is known. The check should be conducted independent of, and uninfluenced by, the reporting scientist, and should use original unmodified hard copy or electronic results that are free from annotation. This is a critical step and is recommended for DNA profile interpretation in all circumstances.

Case Assessment and Interpretation. Comparison of expected, pre-assessed outcomes with actual results under appropriate hypotheses. Some documented indication of expected outcome is recommended in all cases.

Careful selection of case stains/samples for testing to minimise the occurrence of mixtures and low template issues. Selection should be informed by case information and is good practice whenever case circumstances present a choice of DNA case stain targets.

Duplicate (or multiple) analyses to assess stochastic effects in low template samples. Replication is often used in conjunction with interpretation in a consensus framework, but can also be used prior to probabilistic evaluation of the results separately. Replication should be applied whenever a poor quality profile is to be relied upon to progress an investigation or provide evidence against a suspect. It assists in evaluating reproducibility, identifying spurious peaks and informing conclusions relating to the likelihood of allelic drop-out and the number of contributors. Replication allows a fuller understanding of the nature of the sample and reduces scope for conjecture and the risk of misinterpretation; it improves the scientist's ability to accurately gauge whether or not the sample is suitable for any form of comparison or statistical evaluation.

Analysis and interpretation is carried out blind, in the complete absence of any information about the case. This approach is practiced in some jurisdictions and eliminates the risk of some types of bias. It does present the practical challenge of separating case strategy, hypotheses testing, stain selection etc. from result interpretation and reporting in the context of the case. The risk of missing identification of realistic alternative explanations for the evidence given the case circumstances may be greater using this approach.

Use of recently developed interpretation software for complex mixtures[53] such as LikeLTD[54], STRmix™ (Institute of Environmental Science and Research (ESR) or TrueAllele® (Cybergenetics). Ideally should be used with all suitable results whenever other objective numerical methods are not appropriate. Efforts should be made to ensure practitioners are able to use them reliably whenever required.

---

[53] Suitable validation of all such methods would be expected prior to introduction in casework.

[54] A software package developed by David Balding, Adrian Timpson, Christopher Steele, Mayeul d'Avezac and James Hetherington. Further details available from: http://cran.r-project.org/web/packages/likeLTD/likeLTD.pdf [Accessed 27/08/2014]

Appropriate training of practitioners in the method employed, who can demonstrate initial and ongoing competency. This is a critical step and is recommended for DNA profile interpretation in all circumstances.

Transparency and disclosure of appropriate experimental data used to support conclusions and opinions. Research work should ideally be published in a peer reviewed scientific journal.

9.5 Further recommendations for good practice

In addition to the good practice described in 7.4 we also recommend the following:

When a numerical evaluation is not possible, it remains of crucial importance that qualitative and subjective judgments of pertinent profile features and their combined likelihood are assessed under the hypotheses framed by both the prosecution hypothesis (Hp) and defence hypothesis (Hd) separately. The final opinion of evidential weight must be based on how much, if any, comparison of separate assessments favours one hypothesis over the other, as with a likelihood ratio. For example, consider a complex mixture that cannot be conditioned on the presence of a known profile: If it is not possible to form a properly reasoned and reliable view about the probability that the mixture could arise if it came from a combination of unknown individuals (Hd), then the result can be of little, if any, probative value because half of the LR is unknown. If this approach is always adopted, it helps practitioners to identify when an observation favours neither prosecution nor the defence and is likely to prevent issues like those described in case examples 7.3.2 and 7.3.5.

Use a completely "blind" checker who repeats the full interpretation described in 7.4.2 but in the absence of any contextual information relating to the case. This may present practical challenges, particularly within smaller organisations. However, it will assist in a continuous learning and improvement cycle, where Reporting Officers can identify instances where they may have been affected by bias. Further, it provides assurance for the courts that the interpretation is free from contextual bias.

If there is no suitable option for objective evaluation, only employ qualitative and subjective based approaches that have been validated and therefore have demonstrated the robustness of resultant conclusions and opinions. Such procedures should include system performance data indicating when the approach breaks down and is no longer valid. The approach should be quality managed with defined standards and safeguards using trained staff who demonstrate initial and ongoing competence. It is also recognised that some scientists perform better than others under cognitive pressures and if a suitable measure can be adopted by providers this would help to mitigate the risks through improved staff selection, training and self-awareness.

Training and education in relation to the risks of cognitive bias generally and specifically in relation to complex DNA interpretation.

9.6 Further Research

The wider use of software packages (see note 50) capable of numerical evaluation of complex DNA results is likely to reduce the frequency with which

issues relating to subjectivity are encountered. However, such software does not yet offer a complete solution and there will continue to be a gap filled by non-numeric interpretation. Whilst best practice will minimise the inherent issues it is likely that there will continue to be a risk of cognitive bias and general disagreement between experts. We recommend continued research into objective methodology that will increase the power of DNA technology and improve the reliability and robustness of the evaluative processes for the benefit of criminal justice.

10. FINGERPRINTS GUIDANCE

10.1 Brief Outline of the Forensic Process

Every finger, palm or sole of foot comprises an intricate system of ridges and furrows, known as friction ridge skin. The arrangement and appearance of features within friction ridge skin are unique to each individual, persist throughout life and are accepted as a reliable means of human identification. Fingerprint Examiners are trained to interpret arrangements of ridge features and to report their opinion as to the common origin or otherwise of any two areas of friction ridge.

The fingerprint examination process consists of stages frequently referred to as Analysis, Comparison, Evaluation and Verification (ACE-V), terms which provide useful descriptors of the cognitive process undertaken by the examiner in arriving at their final opinion.

Each mark is analysed to establish the quality of detail visible within the mark and to determine its suitability for further examination taking account of variables such as:

a. The surface on which the impression was left
b. Any distortion arising from pressure applied when the impression was deposited
c. The clarity, quality and quantity of detail visible in the print.

During the comparison stage the examiner will systematically compare the ridge pattern and sequence of ridge characteristics in an impression from an unknown source with that of a known source impression. They will establish their opinion of the level of agreement or disagreement between the unique sequence of ridge characteristics visible in both impressions.

During the evaluation stage of the process the examiner will review all of their previous observations and come to their final opinion and conclusions about the outcome of the examination process. The ACE-V process is iterative in application with the analysis and comparison stages overlapping on occasion. The examination of a latent print against a known reference print may allow examiners to observe further features within the mark by directing their attention to areas, which require particular attention and further processing. This comparison activity may cause the examiner to reconsider their initial analysis of the mark and which could require further documentation by way of technical notes. The evaluation stage however remains a separate and distinct phase of the ACE-V process.

If the quality and/or quantity of detail visible within either or both impression is lacking, the examiner will record the impression(s) as **insufficient** and generally no further examination will occur. If the examiner is satisfied that the level of agreement between both impressions is sufficient to determine that they were made by a common donor, then they will consider the unknown impression **identified** to a particular individual. If the examiner feels that the level of disagreement between the two impressions is so significant that they are able to determine that both impressions could not have been made by the known donor, then they will consider that particular individual **excluded** as a potential donor of the unknown print. The examiner may conclude that, although there may be some agreement evident, the extent of disagreement and/or the quality and quantity of detail visible in both or either impression is such that it is not possible to come to a definitive conclusion at this time. In such a circumstance the examiner would consider the outcome of that examination to be **inconclusive**[55].

Although the process is often described sequentially, it is important to note that fingerprint examination is iterative in practice and each stage is not mutually exclusive throughout the process.

It is common practice across the fingerprint discipline globally that identifications are subject to verification by further examiner(s) who will conduct a personal analysis, comparison and evaluation of the impressions under examination.

Due to the subjective nature of the interpretative cognitive process undertaken by the examiner in arriving at their final opinion, it is accepted that the information used to come to conclusions may vary between examiners. For example, individual examiners may approach their examination from different starting points or consider the visible features in differing sequences; however, the original conclusions are shown to be reliable through demonstrating consistent end results from all subsequent examiners.

10.2 Risks of Cognitive Bias

The subjective, iterative and interpretative elements inherent within the fingerprint examination process expose the fingerprint examiner to a range of cognitive influences which, if not properly managed, could impact on the reliability of examination outcomes and examiner opinion.

Significant research has already been undertaken across the fingerprint discipline to explore the impact of cognitive influence and human factors on the examination process and the examiners personal decision-making behaviours. Studies undertaken to date have established that fingerprint examiners will, on occasion, alter their original opinions and conclusions in circumstances when

[55] Not every UK bureau use the same toolbox terminology at this time and 'inconclusive' may not be an option for some to use. This places a cognitive burden on the examiner to side with decisions that may lead to stronger biasing implication. To this extent 'inconclusive' could be a valuable tool to the decision-making armoury.

the original material is presented in a different context[56]. Further research has indicated that this influence is more prevalent when the impressions under examination are of poorer quality[57].

The risks of cognitive bias inherent in the fingerprint examination process can be categorised as contextual, confirmation and cultural.

Contextual bias

Fingerprint examiners are exposed to a wealth of contextual information which will impact on their decision making process such as;

a. Nature and details of the crime including background information
b. Association with or personal knowledge of the victim or their circumstances
c. Status of suspects or person(s) already in custody for the crime
d. Previous criminal activity of suspects or persons of interest
e. Location of the crime (an area close to their home)
f. Media or public interest associated with the crime
g. Personal moral codes or behaviours
h. Time pressure from investigating officers or office managers

For many organisations, contextual influence relating to crime type is in fact imbedded within their standard operating procedures. Crimes of a serious nature such as murder, rape and sexual assault are often given priority over other case work, have additional quality assurance measures in place or have specialist teams dedicated to this type of case work.

Prior knowledge of contextual information can influence the decision making process of a fingerprint examiner. For example, during an analysis an examiner may be more likely to retain an impression of borderline quality submitted as part of a serious crime than if the same impression was submitted as part of a low level volume crime. Prior knowledge of the status of an arrested person can lead to particular focus or emphasis on that individual to the exclusion of others.

Confirmation Bias

Within operational fingerprint bureaus, the majority of examination requests are received from police officers or prosecution services, with both hoping that the examination outcomes will help "solve the case" or "secure a conviction". Contributing to the detection of crime is considered a fundamental aspect of fingerprint bureau service delivery. Also, personal identification or "hit" rates are used as key performance indicators at both organisational and individual level.

---

[56] Dror, I. et al (2006 check) Contextual Information Renders Experts Vulnerable to Making Erroneous Identifications: Forensic Science International 156 74-78

[57] Dror, I. et al (2005) When Emotions Get the Better of Us: The effect of Contextual Top-down Processing On Matching Fingerprints, Applied Cognitive Psychology, Wiley InterScience DOI:10.1002/acp 1130

Combined with a personal moral code to "do the right thing," this emphasis on "identification" as the most favoured hypothesis will exert powerful cognitive influence on examiner decision making.

Having prior knowledge of the previous examiner's findings and conclusions may also expose fingerprint examiners to the risk of confirmation bias and this will have a particular importance during the verification process.

At a technical level, examiners can be unduly influenced by confirmation bias when, having found a number of features from an unknown impression to agree with features in an impression from a known source, the examiner will then begin to reason backward, finding features in the unknown impression which are suggested by those in the known print rather than being visible without reference to the known source material.

Dror's paper "Practical Solutions to Cognitive and Human Factor Challenges in Forensic Science"[58] discusses the issue of base rate regularities and the impact of new technology into the fingerprint examination process. Within the context of automated fingerprint identification systems (AFIS) examiners become accustomed to having positive hits positioned at or near the respondent list. AFIS systems are designed to return those candidates most similar to the mark under search. The combination of heightened expectation of an identification being at top of the list along with the most similar candidates being returned at the top of the list carries with it an increased risk of cognitive influence on the decision making of fingerprint examiners.

Cultural Bias

Individual perception is influenced by the environment in which they are operating. Prior to the publication of The Fingerprint Inquiry Report in 2011, there was a tendency to represent the findings of fingerprint examiners as statements of objective fact rather than expressions of informed technical yet subjective opinion, albeit an opinion based on sound training and experience.

Historically, investigating officers and courts have accepted fingerprint evidence without challenge, which further contributed to the perception that fingerprint examination enjoyed "practical infallibility".

Operating in environments where differences of opinions are perceived as disputes with a "right" or "wrong" answer can also exert a powerful cognitive influence on examiners, leaving them reluctant to challenge their own or the findings of others.

Further examples of cultural influence which can impact on the decision making process include;

a. Strict hierarchical structures based on time served rather than competence.
b. Over confidence in individual or organisational competence.

---

[58] Dror, I.E. (2013) Practical solutions to cognitive and human factor challenges in forensic science. Forensic Science Policy & management 4 p1-9.

c. Lack of interaction with peers or exposure to alternative methods of working.
d. Lack of acceptance of the potential for errors or effective root cause analysis of errors.

The Fingerprint Inquiry report called for the profession to move away from any presentation of fingerprint evidence with 100% certainty, to fully explore the cogency of explanations offered for any evident differences between impressions and most importantly to recognise that fingerprint evidence is opinion evidence and as such is inherently subjective.

Any process which relies on the subjective personal interpretation of data as part of the decision making process is at risk from the influence of cognitive bias. This influence is typically exerted at an unconscious level and examiners often believe that their personal strategies are sufficient to mitigate any associated risk of cognitive bias. However experience has shown this not to be the case.

The challenge for the fingerprint profession is to adopt effective risk management strategies at individual and organisational level but without impacting on service delivery.

10.3 Examples where cognitive risks have become an issue

Brandon Mayfield Case 2006

In May 2004 Brandon Mayfield, an Oregon attorney, was arrested by the Federal Bureau of Investigation (FBI) as a material witness in an investigation of terrorist attacks on commuter trains in Madrid, Spain. In March 2004, the FBI fingerprint department had conducted a computer database search of an impression found on a bag of detonators and identified the impression to Brandon Mayfield. Two weeks after Mayfield's arrest, the Spanish National Police (SNP) informed the FBI that they had in fact identified the print to an Algerian national called Daoud.

The FBI compared Daoud's prints with the impression on the bag of detonators and agreed the findings of the SNP. They subsequently withdrew their previous identification of Brandon Mayfield.

The U.S. Department of Justice, Office of the Inspector General (OIG) launched a review into the FBI's handling of the case and provided an assessment of the causes of the misidentification. FBI examiners initially found 10 features they believed to be in agreement with Mayfield's prints. The OIG report [E] concludes; "...the unusual similarity in position and ridge counts was a critical factor that misled four examiners and contributed to their overlooking other important differences between LFP 17 and Mayfield's fingerprint" (Executive Summary). This conclusion implies that due to the unusual level of similarity, examiners were less focused on information which would negate the hypothesis of identification. The report further states; "There were also other subtle but important differences between the prints in the positioning of the features. But the unusual similarity in position and ridge counts was a critical factor that.....contributed to their overlooking other important differences" (Executive

Summary). It would appear that the examiners applied a lower level of scrutiny to the information which supported their favoured hypothesis of identification.

The OIG found that the examiner's interpretation was also influenced by circular reasoning, working backward from the known source material; "Having found as many as 10 points of unusual similarity, the FBI examiners began to 'find' additional features that were not really there, but rather were suggested to the examiners in the Mayfield prints" (Executive Summary). Again the examiners would seem to be unconsciously seeking out information to confirm their favoured hypothesis of identification and this is a consistent theme throughout the assessment of the causes of the errors, particularly with regard to the explanation offered by the examiners for observed differences between the prints. "This explanation required the examiners to accept an extraordinary set of coincidences. The OIG found that the support for this explanation was, at best, contradictory" (Executive Summary).

Shirley McKie Case 1999

During the 1997 trial of Mr. David Asbury for the murder of Miss Marion Ross, Ms. McKie, one of the investigating officers, did not accept that an impression from the crime scene, identified to her by experts from the then Scottish Criminal Records Office (SCRO) could have been made by her.

Ms. McKie was subsequently charged with perjury in 1999 and at her trial the SCRO identification was challenged and refuted by American Fingerprint Experts, Mr. Pat Wertheim and Mr. David Grieve. These experts also challenged the identification of an impression which had been presented as part of the prosecution case against Mr. Asbury.

The jury unanimously found Ms. McKie not guilty; however the fingerprint evidence remained a matter of dispute and controversy across the national and international fingerprint community for the next decade and was subject to a Scottish Government Justice Committee Inquiry in 2006. In March 2008 Sir Anthony Campbell was appointed to hold a public inquiry into the identification and verification of the fingerprints associated with HM Advocate v McKie 1999. The Fingerprint Inquiry Report was published in December 2011 stating that two misidentifications had occurred and also presented an in-depth scrutiny of fingerprint examination methodology and associated issues.

On discussing the causes of the errors Sir Anthony Campbell stated; "The method of work described by the four SCRO officers displays a number of recognised risks factors and in the case of Y7 and QI2 Ross it is likely that these risks crystallised into the misidentification"[59].

Amongst risk factors identified in the SCRO methodology listed below are those which are relevant to the cognitive bias issues under discussion in this paper:

---

[59]Campbell, A. (2011). The fingerprint inquiry report. Available at: http://www.thefingerprintinquiryscotland.org.uk/inquiry/3127-2.html

Practitioners being taught 100% certainty which could be attained prematurely in the examination process on the basis of relatively few characteristics.

Establishes an inner conviction which can lead to a circular argument discounting differences which must be capable of explanation even if the examiner is not sure what that explanation is.

Diminishes the independence of the verification process because a verifying examiner might tend towards confirming the view of the first examiner particularly if the examiner is senior in experience or rank.

Diminishes the usefulness of asking an examiner to reconsider their findings – if they have already reached a conclusion with 100% certainty then unsurprising that a re-examination would typically lead to a confirmation of the initial findings

The ethos in the SCRO fingerprint bureau where pride was taken in an ability, particularly on the part of more experienced officers, to identify marks that other bureaus might not consider sufficient for identification[60].

An inappropriate hierarchical philosophy

Examiners could be influenced to make identifications or confirm identifications of senior officers, where the quality and volume of information did not properly support identification.

The application of inappropriate tolerances in the observation and interpretation of detail in marks and prints, reverse reasoning and the influence of repeated viewing of known prints.

Contextual information from the police, which may subconsciously influence the conclusions of fingerprint examiners.

10.4 Examples of mitigation strategies.

IPOL Unit, Netherlands Police Service, Zotermeer

The IPOL unit has introduced a structure and workflow process specifically designed to mitigate the risks associated with cognitive bias.

The fingerprint unit is established around regional centres and a central hub. Latent images are input by staff at the regional centres, sent for search on the automated fingerprint recognition system and then processed by examiners at the central hub. These examiners receive only the on-screen image, with all lifts and case information retained at the regional centres.

This structure effectively removes any risk of contextual influence affecting the examiner's technical decision making.

---

[60] This topic is discussed in some detail in: Charlton, D., Fraser-Mackenzie, P.A.F. & Dror I.E. (2010). Emotional experiences and motivating factors associated with fingerprint analysis. Journal of Forensic Sciences, 55, p385-393

Prior to processing the search, the examiner must conduct an onscreen analysis without reference to any comparison print. They are required to demonstrate a minimum of 12 unique features in the print before proceeding with the features graded for suitability for use in the initial findings. Any further features identified at comparison phase are highlighted as such and appropriate tolerances applied. This type of workflow mitigates the risks of cognitive influence associated with the application of inappropriate tolerances in the observation and interpretation of detail in impressions.

Federal Bureau of Investigation (FBI) Latent Print Unit

Following the procedure review instigated as a result of the Brandon Mayfield Case, the FBI introduced a system of blind verification. They have defined blind verification as "the independent application of Analysis, Comparison, and Evaluation (ACE) to a friction ridge print by another qualified examiner who does not know the conclusions of the primary examiner"[61]. The FBI further state that blind verification should; "eliminate confirmation bias and limit contextual bias in the examination process".

Blind verifications take place in cases with a single mark conclusion, circumstances where there are conflicts between examiners and also on decisions of "value" or "no value". The FBI are clear that blind verifications cannot be performed by any examiner who has previously been consulted by the primary examiner, who has knowledge of the previous examiner's conclusions, any knowledge of the information used by the primary examiner or and specific background case details.

The FBI accepts that some consultation is necessary for the sharing of expertise and that not every consultation between examiners is indicative of a complex analysis. However an analysis is considered complex when dissimilarities or factors influencing the quality of the print could interfere with the proper interpretation of the impression. When a complex analysis or conclusion results in an identification, examiners are required to document any explanation for differences caused by apparent distortion and identify the supporting data for their explanation in the case record.

Scottish Police Authority Forensic Services (SPA FS), Fingerprint Units

In anticipation of the publication of The Fingerprint Inquiry Report 2011 SPA FS established a series of work streams to consider good practice in relation to the cognitive influence issues raised as a result of the McKie case.

It was accepted that a certain amount of case context is required to allow the initial examiner to develop an effective case assessment strategy, however SPA FS recognised that it was not essential for subsequent examiners to have access to this information on every occasion.

---

[61] Dror, I.E., & Cole, S.A., (2010). The vision in "blind" justice: Expert perception, judgment, and visual cognition in forensic pattern recognition. *Psychonomic Bulletin & Review* **17(2)**, 161-167

A proportionate risk management approach was adopted to mitigate risks of cognitive influence without impacting on service delivery. A range of measures was developed;

a. Improved note taking, including demonstration of features used in lead identifications.
b. A complex marks process to manage variance in opinion between examiners. This process includes a blind technical review process, where examiners are required to prepare technical reports and supporting visuals following a completely independent review of the relevant impressions. Those involved in the technical review process have no prior knowledge or access to case-related information or the technical findings of any other examiners.
c. A blind verification process for lead identifications in which verifying examiners have no knowledge of the technical findings of any previous examiners.
d. The removal of any case context information or related communication documentation from the verification process in any circumstance.
e. Regular dip-sampling of all completed case work.
f. Training programmes for examiners exploring cognitive bias and its impact on the human decision making process.

Surrey and Sussex Forensic Identification Services Unit (FISU)

Surrey and Sussex Forensic Identification Services Unit have followed similar processes to SPA, and have also introduced cognitive profiling recruitment tests which have proven very effective at predicting cognitive skills of new staff, thus improving effectiveness and efficiency in managing cognitive influence.

Other parameters under consideration by FISU are longitudinal studies to underpin cognitive issues with overall accuracy and performance, and embedding cognitive processes to mitigate risks in using new technologies (remote transmission and on screen annotation tools).

10.5 Recommended good practice

The Codes (section 20.4) states that once a method has been designed or determined, there should be an assessment to identify any risks including; "identifying areas where the operation of the method, or interpretation of the results, requires specialist skills or knowledge to prevent ambiguous or misleading outputs or outcomes". An organisation should therefore adopt a risk management approach to the fingerprint methodology as applied within their organisation to identify, assess and evaluate the threats and consequences posed by the issue of cognitive bias. Practical solutions could include the introduction of a blind element to the verification process or randomising the respondent lists delivered through AFIS searches[62].

[62] Dror, I.E. (2013) Practical solutions to cognitive and human factor challenges in forensic science. Forensic Science Policy & management 4 p1-9.

Further generic guidance from The Institute of Risk Management states that; "Risk Identification should be approached in a methodical way to ensure that all activities within the organisation (or method) have been identified and all the risks flowing from these activities defined"[63]. Once identified, the risks should be displayed in a structured format, which can then be used to evaluate the consequences of the risk including the probability of occurrence. Risk assessment in this manner allows the organisation to break down each stage of the process and consider how best the impact can be mitigated. Areas to be considered can include:

a. Name of Risk
b. Scope of Risk
c. Nature of Risk
d. Stakeholders
e. Quantification of Risk
f. Risk Tolerance
g. Risk Treatment & Control Mechanisms
h. Potential Action for Improvement.

Suitable Risk Treatment and Control Mechanisms for consideration with regard to fingerprint examination are listed below:

a. Survey and breakdown extent of current contextual information available to examiners & assess added value each piece of information brings to the examination process.
b. Remove or limit contextual information which adds no tangible value to the fingerprint examination process.
c. Remove or limit contextual information made available to verifying or subsequent examiners.
d. Introduce a blind verification process for identified case work assessed as at greatest risk from contextual, confirmation and/or cultural bias.
e. Introduce a blind element to a technical review process for analyses, comparisons and/or evaluations which are considered complex or cause a variance in opinion between examiners.
f. As part of a technical review process for complex marks or circumstances where examiners have a variance in opinion, introduce an appropriate and proportionate note-taking strategy which requires examiners to provide written and visual accounts of their reasoning and findings.
g. Develop bespoke training programmes to raise awareness of the cognitive issues involved in human perception, judgement and decision making.
h. As part of an established quality management system, instigate an effective review and monitoring process to provide assurance that the risk treatment and control measures continue to provide effective risk management.

---

[63] Institute of Risk Management (2002) "A Risk Management Standard" IRM

11. FOOTWEAR, TOOL MARK AND FIREARMS COMPARISON AND FIREARMS CLASSIFICATION GUIDANCE

11.1 The generic marks comparison process

Introduction

The generic forensic process that is outlined below encompasses the interpretation and reporting of 'marks' comparison cases. It is applicable to a wide range of evidence types such as firearms, footwear, and tool marks and outlines a practical strategy that can be used to counter potential cognitive bias when carrying out 'marks' comparison cases:

With regards to tool mark comparison this section should be read in conjunction with Regulator Codes of Practice and Conduct – Draft Appendices Toolmarks – HOS/12/027

With regards to footwear marks related comparisons this section should be read in conjunction with Regulator Codes of Practice and Conduct – Draft Appendices Footwear – (HOS/11/059)

With regards to firearms related comparisons this section should be read in conjunction with the Regulator Codes of Practice and Conduct – Draft Appendices Firearms – HOS/12/026, Microscopy and Firing Marks.

The strategy also addresses the possible low expectation of a 'hit" when screening through a firearms Open Case File (OCF)[64]

Confirmation bias in firearms classification examinations is also addressed. In this context this section should be read in conjunction with Forensic Science Regulator Codes of Practice and Conduct – Draft Appendices Firearms – HOS/12/026, Classification of Firearms and Ammunition.

Process outline

Items are recovered from the crime scene and may consist of the original item or a 'true' copy of the mark generated by other methods.

Items are received along with case information and questions to be addressed by the scientific work.

The case information, supplied by the customer, is used to direct the item examination recovery and analysis strategy, ideally within a framework of appropriate propositions.

a. Examination of the item/mark recovered from the crime scene.
b. Use of recovery and enhancement techniques as required.
c. Generation/Examination of the 'control' item
d. Make test marks if required in the appropriate manner.
e. Undertake a comparison using appropriate methods and equipment

[64] An OCF is defined as an organised collection of ammunition components derived from crime scenes that is intended to be compared against test fired and crime scene ammunition samples in order to establish whether or not a single gun has been used at one or more scenes.

f. Interpret and evaluate findings
g. Verification of result
h. Findings are described in a statement or report.
i. The scientist may be called to court to give oral testimony.

11.2 Risks of cognitive bias

A marks comparison seeks to establish if a 'mark' (the unknown) has been made by the submitted exhibit (the known) or has been made by the same item e.g. a revolver which has not been recovered could be responsible for discharging multiple bullets recovered from multiple scenes. It is based on the comparison of detail and is therefore observational. The scientist is looking to determine if the detail present in the mark matches characteristic detail on the item or in a test mark or is significantly different. An assessment of what the detail is and how it has been produced must consider general characteristics common to a set of items (CLASS), unintentional manufacturing marks present on a sub-set of items (SUB-CLASS) through to random damage/wear and tool mark characteristics (INDIVIDUAL). Any examination is therefore dependent upon the visual quality and clarity of the detail that is observed by the examiner. The process is one of pattern recognition aided by the use of equipment such as photographic/imaging, low power microscopy and comparison microscopes. The final assessor of the level of significance of any agreement between the marks is the human operator; there is no significant instrumental analysis [W]. In footwear mark comparisons, the methods employed by footwear practitioners are normally side-by-side comparisons or overlay. In this way the footwear expert assesses the level of agreement in terms of the pattern, pattern configuration, mould/moulding detail, wear and damage. The assessment is subjective, although reference material and data can be used to support the evaluation of the findings. In tool mark/firearms comparisons there are currently two methods; traditional pattern recognition where the examiner's opinion is based on the relative extent of detailed agreement with a best known-non-match and Consecutive Matching Straie (CMS) where the examiner applies a conservative criteria of runs of aligned straie to establish a possible match. Both techniques use subjectivity.

The interpretation and evaluation of a 'marks comparison' may potentially be affected by some form of unintended bias. In the interpretation process there are no results produced by a 'black box'; opinions and decisions are based on the individual's, relevant experience, depth of knowledge and skill as well as their disposition at the time. Every effort must be made to make it logical, transparent, balanced and robust. Usually the opinions are formed in the context of supplied case information, introducing the possibility of contextual bias.

Within marks interpretation it is considered that there is a spectrum of bias risk (table 2).

Risk factor	Low risk	High risk
Detail	The detail in the mark(s) is clear, well defined and unambiguous	Marks are confused and complex, of poor quality and the detail present is poorly defined.
Equipment	Optimum visualisation of the detail in a mark using appropriate equipment/imaging and enhancement techniques.	Poor or inappropriate equipment/imaging and enhancement techniques.
Approach/Examiner	There is a methodical approach with defined standards built on principles that have been tested and validated. Possible confirmation bias may reduce as a consequence of the comparison reviewer having less contextual information[65]	When the approach is un-researched, ad hoc and personal to the operator. When the expectation of an OCF hit is very low.
Scientist/Examiner	Scientist/examiners are well trained, experienced and continuously meet acceptable standards of competence	Scientist/examiners are inexperienced, unmonitored and left to adopt their own approach.

Table 2: Spectrum of bias risk in marks interpretation

a. Risks are low when results are clear and unambiguous and greater when results are complex, of poor quality and there is an increased reliance on subjective opinion.
b. Risks are lower when there is a methodical approach with defined standards built on principles that have been tested and validated and greater when the approach is un-researched, ad hoc and personal to the operator.
c. Risks are lower when equipment is well maintained and functioning to the required standard.
d. Risks are lower when operators are well-trained, experienced and continuously meet acceptable standards of competence and results are peer reviewed, and greater when operators are inexperienced, unmonitored and left to adopt their own approach.
e. Contextual and confirmation bias risk is lower when the contextual information is minimised, particularly at the comparison review stage and the reviewer is unaware of the examiner's opinion, or other evidence that relates to the 'marks' examination.

---

[65] Kerstholt, J., Eikelboom, A., Dijkman, T., Stoel, R., Hermsen, R., van Leuven, B., Does suggestive information cause a confirmation bias in bullet comparisons? (2010) *Forensic Science International* **198** 138–142

f. Expectation bias manifesting in the missing of an OCF hit is lower when there is an expectation of success[66].

Other more general bias risks within "Marks" and firearms examination and classifications:

a. Observations that support the defence case are less rigorously considered or evaluated and are not given their true weight.
b. Interpreting the Firearms Act 1968 when classifying potential component parts or antiques. Confirmation bias on the status of firearms should be avoided; this is particularly pertinent where the prosecution expert relies upon Home Office Guidance, which is not explicitly reflected in the legislation.
c. Reluctance to express doubt particularly during oral evidence at court.
d. Reluctance to clearly understand and express the limitations of a comparison after a time delay between the offence and the recovery of a suspect item.
    i. The comparison of footwear a footwear mark recovered at a crime scene to footwear recovered months later.
    ii. The assessment of the significance when there is matching and non-matching characteristic detail in the mark.
e. Failure to express alternative explanations, such as possible sub-class origins and arguments for alternative firearms legal classifications.
f. A failure to assess detail correctly due to a lack of knowledge and the inability to investigate due to location of manufacturing plant or time and cost considerations.

11.3 Examples where risks of bias have become an issue

a. The identification of a tool being responsible for cutting a wire fence, where detail was clearly visible that excluded the suspect tool.
b. Situation where critical findings checks were being undertaken on a basis of 'I will check yours if you check mine'. An independent approach was not maintained.
c. The association of two crime scenes in the same geographic area, involving crimes of similar *modus operandi*, calibre, make and model of gun. Possibly due to confirmation and contextual bias compounded by lack of awareness of differences between sub-class and individual characteristics.
d. The automatic classification of vintage firearms as not being subject to the section 58(2) exemption provided for antique firearms, due to the prosecution expert relying on "official" guidance as opposed to statute, possibly as a result of confirmation bias.

---

[66] Nennstiel R., (2010). The Human Factor in Detecting Cold Hits, Association of Firearms and Toolmarks Examiners Annual Training Seminar. Henderson, Nevada, USA, 2nd – 7th May 2010.

e. Classification of possible component parts of a firearm as being subject to the 1968 Act without consideration of any alternative hypothesis most probably due to confirmation bias.

11.4 Mitigation strategies currently deployed in the UK and overseas

Examples of mitigation strategies that are variously in current practice are listed below. These are considered to be good practice in appropriate circumstances:

a. Case Assessment and Interpretation. Comparison of expected, pre-assessed outcomes with actual results under appropriate hypotheses.
b. Full disclosure of all data used in the evaluation.
c. In all firearms classification cases, the reviewer should clearly set out what is official guidance and what is statute, ensuring that alternative classification hypotheses are addressed to counter any confirmation bias.
d. Use a completely "blind" checker who repeats the full interpretation, but in the absence of any contextual information relating to the case. Initially, the checker should not be aware of the opinion of the reporting scientist.
e. An acceptable alternative is that result will be subject to a critical findings check by a second authorised examiner. The initial practitioner completes the comparison and records what items they have examined, their findings together with their conclusion. The checker then undertakes a detailed independent review wherever possible without knowledge of the previous practitioner's conclusion. The aim of the check is as follows:
    i. The examiner has followed the appropriate documented examination process and applied the appropriate relevant scientific methodology and techniques.
    ii. The work and findings of the examination are reflected in the conclusion of the report. The results must support the conclusion and clearly there should be an understanding or statement of the findings.
    iii. The maximum evidence has been obtained, that nothing has been overlooked and there are no other marks that may change the outcome.
    iv. The submitting authority's question has been fully addressed.

In addition to the good practice described above the following are also recommended:

a. Validation testing of qualitative and subjective based approaches to demonstrate the robustness of conclusions and opinions.
b. Development of standards and quality managed procedures for qualitative and subjective based methods, including system performance data indicating when the approach breaks down and is no longer valid.
c. Practitioner training in the specific method used, together with initial and on-going competency assessment.
d. Training and education in relation to the risks of cognitive bias in firearms classification and marks comparison generally.
e. An approach to quality that includes the assessment and monitor of on-going competence of practitioners including the use of proficiency tests, declared and undeclared trials.

f. Providers should ensure that a validated form of Context Management is applied.
g. The use of blind trials should be introduced to increase the "success" rate of cold OCF hits.

12. TRACE EVIDENCE (INCLUDING HAIR AND FIBRE) GUIDANCE

12.1 Outline of the Forensic Process for Trace Evidence analysis

The examination of trace evidence covers a wide range of materials including particulate material such as glass, paint, hairs and fibres. However whilst the range of trace materials is wide, the analysis of such material essentially follows the same process which involves comparison of crime (unknown/recovered) material with one or more known/reference samples. This process can briefly be described as follows:

Item receipt: items are received along with case information and questions to be addressed by the scientific work. When dealing with contact traces, taking and submitting the right reference samples (from the crime scene or individuals) is critical as it can have a fundamental impact on the subsequent comparison.

Case assessment: case information is used to direct the strategy for item examination and trace evidence recovery and analysis. Ideally case assessment should be carried out with in a framework of appropriate propositions. By its nature trace evidence examination is time consuming, so practicality and cost have to be considered. Case assessment can assist with targeting the exhibits most likely to yield probative evidence.

Recovery of trace materials using appropriate techniques

Identification of target material and comparison with reference sample(s):

a. Whichever recovery technique is used, the examiner is often presented with a large amount of debris which may potentially contain some of the target material. Where there is a limited amount of target material of interest which can be immediately identified, e.g. glass fragments, paint fragments, this material can be recovered in its entirety or a sample taken. The material can then be compared with the relevant reference sample(s) using the appropriate microscopy and instrumental/analytical techniques.
b. With other evidence types, for example fibres and hairs, there will often be a large amount of material collected which is of no relevance to the case. For this reason it is necessary to review the reference sample(s) and use features to enable an initial search of the recovered material to locate that which is of potential interest. For example, for hairs and fibres a search of tapings under a low power microscope would be conducted to locate hairs/fibres with similar macroscopic features (colour, length etc.) to the recovered hairs/fibres. This material can then be recovered for more detailed comparison with the reference samples using the appropriate microscopy and instrumental/analytical techniques.

c. Evaluation of the scientific findings and interpretation within the context of the case specific information available (may be at source or activity level as appropriate).
d. Provision of report or statement describing the findings and providing opinion on their significance.
e. Oral testimony - the scientist may be called to court to give evidence.

12.2 The Risk of Cognitive Bias in Trace Evidence analysis

As in other areas of forensic science, trace evidence analysis can potentially be affected by some form of subconscious and unintended bias and will be a particular risk where subjective interpretations are required. Trace evidence examinations can broadly be divided into two groups:

Those that are entirely subjective and based on mainly observational skills, for example, the microscopic comparison of hairs or the comparison of the layers of paints in a microscopic fragment, which relies exclusively on a subjective assessment of whether the crime and reference samples match.

Those that may include an initial subjective element, followed by the use of objective instrumental techniques to confirm or eliminate matches. For example, analysis of paint after a visual comparison and fibre comparisons where the subjective microscopic examinations can usually be followed by the use of a range of instrumental/analytical techniques including Microspectrophotometry, Fourier Transform Infrared, Raman spectroscopy and Thin Layer Chromatography. Hair comparisons have no similar follow up tests (unless dyed), other than DNA analysis (nuclear or mitochondrial DNA) which, because of the cost and the destructive nature of the testing, is often not an option.

Additionally, opinions are formed in the context of the information supplied about the case and the samples submitted e.g., where and how the glass was broken, how close the person was to the breaking glass, how long after the incident/alleged contact clothing was recovered etc. This may introduce contextual bias[67]. Regardless of contextual case information, practitioners may have a higher expectation of observing matching hairs, fibres, glass etc., simply because the samples have been submitted by the police investigators.

Due to the nature of trace evidence, the recovery and comparison is time consuming and requires a high level of skill, knowledge and often patience. In all cases involving contact traces, there is a requirement for relevant case information to be available to the practitioner to allow effective case assessment. Where fibre evidence is being considered, without information it would be impossible in all but the simplest cases to effectively target those fibre transfers which are viable and would be most probative, thus keeping the time expenditure at a level commensurate with the requirements of the case. This will also apply to hair examinations, where the population of hairs potentially of interest is large.

---

[67] Miller, L. (1987) Procedural Bias in Forensic Science Examinations of Human Hair, Law and Human Behaviour 11(2) p157-163

Within trace evidence examinations, there is a spectrum of bias risk:

Risk Source	Low risk	High risk
Case Assessment	Full case assessment considering potential outcomes, preferably considering at least two competing hypotheses	No case assessment; only one hypothesis considered.
Examination process	Empirical analysis using instrumental techniques	Subjective microscopic analysis only
Result Quality	Results are clear and unambiguous	Results show wide intra-sample variation, are of poor quality and there is an increased reliance on subjective opinion.
Interpretation Approach	There is a methodical approach with defined standards built on principles that have been tested and validated	The approach is un-researched, ad hoc and personal to the operator.
Operator Competence	Operators are well trained, experienced and continuously meet acceptable standards of competence	Operators are inexperienced, unmonitored and left to adopt their own approach.
Checking	Independent confirmation of critical observations. Full independent reinterpretation	No checking or checking is conducted collaboratively

Table 3: Spectrum of bias risk within trace evidence examinations

a. Risks are high where no case assessment is carried out with respect to the potential outcomes of the examinations and the expectations of the examiner, preferably considering at least two competing hypotheses. Risks are reduced significantly where a documented assessment is carried out, the potential outcomes of the examinations are considered in the light of the relevant contextual information available, and the expectations of the examiner are recorded.
b. Risks are low when empirical analysis forms part of the examination processes, and greater where there is an increased reliance on subjective observational analysis.
c. Risks are low where results are clear and unambiguous (for example with a strongly coloured manmade fibre sample which shows little intra-sample variation) and is higher where there is wide intra-sample variation

(for example with a shoddy mix of fibres where it may not be possible to use instrumental techniques to confirm microscopic matches).

d. Risks are low if there are sufficient reference samples showing all possible variations for example within a painted surface, hair from different parts of the head, all broken windows have been sampled etc. Risks are higher if only a limited reference sample is available and may result in the practitioner making a subjective assessment of the match.
e. Risks are lower when there is a methodical approach with defined standards built on principles that have been tested and validated and greater when the approach is un-researched, ad hoc and personal to the operator.
f. Risks are lower when operators/checkers are well trained, experienced and continuously meet acceptable standards of competence; they are greater when operators/checkers are inexperienced, unmonitored and left to adopt their own approach.
g. Risks are lower when critical observations, such as paint layer colours and sequence, are checked independently by another competent practitioner and higher where no critical observation checks are carried out.
h. Risks are lower when interpretation is checked by a competent peer who conducts a separate interpretation, fully independent and without influence from the reporting scientist. Risks are higher when checking is less rigorous and/or conducted collaboratively.

For some trace evidence there are data to support the practitioner. Studies of glass have been undertaken over many years and provide a great deal of data regarding background population, persistence on clothing, breaking windows and the transfer of glass fragments; refractive index information and analytical data for different types of glass are also available. For fibres, there is considerable empirical data to support interpretations, such as population studies and target fibre studies but there is currently no fibre database which provides any guidance with respect to how common a particular fibre might be in the general fibre population. Previous databases (Forensic Science Service) went some way to providing this, but constantly changing fashions and fibre technology changes mean that any database is almost impossible to keep up to date. Therefore, any assessment regarding how common (or otherwise) a fibre might be is essentially subjective and based on the scientist's experience, unless specific industrial enquiries can be made for a particular case.

Fibre, hair and trace evidence analysis generally are becoming less used, and therefore the risk that the examinations are not carried out by practitioners who are dealing with the evidence on a routine basis is increasing. The lack of work in this field has serious implications for the maintenance of scientists' experience and competence and a reduction in the number of practising scientists may ultimately result in there being no one suitable to undertake peer-review.

It is not operationally practical to carry out a full independent check of microscopic fibre matches where large numbers of fibres have been recovered from tapings and individually examined; but where a range of instrumental and analytical techniques are employed which back-up the subjective microscopic

matches this is not necessary. However, where subjective observational methods are the only option, for example in hair comparisons, a full independent check is vital.

With budgetary constraints a certain amount of 'pre-assessment' is often carried out by police forces before selected items are submitted to a forensic provider for examination. There is a bias risk inherent in this process, particularly where the practitioner is not fully informed. For example, other items seized but not submitted for examination may be potentially be an alternative, legitimate source of matching fibres.

12.3 Case Examples where Cognitive Bias May Contribute to Error

The analytical processes for trace evidence have largely remained the same for several decades. As a result methods have been validated and well-tested in forensic casework. The authors are unaware of any specific examples where the results of the microscopic comparison of trace evidence, or subsequent analytical testing of the material has been an issue in case work in the UK. The area of high risk with respect to bias in trace evidence analysis is that of the case evaluation and interpretation where contextual bias might be introduced. Whilst no specific casework examples can be provided where cognitive bias may have contributed to interpretational error, the following hypothetical examples involving glass and fibre examinations are offered where bias might be observed:

Absence of matching glass fragments concluded as being inconclusive

Clothing is submitted from a suspect who is believed to have been seen breaking a glass window and who was arrested shortly after the incident. The practitioner would have a high expectation of finding glass fragments on the persons clothing (choice of clothing to examine would depend on the height of the window). If the relevant clothing was examined and no glass is found then what should the practitioner conclude? As a simple observation then it could be said that no glass was recovered, however this provides no evaluation of the significance of the evidence. Often it is concluded that the findings are inconclusive as it is not possible to comment as no glass was found. If the practitioner evaluates the evidence using a structure of alternative propositions, one reflecting the prosecution view and one the defence view (or a hypothetical defence view if appropriate) the lack of any glass fragments may well support the view that the suspect was not involved in breaking the window as alleged. Therefore reporting the findings as inconclusive might be considered biased.

Absence of matching fibres concluded as being neutral

The examination of car seat tapings for a transfer of fibres from the clothing of an individual who is alleged to have stolen and driven the car for some hours results in no matching fibres being found. The defendant has made no comment. In this situation, it is tempting to conclude that the absence of matching fibres is neutral and does not assist in addressing whether or not the individual had been in the car. However, if the information available provides no explanation for the absence of matching fibres (for e.g., the defendant might have had had time to change clothing before arrest) and the scientist had a high expectation of finding matching fibres if the contact had occurred as alleged, the

absence of matching fibres may well support the view that the defendant had not been in the car. Even where a 'no comment' interview has been offered by the defendant, a good case assessment at the outset requiring consideration of the full range of outcomes and potential defence scenarios, including the absence of any matching fibres, would be likely to result in this type of bias being eliminated.

Difference in treatment of crime and reference material post transfer

A fibre examiner faces considerable difficulty in dealing with cases where clothing has been altered at a chemical level in the period between the offence and seizure of the clothing, for example where the body of a victim has been submerged in a river or at sea for some time, causing the dye in the clothing to fade. In this situation, the challenge for a fibre examiner is firstly searching for fibres without a reference sample that is representative of the fabric at the type of the offence, and then having to interpret a population of fibres on a suspect's garment which does not match the control, but perhaps did at the time of the offence.

A European Textile and Hair Group (ETHG) collaborative exercise in 2004 involved a hypothetical scenario involving blue pigmented viscose fibres found on the victim's clothing, which appeared the same as those from the putative source when compared under transmitted light, but differed markedly under UV light. Clearly these fibres did not match. Subsequent experimentation to test a theory that when the T-shirt had become wet, the fibres had 'taken up' washing detergent residues on T-shirt which contain optical brighteners causing them to fluoresce, demonstrated that this was possible. But the issue that the experiment does not address is how we tell whether the fibres on the T-shirt fluoresced the same as those from the mattress prior to the absorption of detergent. It is entirely possible that the fluorescent behaviour observed under the microscope is exactly what the fibres were like at the point of transfer. Whilst it is fair to explore the possibility that fibres have been changed at a chemical level and pursuing experiments to assess that, it would be biased for a laboratory to state that on the basis of such experiments more support is provided for the view that the fibres recovered from the T-shirt came from the mattress rather than from another source.

12.4 Mitigation strategies deployed both within the UK and overseas

The following are examples of mitigation strategies that are variously used in current practice. All are examples of good practice in appropriate circumstances and should be applied as described.

Independent checking – where only subjective observational assessments of a match are possible (for example hair comparisons, paint layer colours and sequences), full independent checking should be carried out and clearly documented. The check should be carried out independently of the original examiner.

Independent checking of analytical results – where instrumental techniques are used, either alone or to back up subjective microscopic matches, and the results are subject to interpretation by the operator (e.g., Microspectrophotometry result for analysis of colour of fibres, refractive index

measurements for glass, chemical analysis of glass fragments and paint layers), the interpretation of the results should, where possible, be carried out by two competent and experienced scientists, (operator plus one other) independently of each other.

Use of statistical approach to evaluation – to assess whether the refractive index of suspect glass fragments match that of reference glass sample(s) a statistical approach can be applied rather than relying on the experience of the practitioner.

Case Assessment and Interpretation – a robust and documented comparison of expected, pre-assessed outcomes with actual results under appropriate competing hypotheses. Some documented indication of expected outcome is recommended in all cases. Where results are at the least likely end of the expected outcomes, for example the absence of matching fibres where the most likely outcome was to find lots of matches, an independent review of the tapings would be advisable.

Training – appropriate training of practitioners in the methods employed who can demonstrate initial and ongoing competence.

Quality assurance trials – participation in internal and external quality assurance trials. Members of the ENFSI European Textile and Hair Group (ETHG) participate in an annual collaborative exercise which seeks to test various parts of the process of fibre examination. Membership of the ETHG is limited, and participation is only available to members. Forensic Science Providers (FSP) in the UK also participate in CTS (Collaborative Testing Services Inc.) trials which are available by subscription and cover fibre, paint and glass analysis. These trials are considered to be fairly basic and test the microscopic and analytical procedures employed, but do not assess the approach to evaluating the significance of the findings. At least one of the UK FSPs carrying out fibre work also carries out internal quality assurance testing with each of their scientists undertaking a mock case every 2 years to test their competency. Only some of these trials will be relevant with respect to assurance that bias is being avoided, however all provide some level of assurance of the ongoing competence of the scientists involved. There is a gap in the current system with respect to 'blind' trials – small organisations do not have the resources to conduct such testing.

Further recommendations for good practice

In addition to the good practice described in 11.4, also following may be considered:

a. Use of a completely independent ("blind") checker who repeats the examination/interpretations described in 11.4.1 and .2 but in the absence of any contextual information relating to the case. This may present practical challenges, particularly within smaller organisations. However, it will assist in a continuous learning and improvement cycle, where reporting scientists can identify instances where they may have been affected by bias. Further, it provides assurance for the courts that the interpretation is free from contextual bias.
b. Documented case assessment and interpretation in all cases involving trace evidence analysis, preferably carried out independently by a

second scientist, but at the very least to be peer reviewed. Elements of the interpretation should also be included in the scientist's statement to explain to the court how their conclusion has been reached.

c. With a reduction in the use of trace evidence analysis in casework in the UK, maintaining competency and having sufficient trained and competent staff to allow independent checks and peer reviews will be a challenge, particularly for smaller organisations. Clear documentation of case assessment, interpretation and a report/statement which clearly states the limits of the examinations used (i.e. where appropriate their subjective nature, limitations of small amounts of reference material (hairs) and whether findings and interpretation have been reviewed) should be a requirement. Such transparency and disclosure provides the opportunity for scrutiny and the identification of potential bias.
d. Where items submitted to a forensic provider for examination have been the subject of 'pre-assessment' by the submitting force, ideally a list of other items seized should be made available to the scientist on request to allow consideration of potential alternative sources of transferred material.
e. Training and education in relation to the risks of cognitive bias in trace evidence examination generally and specifically in relation to highly subjective examinations.
f. A program of 'blind' or undeclared quality assurance trials in the UK submitted to all FSPs could address the issue of bias thus providing assurance to the courts that procedures are robust and areas of potential bias are identified and managed.

13. VIDEO AND AUDIO

13.1 Introduction

A video or audio comparison often seeks to establish if the image or signal associated with a suspected crime (the "item") is of a specific article or person (the "target"). This may be for example a person's face captured on CCTV, an item of clothing being worn by the perpetrator, a vehicle or indeed any other object that may be relevant to the crime scene. This is undertaken by comparison against a reference image or signal from the target, ideally which has been generated under identical conditions to the original item. The comparison may be subjective and may utilise either purely visual side by side comparisons, or may include use of tools to aid comparison, such as overlaying of the images and switching between the two to highlight any potential differences. Alternatively comparison may be aided by objective measurements of the images (photogrammetry) for example in facial comparison in which spatial proportions of facial features are compared using measurements of distances and angles between facial landmarks in order to quantify any differences or similarities observed. Elimination should be the fundamental aim in any comparison and presence of a single difference for which there is no viable explanation should be sufficient for an exclusion. Conversely where a number of features are seen to be in common and no differences are observed, then this can provide corroboration to other evidence of inclusion.

Any examination is therefore dependent upon the visual quality and clarity of the detail that is observed by the examiner plus how inherently discriminable the object is from other objects of the same type. In combination these ultimately impact on the strength of the conclusions that may be drawn. For example with a good quality image of a motor vehicle it may be possible to identify the make and model with confidence by observing a combination of class characteristic features such as the shape of the windows, lights, bumpers, doors, overall shape etc. However, narrowing the identification to a single specific car would require much more detail in the images in order to observe individual characteristics or features that differentiate one individual car of the same make/model from another e.g. registration number, intentional alteration such as cosmetic modifications, wear and tear such as scratches or other damage features[68].

The basis for opinions and conclusions reached lies in the detection of correspondence or discordance of features determined to be reliable. These in turn rely on the individual's, relevant experience, depth of knowledge and skill as well as their disposition at the time. Every effort must be made to ensure that opinions and conclusions are logical, transparent, balanced and robust. In some cases a statistical model may be applied to provide a formal probabilistic basis for a conclusion. In other cases a statistical model may not be feasible but this does not necessarily preclude reaching a sound conclusion where for example a CAI approach is adopted.

13.2 Generic video and audio process outline

The generic forensic process that is outlined below encompasses the interpretation and reporting of video and audio comparison cases. It is applicable to a wide range of evidence types including photographic evidence with motion and still images, plus audio recordings associated with a suspected criminal act under investigation:

a. Recovery of video, photo or audio material related to the crime scene consisting
b. Items are received by the analyst along with relevant case information and questions to be addressed by the scientific work.
c. Generation of an exact copy of the original then use of techniques as required to clarify or clean up the copy of the image or audio signal
d. Examination of the copied material recovered from the crime scene and notation of features determined to be reliable
e. Examination of the 'control' item
f. Undertake a comparison using appropriate methods and equipment
g. Interpret and evaluate findings
h. Verification of result
i. Findings are described in a statement or report.
j. The scientist may be called to court to give oral testimony.

---

[68] Scientific Working Group Imaging Technology (SWGIT) (2013) Best practices for forensic photographic comparison V1.1 Section 16

13.3 Risks of cognitive bias

Within video and audio comparison, there is a spectrum of bias risk:

Risk factor	Low risk	High risk
Detail & Presentation	The images/signals are clear detailed and unambiguous with item and reference images generated under identical conditions	The images are of poor quality and the detail present is poorly defined, and the images being compared have been generated under very different conditions
Equipment	Optimum visualisation of the detail in an image using appropriate equipment/imaging and enhancement techniques.	Poor or inappropriate equipment/imaging and enhancement techniques.
Approach	There is a methodical approach with defined standards built on principles that have been tested and validated. Item is characterized prior to exposure to reference image	When the approach is un-researched, ad hoc and personal to the operator. Item is characterized after exposure to reference image
Scientist/Examiner	Scientist/examiners are well trained, experienced and continuously meet acceptable standards of competence	Scientist/examiners are inexperienced, unmonitored and left to adopt their own approach.
Verification of results	Independent review of critical findings	There is no independent review, or reviewer knows findings and conclusions drawn from original assessment

Table 4: Spectrum of bias risk in video and audio comparison

13.4 Mitigation strategies and good practice guidance

Avoiding psychological contamination in the processing of material

One of the greatest risks of introducing cognitive bias is in the way the material is provided for assessment. Examiners should only be provided with the information relevant to the examination of the item image, and in the first instance and they should only be asked to describe what they see. The latter guards against confirmation bias, which is almost inevitable if the question asked is along the lines of "do you agree that this is item/individual x?", or the examiner asks to be told what the item is so that they can consider whether or not they agree. Not being provided with the case notes and other extraneous information prior to the examination and comparison task at hand helps safeguard against contextual bias. For the same reason it is better for the

analyst to receive written briefing regarding the comparison to be made rather than being in direct verbal contact with the investigator, so that opportunity for transfer of non-relevant and potentially biasing information (both contextual and confirmatory) can be avoided.

Wherever possible, the item should be assessed prior to observing the reference image or signal, again so that confirmation bias can be guarded against. If a series of images are submitted of what is believed to be the same item, these should be assessed in sequence starting with the worst image first, so that the potential for confirmation bias between these images is avoided. Where a discriminatory feature is identified in the item only after comparison with the reference, this should be fully explained in the examination records, so that transparency of the assessment is maintained at all times.

Independent assessment of critical findings is also crucial. Independent checking that minimizes the risk of cognitive bias entails assessment without knowing the outcome of the initial analysis, or even where possible the identity of the original examiner in order to avoid confirmation bias.

Use of validated processes

All forensic processes should be validated prior to use in casework. Section 20 of the FSR Codes provides guidance on validation with more detailed explanations given in validation appendix currently due for publication by the FSR in September 2014 plus guidance on how to approach validation of digital forensic techniques in an currently being drafted for consultation by the FSR. Scientific validation is the process by which a new method or technique is assessed to ensure that it is fit for purpose and that once implemented will continue to function as such. This principle applies whether a system provides objective highly automated analysis and comparison of materials, or at the other extreme where the process relies almost entirely on subjective comparison and assessment by an analyst.

Bias is less likely when images are clear and well defined, whilst the risk of bias increases as images become less defined and ambiguity regarding interpretation increases. Therefore use of appropriate and validated methods to clarify images/signals may help reduce risk of bias. However certain techniques for image manipulation are "lossy" and can result in the loss of potentially discriminable detail (increasing the risk of false inclusion) whilst other enhancement techniques can create artefacts, thereby increasing the risk of false exclusion. It is crucial therefore that any manipulation processes are validated. This should include full characterization of the processes applied including determination of the limits within which the application can be reliably used and demonstration through experimentation not to increase the risk of false inclusion or exclusion. Likewise during application to casework, and especially in the enhancement of audio signals the analyst should frequently check back during processing against the original to ensure that the signal has

not become over-processed[69]. Likewise, when using colour as a comparator, the limitations of the approach should be fully evaluated and understood: under certain lighting conditions (e.g. sodium lamp), 2 items that are different in colour under natural illumination may appear to be the same, whilst the same item under different lighting conditions may appear to be markedly different in colour.

Techniques deployed to aid in the side by side comparison of images must be validated to ensure they do not introduce bias. For example overlaying techniques for comparison can highlight differences between images by rapid flicking between images. However a gradual transition between two overlaid images may cognitively mask any differences from the observer. Wherever possible the same context should be used to generate reference images for comparison against the original crime scene image by for example re-constructing the scene and capturing the reference image using the same equipment, lighting conditions, camera angles, environmental conditions etc. Where this is not possible, the resultant limitations in making a comparison should be declared in any statement.

Proficiency testing/ QC measures

The fact that the police have asked for a comparison to be made between two images or an image and an item can in itself create a bias towards confirmation. The use of appropriate procedures, plus the training, experience and competence of the examiner should in combination ensure that in this is being safeguarded against in practice, but these measures should be both strengthened by and demonstrated to be effective through the use of effective QA/QC measures. These measures include the following:

Initial competency assessment of an individual prior to commencing forensic casework: the individual is subjected to proficiency testing using characterized test material of known provenance to demonstrate that they, in combination with validated working practices, generate reliable unbiased outcomes.

Ongoing competency assessment through use of declared and undeclared trials. Undeclared or blind trials are of particular value as these are more likely to give a truer indication of typical performance and behaviours, unlike a declared trial where the individual knows that they are being observed, and may consequently behave differently to normal by for example being more cautious in their evaluation.

Provision of an image line up using "fillers". This is akin to an identity parade in which for example the analyst may be presented with a number of images comprising that of the target plus a number of other broadly similar "innocent" items, and asked to determine which if any constitutes a match to the image corresponding to the crime scene[70]. A further refinement is to split this

---

[69] Manchester, P. (2010) An introduction to forensic audio. Sound on Sound. January 2010 http://soundonsound.com/sos/jan10/articles/forensics.html

[70] Kassin, et al (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. Journal of Applied Research in Memory and Cognition. 2, p42-52

comparison into two sets so that the examiner does not know whether an individual set contains the target image.

14. ABBREVIATIONS

ACE-V	Analysis, Comparison, Evaluation and Verification
FBI	Federal Bureau of Investigation
ENFSI	European Network of forensic Science Providers
ETHG	European Textile and Hair Group
FSP	Forensic science provider
Hd	Defence hypothesis
Hp	Prosecution hypothesis
LR	Likelihood Ratio
OCF	Open Case File

15. ACKNOWLEDGEMENTS

This draft appendix was produced following a competitive tender, by Kevin Sullivan (Principal Forensic Services Ltd.

The author would like to acknowledge input from the following in the formulation of this guidance: M. Cass, TRL; K. Aduse-Poku, Teeside University

Science and Justice xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Science and Justice

journal homepage: www.elsevier.com/locate/scijus

Editorial

Research focused mainly on bias will paralyse forensic science

1. Introduction

There is now a body of research that has reinforced what many (including forensic scientists) had experienced before: decision making in forensic science is not immune to bias. Confirmation bias, associated with the potential adverse impact of contextual case information, has a prevalent position among the potential cognitive influences. It has been the object of research (for a recent overview, please refer to the feature article by Kassin et al. [1] and the follow-up commentaries [2–12]) and figures in the top priorities of many organisations. For example, the theme of bias received top research priorities in the 2009 NAS report recommendation 5 [13].

Forensic journals have also put bias at the forefront of their publication agenda and we can observe a constant feed of papers testifying to various degrees on the risk of having forensic scientists' judgement tainted by inadequate bias. The trend goes also across all forensic disciplines as attested by papers published since 2011 on fingerprints [14–16], DNA [17], anthropology [18], handwriting [19,20] or odontology [21].

I do not want to minimize the importance of the above and how it contributes to a better management of forensic science, but should research remain focused on processes, or should we not move on to the basic understanding of the forensic traces?

I can foresee the following risks of being focused on bias only:

(a) The risk of enforcing the view that the forensic scientists should be detached, blind and immune from any external influences (especially from the inquiry).
(b) The risk of enforcing the view that forensic experts can continue to operate as "black boxes" provided they work according to regulated standard operating procedures, designed to cure for bias and that estimates of the error rates associated with their decisions are disclosed. A corollary is the risk to ignore the needed requirement to develop fundamental research in areas dominated by decision-making processes based largely on human perception and skilled judgement.

I view both of the above risks as major obstacles to what forensic science could offer to the criminal judicial system. Let me give you a few personal arguments. I am conscious that they may provoke reactions and, yes, these opinions involve judgement and as such could be considered as biased!

2. The risk of a "blind" forensic scientist

Forensic science laboratories are moving quickly into becoming providers of service commodities: they receive pre-processed samples and are just asked to apply a given analytical technique to the content of these test tubes. The forensic work is segmented without any encouragement towards an integrated approach. The mechanisms offered to mitigate contextual bias just validate such a vision of forensic science: it is proposed to blind examiners from any domain-irrelevant information or to adopt sequential unmasking procedures. The whole enterprise is driven by a risk adverse strategy focused on the micro-management of detected errors. Each new error will be more expensive to fix, will bring its new sets of procedures, and will hinder any development. Opportunities will systematically be first vetted against the risk of bias (even the unconscious ones) and then only its potential to improve policing and the criminal justice system. Is this the future of forensic science? Is this how forensic scientist wants to contribute? What about the 'science'?

As noted by Dror [22], an appropriate balance needs to be found between the risks and benefits. However, at the moment, I do not see any sign that the forensic community has found this appropriate balance.

For example, good practice of case assessment and interpretation [23] invites the forensic scientist to inquire about the needs of the case (beyond the police request) and to obtain contextual elements in order to help formulate propositions against which the forensic findings will be assessed. This step requires obtaining information regarding the activities alleged by the parties (hence requirement for some contextual information). To prevent that risky exchange of information, Risinger [24] urged forensic scientists to deal exclusively with source level issues and leave the rest to the factfinder. Defaulting to source level issues on the ground that case information should not be disclosed to the forensic scientist is very dangerous in my opinion. I recently expressed that view in relation to the interpretation of small quantity of trace DNA [25]: there is a risk with leaving the presence of DNA to be assessed by others, left to advocacy, when the scientist can bring decisive knowledge including highlighting how complex the task may be. This discussion is not new and let me clarify that the risk I am referring to here is the risk for potential miscarriage of justice due, partly or fully, to the strict and blind segmentation between the forensic scientist and the investigation. As Roberts and Willmore already put it [26, p. 137] in 1993: "Our research suggests that the superficially attractive objective of shielding the forensic scientist from information which might inappropriately influence her scientific judgment should be abandoned in favour of more productive efforts to improve the extent and quality of the information exchange between FSS scientists and instructing lawyers."

I observed another worrisome trend when commentators looked into the broader investigative usage of forensic science. It has been rightly noticed by Laurin [27] that the 2009 NAS report did not gave any in-depth consideration of the use of forensic science as a police intelligence and strategic tool. Indeed the NAS report proceeded under a very narrow view of a laboratory providing services to generate forensic findings to be used potentially in a court of law. The work of my colleagues in this investigative and crime analysis area [28–30] and

http://dx.doi.org/10.1016/j.scijus.2014.02.004

translated into operational benefits by various police forces in Switzerland testifies to the fundamental merit of the approach. A systematic collection of forensic traces (e.g. biological traces, footwear marks, earmarks, tool marks) allows connecting apparently unrelated cases. When this information is structured in the context of time, geography, types of targets and modus operandi, it successfully allows identifying and following criminal phenomena. But when Cole [31] responded to Laurin [27], he described the approach as a nostalgic and "unabashed attempt to recapture a lost vision of both forensic science and scientific policing." The risk of bias is raised again and it is posited as a matter of principle that there is a requirement to separate forensic science from investigation. But the perceived risks are only postulated and have never been measured. The bias is becoming the attractive swiping argument to legitimate a paralysed vision of a detached forensic laboratory working in silos, even between forensic scientists, to avoid any exchange of knowledge.

3. The risk of the "black box" expert

Research on bias promotes a view of a forensic scientist delivering decisions on the issue, most of the time with yes/no decisions regarding for example the source of the examined items. It perpetuates a status quo of the forensic examiners empowered to make decisions. In that paradigm, the experts (through training and experience) have acquired a status of adjudicator by delegation of the court and we just want to monitor/calibrate them. The efforts towards an understanding of how they make decisions become secondary because the system is satisfied that experts can come to the correct decisions under controlled conditions.

Biedermann et al. [32] make a strong case for the use of probabilistic statements in the forensic identification disciplines, rather than stating blunt (or apparent) certainties. They rightly insisted on the probabilistic nature of the endeavour. But despite some calling for a change of culture [33] or reporting practice [34], the dominant view is for experts to keep reporting opinions amounting to a factual establishment of sources. I do not understand why we are so far from an application of likelihood ratio associated with fingerprint evidence. The recent paper by Neumann et al. [35] gave the perfect signal for a development but unfortunately (and partly due to the closure of the Forensic Science Service), instead of pursuing, I sense that all future efforts will concentrate on measuring experts' performance and not in changing how they interpret and report their findings.

Measuring error rates from experts will provide needed indicators for quality but I can hardly see this as the panacea and it may even serve as a proxy for more fundamental research on the forensic trace itself. Take the most recent study by Ulery et al. [36], the reported rate of false positive is 0.01%. When presented in court with a decision of identification, the weight associated with the decision will not be measured against that rate. The rate will just serve as initial pass criterion. Does the court trust the discipline and its practitioners? If nothing indicates that the testifying examiner deviates from the practice espoused by the experts tested by Ulery et al., we can predict that the testimony will be trusted. By trusted, we mean an absolute confidence on the strength of the conclusion. In other words an expression of a likelihood ratio that is so high in favour of a common source that the chance of an error is considered as so small as to be dismissed. The problem here is that there is no appropriate weighing of the contribution of the forensic findings. Only structured and systematic research on the features themselves (and not of the experts' decisions) can lead to such a state. Procedures guarding against bias and measurement of experts' error rates will only provide *satisfecit* allowing courts to trust the expert's opinion. But that opinion will remain being delivered *ipse dixit*. That process offers no mechanism to effectively measure the actual weight to be attached to the forensic findings.

Research favouring a systematic acquisition of data associated with the features used holistically by experts should be at the forefront of the agenda. The research should not be designed to validate practice that prevailed for years in a given area, but to support a more fundamental change in the way forensic evidence is delivered in court.

To put some context on the above argument I will use an example outside the usually discussed forensic disciplines. The provision of evidence based on the examination of earmarks and earprints is still in its infancy [37]. We could choose two research strategies:

(1) Establish a community of experts, train them to the task, develop examination protocols that will limit bias and measure their decision performance. Regardless on the training efforts, I can predict that the false positive rate in their decisions will be about 1%. Will the court be able to handle an identification decision delivered by an expert qualified with a 1% error rate? Especially when informed that earmarks may vary drastically in their quality and that when the information from the mark is limited the probability of an error could reach more than 20%.
(2) Measure systematically the earmark/earprint features on adequate samples, acquire new knowledge, and strive to assign a likelihood ratio to a comparison between a mark and a print. Recent research [38] has shown that on the average likelihood ratio when comparing marks and prints from the same source is of the order of 10^3. And needless to say that in a given case, the case-specific likelihood ratio (based on the intrinsic merit of the mark) will be quoted.

In my opinion, the second option should be on the top of the research agenda. And I am happy to generalise this to all forensic domains where currently a full holistic expert-based approach is used (such as fingerprints, handwriting, tool marks and firearms, bite marks or footwear marks).

To conclude, I argue that we should move away from the "black box" approach and study more deeply, in a systematic approach, the forensic traces themselves.

References

[1] S.M. Kassin, I.E. Dror, J. Kukucka, The forensic confirmation bias: problems, perspectives, and proposed solutions, J. Appl. Res. Mem. Cogn. 2 (2013) 42–52.
[2] L. Butt, The forensic confirmation bias: problems, perspectives, and proposed solutions — commentary by a forensic examiner, J. Appl. Res. Mem. Cogn. 2 (2013) 59–60.
[3] D. Charlton, Standards to avoid bias in fingerprint examination? Are such standards doomed to be based on fiscal expediency? J. Appl. Res. Mem. Cogn. 2 (2013) 71–72.
[4] S.D. Charman, The forensic confirmation bias: a problem of evidence integration, not just evidence evaluation, J. Appl. Res. Mem. Cogn. 2 (2013) 56–58.
[5] S.A. Cole, Implementing counter-measures against confirmation bias in forensic science, J. Appl. Res. Mem. Cogn. 2 (2013) 61–62.
[6] I.E. Dror, S.M. Kassin, J. Kukucka, New application of psychology to law: improving forensic evidence and expert witness contributions, J. Appl. Res. Mem. Cogn. 2 (2013) 78–81.
[7] E. Elaad, Psychological contamination in forensic decisions, J. Appl. Res. Mem. Cogn. 2 (2013) 76–77.
[8] B.L. Garrett, Blinded criminal justice, J. Appl. Res. Mem. Cogn. 2 (2013) 73–75.
[9] R.N. Haber, L. Haber, The culture of science: bias and forensic evidence, J. Appl. Res. Mem. Cogn. 2 (2013) 65–67.
[10] R. Heyer, C. Semmler, Forensic confirmation bias: the case of facial image comparison, J. Appl. Res. Mem. Cogn. 2 (2013) 68–70.
[11] M. Triplett, Errors in forensics: cause(s) and solutions, J. Appl. Res. Mem. Cogn. 2 (2013) 63–64.
[12] G.L. Wells, M.M. Wilford, L. Smalarz, Forensic science testing: the forensic filler-control method for controlling contextual bias, estimating error rates, and calibrating analysts' reports, J. Appl. Res. Mem. Cogn. 2 (2013) 53–55.
[13] National Research Council, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington, D.C., 2009
[14] P.A.F. Fraser-Mackensie, I.E. Dror, K. Wertheim, Cognitive and contextual influences in determination of latent fingerprint suitability for identification judgments, Sci. Justice 53 (2013) 144–153.
[15] I.E. Dror, K. Wertheim, P. Fraser-Mackenzie, J. Walajtys, The impact of human–technology cooperation and distributed cognition in forensic science: biasing effects of AFIS contextual information on human experts, J. Forensic Sci. 57 (2012) 343–352.
[16] I.E. Dror, C. Champod, G. Langenburg, D. Charlton, H. Hunt, R. Rosenthal, Cognitive issues in fingerprint analysis: inter- and intra-expert consistency and the effect of a 'target' comparison, Forensic Sci. Int. 208 (2011) 10–17.
[17] I.E. Dror, G. Hampikian, Subjectivity and bias in forensic DNA mixture interpretation, Sci. Justice 51 (2011) 204–208.

Please cite this article as: C. Champod, Research focused mainly on bias will paralyse forensic science, Sci. Justice (2014), http://dx.doi.org/10.1016/j.scijus.2014.02.004

[18] S. Nakhaeizadeh, I.E. Dror, R.M. Morgan, Cognitive bias in forensic anthropology: visual assessment of skeletal remains is susceptible to confirmation bias, Sci. Justice (2014) in press, http://dx.doi.org/10.1016/j.scijus.2013.11.003.
[19] R.D. Stoel, I.E. Dror, L.S. Miller, Bias among forensic document examiners: still a need for procedural changes, Aust. J. Forensic Sci. 46 (2014) 91–97.
[20] B. Found, J. Ganas, The management of domain irrelevant context information in forensic handwriting examination casework, Sci. Justice 53 (2013) 154–158.
[21] M. Page, J. Taylor, M. Blenkin, Context effects and observer bias—implications for forensic odontology, J. Forensic Sci. 57 (2012) 108–112.
[22] I. Dror, Letter to the Editor—combating bias: the next step in fighting cognitive and psychological contamination, J. Forensic Sci. 57 (2012) 276–277.
[23] G. Jackson, C. Aitken, P. Roberts, Case Assessment and Interpretation of Expert Evidence: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses, The Royal Statistical Society, London, 2013.
[24] D.M. Risinger, Reservations about likelihood ratios (and some other aspects of forensic 'Bayesianism'), Law Probab. Risk 12 (2013) 63–73.
[25] C. Champod, DNA transfer: informed judgment or mere guesswork? Front. Genet. 4 (2013)http://dx.doi.org/10.3389/fgene.2013.00300.
[26] P. Roberts, C. Willmore, The Role of Forensic Evidence in Criminal Proceedings, Royal Commission on Criminal Justice Research Study, No. 11, HMSO, London, 1993.
[27] J.E. Laurin, Remapping the path forward: toward a systemic view of forensic science reform and oversight, Tex. Law Rev. 91 (2013) 1050–1118.
[28] O. Ribaux, P. Margot, R. Julian, S.F. Kelty, Forensic intelligence, in: J.A. Siegel, P.J. Saukko, M.M. Houck (Eds.), Encyclopedia of Forensic Sciences, Academic Press, Waltham, 2013, pp. 298–302.
[29] O. Ribaux, A. Baylon, C. Roux, O. Delémont, E. Lock, C. Zingg, P. Margot, Intelligence-led crime scene processing. Part I: forensic intelligence, Forensic Sci. Int. 195 (2010) 10–16.
[30] O. Ribaux, A. Baylon, E. Lock, O. Delémont, C. Roux, C. Zingg, P. Margot, Intelligence-led crime scene processing. Part II: intelligence and crime scene examination, Forensic Sci. Int. 199 (2010) 63–71.
[31] S.A. Cole, Response — forensic science reform: out of the laboratory and into the crime scene, Tex. Law Rev. 91 (2013) 123–136.
[32] A. Biedermann, P. Garbolino, F. Taroni, The subjectivist interpretation of probability and the problem of individualisation in forensic science, Sci. Justice 53 (2013) 192–200.
[33] M. Page, J. Taylor, M. Blenkin, Uniqueness in the forensic identification sciences—fact or fiction? Forensic Sci. Int. 206 (2011) 12–18.
[34] C. Champod, Fingerprint examination: towards more transparency, Law Probab. Risk 7 (2008) 111–118.
[35] C. Neumann, I.W. Evett, J. Skerrett, Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm, J. R. Stat. Soc. 175 (2012) 371–415 (with discussion).
[36] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Accuracy and reliability of forensic latent fingerprint decisions, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 7733–7738.
[37] C. Champod, I. Evett, B. Kuchler, Earmarks as evidence: a critical review, J. Forensic Sci. 46 (2001) 1275–1284.
[38] S. Junod, J. Pasquier, C. Champod, The development of an automatic recognition system for earmark and earprint comparisons, Forensic Sci. Int. 222 (2012) 170–178.

Christophe Champod
School of Criminal Justice, Forensic Science Institute, University of Lausanne, Switzerland
E-mail address: christophe.champod@unil.ch.

Available online xxxx

Please cite this article as: C. Champod, Research focused mainly on bias will paralyse forensic science, Sci. Justice (2014), http://dx.doi.org/10.1016/j.scijus.2014.02.004

Science and Justice xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Science and Justice

journal homepage: www.elsevier.com/locate/scijus

Letter to the editor

Regarding Champod, editorial: "Research focused mainly on bias will paralyse forensic science"

Dear Dr. Barron,

In 2009, a report of the (US) National Research Council declared that "[t]he forensic science disciplines are just beginning to become aware of contextual bias and the dangers it poses" [1]. The report called for additional research and discussion of how best to address this problem. Since that time, the literature on the topic of contextual bias in forensic science has begun to expand, and some laboratories are beginning to change procedures to address the problem. In his recent editorial in Science and Justice, Christophe Champod suggests that this trend has gone too far and threatens to "paralyse forensic science" [2]. We think his arguments are significantly overstated and deserve forceful refutation, lest they stand in the way of meaningful progress on this important issue.

Dr. Champod opens by acknowledging that forensic scientists are vulnerable to bias. He says that he does not "want to minimize the importance of [research on this issue] and how it contributes to a better management of forensic science…" He continues by asking "…but should research remain focused on processes, or should we not move on to the basic understanding of the forensic traces?" He then comments on risks of "being focused on bias only." By framing the matter in this way, Dr. Champod creates a false dichotomy, and implies facts about the current state of funding and research that are simply not the case. He seems to be saying that currently all or most research funding and publication is directed towards problems of bias, and little or none towards "basic understanding of the forensic traces." Dr. Champod should know that this is not the case, however, since (among other things) he is a co-author of a marvellous recently-released empirical study on fingerprint analysis funded by the (US) National Institute of Justice [3]. Any perusal of NIJ grants, or the contents of leading forensic science journals, would not support Dr. Champod's apparent view of the current research world.

It would of course be a mistake for all of the available funding for research on forensic science topics to be devoted to the potential effects of bias, but again, this neither is the case currently nor is it in our opinion likely to become the case in the future. To discuss the risks of focusing "on bias only," is simply raising a straw man when no one, not even the most ardent supporter of sequential unmasking or other approaches to the control of biassing information in forensic science practise, suggests focusing research "on bias only."

That said, we do believe that the research record both in forensic science and in a variety of other scientific areas has reached a point that clearly establishes the pressing need for all forensic areas to address the problem of contextual bias. As Andrew Rennison, who was then the forensic science regulator for England and Wales, told the plenary session of the American Academy of Forensic Science in February, "we don't need more research on this issue, what we need is action." This is not to say that further research on bias and its effects is not valuable, and should not be appropriately supported, but merely that it is not required as a precursor to taking steps to control the pernicious effects of biassing information in practise.

Dr. Champod argues against taking such steps, however, claiming that bias reduction efforts create two risks. First, there is the "risk of the blind forensic scientist," which he explains as: "[t]he risk of enforcing the view that the forensic scientists should be detached, blind and immune from any external influences (especially from the inquiry)." In essence, he is concerned that forensic scientists will be isolated from investigators in ways that undermine their effectiveness in supporting criminal investigations. But his argument rests on the incorrect assumption that forensic scientists must choose to play only one of two possible roles — if they remain detached and blind (in order to insulate themselves from "external influences") then they cannot play the broader advisory role that Dr. Champod views as vital for effective investigations.

While Dr. Champod is correct that in a given case the two roles cannot be played by the same person, he fails to acknowledge the obvious response that *the two roles need not be played by the same person*. For example, it has been suggested that different forensic scientists in the individual case be assigned to two different roles: case managers and analysts [4–6]. Case managers would participate in investigations in the manner that Dr. Champod contemplates but would not conduct or interpret examinations themselves. Instead, they would screen the information that is passed to colleagues (analysts) who could thereby remain blind to potentially biassing contextual information while conducting examinations and issuing laboratory reports. A given forensic scientist could be a case manager, or an analyst, or could alternate between those roles (from case to case). We have argued on a number of occasions that separating functions in this manner would largely eliminate the "risk" that Dr. Champod associates with blinding procedures [4–6]. We are perplexed at his failure to address this key point in his editorial.

As Dr. Champod properly notes, there are two broad contexts in which questions can arise concerning what forensic scientists should know in order to do the job assigned to them: contexts in which the expert's conclusions may be used in court, and contexts (such as more generalised intelligence work) where the conclusions generated are unlikely to be so used. The latter is often the case, for example, in regard to computer forensics applications.

In the latter setting, it should be up to the investigating agency to determine the extent to which they want to turn their forensic experts into all-source experts (general detectives with an expertise component, if you will). In such cases there would be no direct implication

http://dx.doi.org/10.1016/j.scijus.2014.06.002

Please cite this article as: D.M. Risinger, et al., Regarding Champod, editorial: "Research focused mainly on bias will paralyse forensic science", Sci. Justice (2014), http://dx.doi.org/10.1016/j.scijus.2014.06.002

for persons charged in a criminal proceeding, assuming the two contexts can be kept sufficiently separate. But it would be wise for whoever is leading such an intelligence operation to realise that using forensic scientists in this way might undermine the reliability of the domain-specific conclusions reached, thus impairing their utility in the more general inquiry.

In the context of any forensic science application where the results will be used as evidence in a legal case, however, and most certainly against a defendant in a criminal case, the agency or laboratory responsible for the results as evidence is no longer free to make its own decisions about the costs and benefits of structuring the process one way or another. Opinions that are influenced by contextual information not relevant to the analyst's forensic expertise invade the province of the factfinder, and run the risk of factfinder confusion as to the scope of the forensic science expertise involved, and of double counting the domain-irrelevant information — counting it once as part of the hidden basis for the "expert" conclusion, and again by direct evaluation by the factfinder. In this context, the risk of error falls most heavily on the criminal defendant, and error reduction is a paramount concern. It is this focus that was properly the focus of the NRC report, and properly the focus of various calls for masking protocols to eliminate or control the effects of biassing information.

No one who has called for such bias reduction measures has sought to deprive forensic scientists of any information relevant to the exercise of their expertise. Indeed, the leading framework for control of biassing information, "sequential unmasking," explicitly builds into its two-stage process a filtration of domain-irrelevant information coupled with the release of domain-relevant information with the potential to induce bias in the least biassing order consistent with maximal accuracy [7]. Nor does this approach deprive law enforcement of investigatory guidance informed by forensic expertise. The control officer who does the filtration is also the interface with the "client" (usually law enforcement, but sometimes the defence), and can freely perform this function. But the forensic scientist doing the characterisation and interpretation of the evidence in the individual case gives maximally accurate results concerning case-specific issues within their expert domain based only on domain relevant information released in the least biassing order. Forensic scientists owe the criminal justice system no less.

Dr. Champod also identifies a second risk, which he dubs "the risk of the black box expert." His concern, in essence, is that efforts to address contextual bias will somehow interfere with the efforts of forensic scientists to develop empirically-based match criteria that can be applied more objectively. In our view, this second "risk" is no risk at all. No one who calls for bias controls is in favour of using bias controls as an excuse not to improve the objectivity and diagnostic value of forensic science methods, or of depriving such efforts of funding. In fact, some of the leading exponents of sequential unmasking were present at the Royal Statistical Society when Cedric Neumann's foundational paper (co-authored with Evett and Skerrett) [8] on improvements in fingerprint methodology was read, and they published highly laudatory commentary upon it [9]. Ultimately research such as that, and the recent extension of it referenced above [3] co-authored with Neumann by Dr. Champod himself (along with Yoo, Gennesay and Langenburg) might someday in the distant future bring fingerprint identification to a point of such mathematised and mechanised perfection that the potential for contextual bias would be trivial. But in the here and now, fingerprint examination is not there yet, and none of the other pattern-matching disciplines are even close. Until then our choices are either to abandon such expertise wholesale (which is not going to happen, nor should it), or to do what we can to insure that their products proffered as evidence are as valid as possible. Protocols to control biassing information are necessary for that, and will remain necessary for the foreseeable future. It is time for every current area of forensics to require the adoption of such standards. It would be hugely unfortunate if Dr. Champod's editorial became an excuse not to do so.

References

[1] National Research Council, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington, DC, 2009.

[2] C. Champod, Research focused mainly on bias will paralyse forensic science, Sci. Justice (2014)http://dx.doi.org/10.1016/S.scijus.2014.02.004.

[3] C. Neumann, C. Champod, M. Yoo, T. Genessay, G. Langenburg, Improving the Understanding and the Reliability of the Concept of "Sufficiency" in Friction Ridge Examination, NIJ Publication Update, 2014.

[4] W. Thompson, What role should investigative facts play in the evaluation of scientific evidence, Aust. J. Forensic Sci. 43 (2011) 123–134.

[5] W. Thompson, S. Ford, J. Gilder, K. Inman, A. Jamieson, R. Koppl, et al., Commentary on: Thornton—a rejection of 'working blind' as a cure for contextual bias, J. Forensic Sci. 56 (2011) 562–563.

[6] D.M. Risinger, M. Saks, W. Thompson, R. Rosenthal, The Daubert/Kumho implications of observer effects in forensic science: hidden problems of expectation and suggestion, Calif. Law Rev. 90 (2002) 1–55.

[7] D. Krane, S. Ford, J. Gilder, K. Inman, A. Jamieson, R. Koppl, et al., Sequential unmasking: a means of minimizing observer effects in forensic DNA interpretation, J. Forensic Sci. 53 (4) (2008) 1006–1007.

[8] C. Neumann, I.W. Evett, J. Skerrett, Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm, J. R. Stat. Soc. 175 (2012) 371–415 (with discussion).

[9] D.M. Risinger, Comment on Neumann et al., Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm, J. R. Stat. Soc. Ser. A 175 (2012) 398;
S.A. Cole, Comment on Neumann et al., Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm, J. R. Stat. Soc. Ser. A 175 (2012) 399;
M.J. Saks, J.M. Votruba, Comment on Neumann et al., Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm, J. R. Stat. Soc. Ser. A 175 (2012) 408;
W.C. Thompson, Comment on Neumann et al., Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm, J. R. Stat. Soc. Ser. A 175 (2012) 409.

D. Michael Risinger
Seton Hall University School of Law, USA
Corresponding author at: Seton Hall University School of Law, One Newark Center, Newark, NJ 07102, USA.
E-mail address: risingmi@shu.edu.

William C. Thompson
Department of Psychology and Social Behavior, University of California, Irvine, USA

Allan Jamieson
The Forensic Institute, Glasgow, Scotland

Roger Koppl
Whitman School of Management and Forensic and National Security Sciences Institute, Syracuse University, USA

Irving Kornfield
University of Maine, USA

Dan Krane
Wright State University, USA

Jennifer L. Mnookin
Program on Understanding Law, Science and Evidence (PULSE), UCLA School of Law, USA

Robert Rosenthal
Emeritus Harvard University, University of California, Riverside, USA

Michael J. Saks
Center for Law, Science & Innovation, Arizona State University, USA

Sandy L. Zabell
Northwestern University, USA

18 April 2014
Available online xxxx

Please cite this article as: D.M. Risinger, et al., Regarding Champod, editorial: "Research focused mainly on bias will paralyse forensic science", Sci. Justice (2014), http://dx.doi.org/10.1016/j.scijus.2014.06.002

Science and Justice xxx (2014) xxx

Contents lists available at ScienceDirect

Science and Justice

journal homepage: www.elsevier.com/locate/scijus

Reply to letter to the editor

Letter to editor in response to editorial by Risinger et al.

Keywords:
Interpretation
Bias
Expert judgment
Context effects

Dear Sir,

The person of interest, S, has a certain property associated with him signified by a colour. In this case S is red. Other people are blue, green etc. There are also other red people in the world. When a person commits a crime there is a chance they will leave their colour behind. In this particular case the crime was committed in a very dark room. A photo of a mark in that room is taken and sent to the laboratory for colour determination. A sample of S's colour is also submitted.

In the photograph the mark can be seen. Its colour is barely visible but it is submitted to a specialist colour analyst for an expert opinion. Should the analyst be told that S is red before she makes her colour assessment? No. Should a photographer be sent back to the scene to rephotograph the mark under acceptable lighting conditions? Yes.

This analogy illustrates one of the points made in the thoughtful article by Champod [1] when he called for renewed energy in the study of traces. We agree with Champod. Neither ourselves nor Champod deny the existence of context effects. Much, but not all, of the work on context effects has gone to showing the existence of such effects. We agree that study on these should continue not because the case needs proving, in our view the case is proven, but because widespread acceptance and action are lacking. We applaud both the sequential unmasking concept of Krane et al. [2] and the context management approach of Found [3]. But, like Champod, we would also greatly welcome improvement in the examination of traces.

Let us see what Champod actually recommends. He recommends the development of systematic ways to measure and characterise the features on a given sample, the acquisition of new knowledge, and the development of methods to assign a likelihood ratio to a comparison. We cannot agree more. How then is it that Risinger et al. [4] find fault? There must be some misunderstanding happening. Risinger et al. appear to misread the Champod article as arguing against methods to counter context effects. They further argue that there is ample research in the areas for which Champod makes a call although they do agree that this work is worthwhile. We feel that there is still a lot to do in these areas. What substantial improvements have been made in the probabilistic assessment of toolmarks, footwear impressions, or fingerprints in the last decade? Progress is not zero but equally these sciences have not been revolutionised. We have seen resistance to probabilistic intrusion into these fields [5,6]. The status quo in these fields is that well trained analysts compare impressions often side by side or in overlay and then make a subjective judgement on the value of the correspondence.

This is an appropriate point for us to remark that we have no difficulty with the notion of subjective judgement. Whereas we accept the need for objectivity in the sense of a judgement uninfluenced by irrelevant context effect. The notion that in any situation there is an assessment of evidential weight that is objective in the sense of being "real" and independent of human judgement is a myth.

This judgement is likely to be structured within a construct designed to improve the reliability of the opinion. One such commonly applied is termed ACE-V which stands for analysis, comparison, evaluation and verification. These experts are most likely to be working in an environment without explicit context management. We can think of the expert and peer-reviewer and laboratory system as if it were some instrument. We put in the evidence at one end and out the other comes an opinion. We can measure the performance of the instrument under known conditions. We can all agree that it is advantageous to remove any biasing influences from this instrument. But Champod is calling for additional effort as well as this removing of biasing information. He is calling for fundamental scientific endeavours that improve the instrument. We agree.

References

[1] C. Champod, Research focused mainly on bias will paralyse forensic science, Sci. Justice 52 (2014) 107–109.
[2] S.F. Krane, J. Gilder, K. Inman, A. Jamieson, R. Koppl, Sequential unmasking: a means of minimizing observer effects in forensic DNA interpretation, J. Forensic Sci. 53 (2008) 1006–1007.
[3] Found B, Ganas J. The management of domain irrelevant context information in forensic handwriting examination casework. Science & Justice.
[4] Risinger DM, Thompson WC, Jamieson A, Koppl R, Kornfield I, Krane D, et al. Regarding Champod, editorial: "Research focused mainly on bias will paralyse forensic science". Science & Justice.
[5] R v T. Neutral Citation Number: [2010] EWCA Crim 2439: Court of Appeal, 2010.
[6] W.B. Bodziak, Traditional conclusions in footwear examinations versus the use of the Bayesian approach and likelihood ratio: a review of a recent UK appellate court decision, Law Probab. Risk 11 (4) (2012) 279–287.

John Buckleton*
ESR, Auckland, New Zealand
* Corresponding author.
E-mail address: john.buckleton@esr.cri.nz

Ian Evett
Principal Forensic Services Ltd, London, United Kingdom

Bruce Weir
Department of Biostatistics, University of Washington, Seattle, WA 98195-7232, USA

25 July 2014
Available online xxxx

http://dx.doi.org/10.1016/j.scijus.2014.07.003

Forensic Science Policy & Management: An International Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ufpm20

A Report of Statistics from Latent Print Casework

Glenn Langenburg[a], Flore Bochet[b] & Scott Ford[c]
[a] Minnesota Bureau of Criminal Apprehension, St. Paul, Minnesota
[b] Brigade de Police Technique et Scientifique, Police Cantonale, Geneva, Switzerland
[c] Tri County Regional Forensic Laboratory, Andover, Minnesota
Published online: 21 Jul 2014.

To cite this article: Glenn Langenburg, Flore Bochet & Scott Ford (2014) A Report of Statistics from Latent Print Casework, Forensic Science Policy & Management: An International Journal, 5:1-2, 15-37

To link to this article: http://dx.doi.org/10.1080/19409044.2014.929759

Forensic Science Policy & Management, 5(1–2):15–37, 2014

ISSN: 1940-9044 print / 1940-9036 online
DOI: 10.1080/19409044.2014.929759

A Report of Statistics from Latent Print Casework

Glenn Langenburg,[1] Flore Bochet,[2] and Scott Ford[3]
[1]Minnesota Bureau of Criminal Apprehension, St. Paul, Minnesota
[2]Brigade de Police Technique et Scientifique, Police Cantonale, Geneva, Switzerland
[3]Tri County Regional Forensic Laboratory, Andover, Minnesota

ABSTRACT Statistics were derived from casework from the Minnesota Bureau of Criminal Apprehension Latent Print Unit. These data represented a portion of the latent print casework completed in the 2003/2004 calendar years (N = 673 cases) and 2009/2010 calendar years (N = 885 cases). The 2003/2004 data revealed latent print recovery rates from various exhibits. Identifiable latent prints were recovered 13% of the time on firearms, 13% of the time on plastic bags, and no identifiable latent prints were recovered from fired or unfired ammunition. The processing of evidence prior to submission to the laboratory increased the chances of latent print recovery. Both data sets were explored for the rate at which identifiable latent prints were reported (61% of cases in 2003/2004 and 54% of cases in 2009/2010) and the rate at which identifiable latent prints were subsequently identified to an individual in the case (23% of cases in 2003/2004 and 25% of cases in 2009/2010). There was no noticeable difference for the identification rate in property crimes versus crimes against people.

The 2009/2010 data were explored for possible effects from analysts having access to contextual information or significant interaction and communication with police officers or prosecutors while working a case. We noted that 2% of cases in the data qualified for this condition—the majority of BCA-LPU cases are worked without contextual information or police interaction. Comparing high context/high interaction cases versus no context/no interaction cases, we found the latent print identification rates to be equal (21% versus 22%, respectively).

KEYWORDS Fingerprints, bias, statistics, recovery rates, firearms, ammunition

Received 13 March 2014;
accepted 27 May 2014.

Address correspondence to Glenn Langenburg, Minnesota Bureau of Criminal Apprehension, 1430 Maryland Avenue East, Saint Paul, Minnesota 55106. E-mail: glenn.langenburg@state.mn.us

Color versions of one or more figures in this article can be found online at www.tandfonline.com/ufpm.

INTRODUCTION

Finding a source for detailed and accurate fingerprint evidence from a crime lab can be difficult. While some sources have provided general trends for forensic service providers, proficiency testing results, or crime justice statistics (5; Peterson et al. 2013), few crime labs actually publish data from their case results. Elsewhere, we have reported data from a field study that focused on the volume of unrecovered evidence and its potential weight of evidence (Neumann et al.

2011), but that study did not examine elements such as recovery rates from various processing techniques, submission trends, AFIS use and success, etc. The aim of the present paper is to provide casework statistics, such as latent print recovery rates and rates of identification, that one would find in a fingerprint laboratory.

With respect to latent print recovery rates, recovery rates on firearms and ammunition in *actual casework* have been reported elsewhere (Barnum and Klasey 1997; Johnson 2010; Pratt 2012; Maldonado 2012). These sources noted consistent recovery rates of 11%, 12%, 10%, and 13%, respectively, for firearms or magazines from firearms, depending on the study. We wish to contribute to those data as well, while adding another layer of information by further subcategorizing our firearms, as was done by Pratt (2012). Recovery rates of latent prints from plastic bags from casework have not been reported to date.

A portion of the present paper was dedicated to the exploration of possible bias effects from significant interaction between the forensic analyst and the case investigator, or from analyst exposure to contextual information about the case—information which has nothing to do with the processing of the evidence. Much has been made of these interactions, and there is general concern for the influence it may have on the accuracy of the results from a crime lab (Kassin, Dror, and Kukucka 2013; Dror 2013; Dror and Hampikian 2011). Yet to date, no source has demonstrated that, in a crime lab that works a high volume of cases, these errors are frequent and exposure to contextual case information is to blame. Contrived research, anecdotal cases, and miscarriages of justice have showcased these dangers (Office of the Inspector General [OIG] 2006; Cole 2006; Dror and Charlton 2006). Yet, in comparable non-forensic, diagnostic testing domains, such as radiological diagnostic testing, there is considerable debate about the advantages and disadvantages of making patient clinical history available to the radiology technician to render an accurate and efficient assessment of the case (Potchen et al. 1979; Potchen et al. 2000; Loy and Irwig 2004; Dhingsa et al. 2004). Furthermore, some research in the forensic domain has pointed toward the benefits of information exchange between analysts and investigators (8; 9; 3; Roberts and Willmore 1993), while still acknowledging the pitfalls of bias effects. This has prompted some authors to argue that shielding a forensic analyst from case information or failing to consider the evidence in the context of the specific case may in fact lead to more error or missed opportunities to critically evaluate the evidence (1; Thornton 2010). They argue, generally, that forensic scientists should enter a professional dialogue with the investigator to develop an appropriate resource-conscious forensic strategy. This strategy can limit the examination and testing just to those evidential items which can impact the investigation.

In the midst of this debate, there has been a call for better quality assurance measures to prevent domain irrelevant information exchange between the analyst and the investigator (National Research Council 2009). These suggested measures have ranged from blinding the analyst from all domain irrelevant information in every case (Haber and Haber 2008) to a sequential unmasking approach, whereby case information is revealed ("unmasked") after critical decision making stages have been completed (Krane et al. 2008). In this scheme, the analyst will eventually have access to all the case information, but only after it cannot influence the analyst's decision. Other variations to these schemes have been proposed such as blind verification in select cases (Cole 2013) or evidence line-up/distractor sample approaches (Wells, Wilford, and Smalarz 2013). Typically, these quality assurance measures must be introduced by a case coordinator, who assigns the case, filters information, and acts as a liaison between the analyst and investigator. This approach raises some questions such as: 1) which information should be kept from an analyst? 2) what if contextual case information could help the analyst make more accurate, efficient, and informed decisions about the case? and 3) at what cost (both monetarily and in terms of benefits versus risks) do these changes bring? (Langenburg 2012).

The present paper explores these issues and identifies which cases may actually present the most danger of error from bias. This will give a clearer picture of what resources are required to address this issue or where best to concentrate efforts and quality assurance measures to limit bias effects.

Demographics of Minnesota and the BCA-LPU

The data in the present paper represent samplings from actual casework for the Minnesota Bureau of Criminal Apprehension Latent Print Unit (BCA-LPU). To properly assess these data, it is important to

understand how the BCA-LPU operates and what are the characteristics of BCA-LPU, the BCA in general, and the State of Minnesota. Before comparing data between agencies, it is important to ensure that what constitutes a "case," similar workflows, and similar processes are compared for a fair apples-to-apples comparison.

There are approximately 5.3 million people living in Minnesota (United States Census Bureau 2012). About 60% of the population (3 million people) live in the Minneapolis/St. Paul metropolitan area and suburbs, and the remainder of the population is spread throughout the mostly rural farmland or heavily wooded and lake abundant state.

The BCA-LPU is the latent fingerprint section for the State of Minnesota. The BCA-LPU services 87 counties. In actuality, since the two largest metropolitan areas in Minnesota, St. Paul and Minneapolis, have their own latent print units, the BCA-LPU does not routinely receive requests from these agencies. In effect, the BCA-LPU receives the cases from the greater metropolitan area and the rest of the State of Minnesota. The BCA-LPU is comprised of two laboratories: the headquarters laboratory in St. Paul and a regional laboratory in Bemidji. The St. Paul lab services the lower half of the state and the metropolitan area and the Bemidji lab services the upper half of the state, which is more rural and less populated. The BCA-LPU currently employs seven analysts (two in the satellite lab and five in the central headquarters). The range of experience of these analysts is from 4 years to 25 years in latent prints. The BCA-LPU is part of an accredited laboratory system, under ISO17025, and offers other testing services (e.g. DNA, firearms, etc.). All of the BCA-LPU analysts are certified latent print examiners by the International Association for Identification (IAI).

The BCA-LPU provides processing, comparison, and AFIS services. Analysts typically process their own evidence, perform photography of any identifiable latent prints, perform the comparisons, enter unidentified latent prints into AFIS, and write their reports. The BCA laboratory offers on a voluntary basis, the opportunity to join the BCA Crime Scene Team, which primarily assists local law enforcement when requested on homicides, kidnapping, officer-involved shootings, etc. Many of the BCA-LPU serve/have served on this team. This is relevant because, in those cases, the attending analyst often will also be the case-working fingerprint analyst. The authors anticipate that the readers will have mixed feelings about this. On the one hand, the attending analyst understands why and how the latent print evidence was collected and which evidence is most critical. On the other hand, there may be concern that this level of interaction, exposure to contextual information, and perhaps even emotional investment, may influence an analyst's decision in the case. The authors specifically wanted to explore that issue in this paper as well.

In the vast majority (over 99%) of the cases received by the BCA-LPU, the analysts receive case submissions from local law enforcement. These local police and sheriff departments have responded to a scene, collected (and possibly processed to some extent) evidence, and submitted it to BCA. The delay of evidence submitted to the BCA-LPU can vary from a few days to sometimes more than a year or more after the crime. Because the evidence is received by "Evidence Specialists," who take evidence into the BCA for all the forensic sections at the BCA, the BCA-LPU analysts rarely have contact with a submitter at the time of delivery. In the course of working the case, the analyst may have a need to contact the investigator with follow-up questions. These questions may occur at the beginning of the process (e.g. "which of these 100 items should I start processing first?" "this person does not have a fingerprint record against which to compare") or near the end of the process (e.g. "I have identified the suspect in the case several times, do I need to continue to compare all the remaining 20 latent prints to this suspect too?"). Often these questions help the analyst to allocate their time and resources effectively. The concern by some commentators is that in the course of those conversations, the potential to be exposed to biasing contextual information exists (Dror, Charlton, and Péron 2005; Mnookin 2010).

The BCA-LPU received approximately 1400 case submissions for the 2012 calendar year. This submission rate has steadily climbed over the last 10 years. The submission rate was around 1,000 to 1,100 cases ten years ago. A Bureau of Justice (BJS) survey in 2005 reported the median number of latent print examination requests in the U.S. was 909 cases for the 194 agencies that responded to the BJS survey (Durose 2008). This places the BCA-LPU slightly above those submission rates, and certainly these numbers have increased since 2005.

MATERIALS AND METHODS

For the present paper, two data sets were prepared by random sampling of completed BCA-LPU cases. The first data set, which focused on recovery rates, is referred to as the 2003/2004 data set. The second data set, which focused on rates of identification, impact of AFIS, and effect of exposure to case context information and interactions between forensic analysts and investigators, is referred to as the 2009/2010 data set.

The 2003/2004 set was prepared by sampling 673 cases from a 12-month period of cases worked by the BCA LPU in mid-2003 through mid-2004. At that time, the BCA LPU was working about 1,000 to 1,100 cases per year. This sample is about two-thirds of the cases worked in that time period. Specifically, the sample represented about 50% of the cases worked in the St. Paul laboratory, and about 70% of the cases worked in the Bemidji laboratory in this time period.

Data were collected through the use of a data sheet prepared for each case. At the end of the data collection period, the data were entered into a Microsoft Access (2003) database for analysis.

The 2009/2010 data set was prepared by sampling 885 cases from a 12-month period of cases worked by the BCA LPU in 2009 and 2010. There were approximately 1,200 cases per year received by the BCA in 2009 and 2010. This sampling represented approximately 75% of the cases worked in St. Paul and 30% of the cases worked in Bemidji. Caution is warranted when comparing the data from 2003/2004 to the data in 2009/2010; proportions should be compared to minimize sampling and population size differences.

The BCA codes a case during its submission based on the submitting officer's description of the case. For the 2003/2004 and 2009/2010 data sets, we pooled case types together to identify four classes of case type. These are:

1) Property crimes: includes burglary, theft, auto theft, fire investigation, forgery, fraud, stolen property, and vandalism.
2) Crimes against people: includes cases with death investigation, homicide, attempted homicide, robbery, criminal sexual conduct, assault, kidnapping, threats, stalking, hit and run, etc.
3) Drugs: includes controlled substances with possession, sale, or manufacture.
4) Weapons: includes cases with unlawful discharge or unlawful possession of a firearm.

In the 2009/2010 data, we assessed the level of interaction between the case analyst and the police/investigator(s)/prosecutor(s). We also assessed the amount of contextual information, such as police reports or investigative information, available to the analyst in the case. To collect these data, a worksheet was completed for each of the sampled cases by reviewing the case reports. We also reviewed the LIMS (Laboratory Information Management System), which tracks case information and would include such things as communiqués between the analysts and investigators, police reports available to the analyst at the time of the examination, and notes regarding the analysts' observations or decisions in a case.

We categorized the level of interaction as "high," "moderate," or "none/minimal." The level of interaction was deemed "high," "moderate," or "none/minimal" based on the following criteria:

- High = significant interaction between investigators or prosecutor, resulting from at least 3 phone calls, at least 3 email exchanges, or attendance at the crime scene.
- Moderate = 1-2 email or phone call exchanges between submitting officer(s), prosecutor(s), or investigator(s) typically where case information and details are exchanged.
- None/minimal = no recorded contact with submitting officer(s), prosecutor(s), or investigator(s), or minimal contact to clarify a case question (e.g., an e-mail to check the spelling or date of birth of a suspect, a phone call asking if the item had already been processed, etc.

This assignment was obviously a judgment call of the researchers. If there was any doubt, and any case information appeared to be exchanged with the analyst and the requesting parties, then the case was classified at a minimum as "moderate" interaction. We also considered the reading of case information to be a type of "interaction." If it was clear in the LIMS that the analyst had read considerable case information (high or moderate context report) then the level of interaction would be increased one level (i.e., "none/minimal" interaction was raised to

"moderate" if the analyst clearly read a detailed case report). Although it should be noted that it was only clear in 9 of 885 cases in the LIMS that the analyst had read the report. It was also possible that a detailed report was present, but it was not read by the analyst.

The amount of contextual information available to the case analyst was categorized as "high," "moderate," or "none/minimal." The level of contextual information was deemed "high," "moderate," or "none/minimal" based on the following criteria:

- High = significant case details were available in LIMS. Typically, in "high context" cases, officers have submitted detailed reports about the scene or the investigation. These reports may include investigator theories, detailed interviews with suspects, suspect statements, or details and observations made by investigators at the crime scene or during collection of the evidence. Cases where the analyst attended the crime scene were also deemed "high context."
- Moderate = short reports or details about the crime or investigation were provided by the investigator in addition to the standard submission forms required by BCA.
- None/minimal = no case details were provided at all, or only minor, domain relevant information, or required information for case submission were provided on standard BCA submission forms.

RESULTS AND DISCUSSION

Latent Print Submissions by Case Type

The BCA LPU received approximately 1,000 to 1,200 case assignments per year in the considered time frames for both data sets. Recent submission rates for 2011 and 2012 have increased by 20% to approximately 1400 per year.

The distribution of cases for both the 2003/2004 data set (recovery rate data set) and the 2009/2010 data set (conclusion rate data set) is shown in Figure 1 below. It can be seen that property crimes were the most common case type submissions for latent prints. The BCA-LPU received over four times as many property crimes as crimes against people or drug cases.

Care must be taken when comparing the two data sets in Figure 1. The two samples have different sizes, $N = 673$ and $N = 885$. A two-sample Z test for proportions can be used to assess the statistical significance of the difference in submissions between the two data sets. There was a significant increase ($Z = -4.18$; $p < 0.001$) for property crime cases from 2003/2004 to 2009/2010; 432 out of 673 cases (64%) in 2003/2004 were property crimes compared to 655 out of 885 cases (74%) in 2009/2010. Simultaneously, there was also a significant decrease ($Z = 3.15$; $p = 0.002$) in crimes against persons submissions for latent print analysis. The differences between the number of weapons and drugs submissions between the data sets were not statistically significant ($p > 0.05$). The shift in property

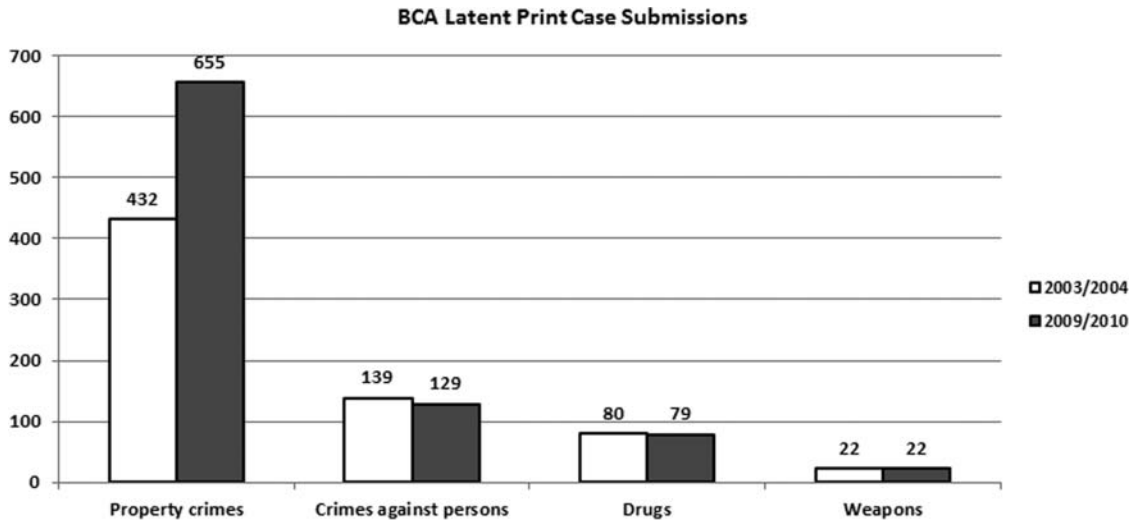


FIGURE 1 Comparison of case submissions requesting latent print examinations by category of case type.

crimes and crimes against people may be due to changes in the types of cases submitted for DNA analysis.

From 2003 to 2009, the BCA saw a significant increase in property crime cases submitted for DNA analysis. In 2003, the BCA received 1,714 DNA assignments; 224 (13%) were property crime cases. In 2009, the BCA received 3,407 DNA assignments; 907 (27%) were property crimes. The sheer volume of casework for DNA had doubled, but the proportion of property crimes for DNA analysis had also doubled. In many of these cases, latent prints were also being requested by the submitters, or a DNA analyst at the BCA would recommend latent print examinations to the submitter in lieu of, or sometimes in addition to, DNA examinations. This collateral effect is clearly seen in Figure 1, both in the increase in submissions, but also in the increased proportion of property crimes. We refer to this as the "DNA trickle down" effect.

It should also be noted that, in 2003, there were 15 BCA DNA analysts to work the 1,714 submissions. In 2009, when the number of DNA submissions doubled to 3,407, the number of BCA DNA analysts had also nearly doubled to 27. In 2003, the BCA LPU had 7 fingerprint analysts In 2009, the BCA LPU had 7 fingerprint analysts. Today at 300 more submissions annually than in 2009, the BCA LPU has 6 (and a half timer) fingerprint analysts.

While funding and backlog reduction funds (e.g., Coverdell grant) have been prioritized for DNA laboratories in the U.S., the same cannot be said for most latent print units. Unfortunately, the latent print sections have not received the benefit of funding and personnel to match their DNA counterparts. As a result, the increase in DNA testing requests has increased the burden on the latent print section without a commensurate investment in latent print personnel or resources.

Identifiable Latent Prints and Identification Rates

When determining the intrinsic value of latent print evidence, an analyst at the BCA-LPU will first note the presence of ridge detail, if any, observed on the exhibit. Then the analyst will determine its "suitability" (or in some agencies "value") for comparison. This is the analyst's judgment of the utility of the impression and the likelihood that they will be able to reach a definitive conclusion ("identification" or "exclusion"). Agencies will vary in how they apply this approach as noted by SWGFAST standards (Scientific Working Group on Friction Ridge Analysis Study and Technology [SWGFAST] 2013). BCA-LPU subscribes to Approach #2 as described in those standards, whereby most impressions are compared with the expectation that they can be identified when presented with the correct source exemplars, but not in all cases. In some cases, the correspondence may be insufficient and an "inconclusive" opinion due to the limited information in the latent print, may be rendered. For the non-technical reader, we have opted for the remainder of the paper to refer to these latent prints that have been deemed comparable by the analyst as "identifiable," although in actual practice at the BCA-LPU we use the term "suitable for comparison." Finally, it must be clarified, that this decision of "suitability" takes place before ever viewing the exemplars of any of the subjects in the case; it takes place during the analysis stage of the Analysis-Comparison-Evaluation-Verification (ACE-V) process (Langenburg and Champod 2011).

In the 2003/2004 data, we recorded if the analyst observed "any ridge detail." This would include cases where ridge detail was observed by the analyst, but not recovered due to the perceived inability to exploit the ridge detail. This question was not asked in 2009/2010, although cases from 2009 comprised the data set used in a previous study (Neumann et al. 2011) where the amount of unrecovered ridge detail was quantified and explored. In the 2009/2010 data, we were only concerned with the proportion of cases with identifiable latent prints. Lastly we examined the proportion of these cases where the identifiable latent prints resulted in "identification" decisions to either the victims or the suspects. The distinction between victim and suspect identifications was not made in the 2003/2004 data, but was explored in detail in the 2009/2010 data. These data are shown in Table 1 and they are further deconstructed by case type.

In the 2003/2004 data, 575 out of 673 (85%) cases had at least one item of evidence that bore some visible ridge detail for the analyst to evaluate for its potential "value." Of these 575 cases, 410 (410 out of 673 total cases = 61%) resulted in latent prints deemed "identifiable." Finally, for these 410 cases where suitable ridge detail was observed, 152 cases

TABLE 1 The Proportion of Cases with Identifiable Latent Prints and "Identification" Decisions are Compared Between the 2003/2004 and 2009/2010 Data Sets. Percentages Reported are Using the Total Number of Considered Cases (N = 673 and N = 885) as the Denominator

	2003/2004 (N = 673)			2009/2010 (N = 885)	
	Number of cases with any ridge detail observed	Number of cases with identifiable latent prints	Number of cases with "identification" reported	Number of cases with identifiable latent prints	Number of cases with "identification" reported
Property crime	398 (59%)	304 (45%)	101 (15%)	384 (43%)	167 (19%)
Drugs	53 (8%)	31 (5%)	17 (3%)	26 (3%)	14 (2%)
Weapons	14 (2%)	4 (<1%)	4 (<1%)	4 (<1%)	1 (<1%)
Crime against persons	110 (16%)	71 (11%)	30 (4%)	66 (8%)	40 (5%)
Total	575 (85%)	410 (61%)	152 (23%)	480 (54%)	222 (25%)

(152 out of 673 total cases = 23%) had at least one "identification" decision.

In the 2009/2010 data, 480 out of 885 (54%) cases bore at least one latent print deemed identifiable. If we compare this to the 2003/2004 data, we see there is a drop from 61% to 54% of cases with identifiable latent prints. This is a statistically significant decrease ($Z = 2.64$: $p = 0.008$), and may be due in part to the previously discussed "DNA trickle down" effect from increased property crime submissions for both DNA and latent prints. It may be possible that some of these exhibits selected for DNA testing may not have been the most appropriate or conducive for latent print evidence, but since the exhibit has been submitted for DNA, the officer requests latent print examination to be done anyway. There is no actual cost to the officer or prosecuting attorney and these decisions may not always be carefully considered. It may be one of the factors leading to some of the observed backlogs in crime labs (Durose 2008). Perhaps an approach closer to the "case assessment model" as proposed by Cook, et al. (1998) may lead to better screening and evidential choices. A discussion between the scientist and the investigator may allow for better choices when selecting which items to analyze, or which tests to perform, despite the potential risk of bias.

Recovery Rates From Various Exhibits

The rate of recovery of latent prints from various substrates and exhibit types was not explored in 2009/2010, therefore the data below only represent the 2003/2004 dataset. The cases were sorted into three categories:

1) Lifts only: these were cases where latent prints were recovered at the scene only by tape lifts or photographs. No exhibits to examine or process were submitted.
2) BCA processing: these cases required processing of exhibits by the BCA. They are the most time consuming due to the sequential application of different development techniques.
3) Submitting agency processed: these cases had exhibits that were processed by technicians prior to the submission to BCA. Processing may have occurred in the field or at the submitter's agency.

Figure 2 shows the relative proportions of cases where "lifts only," "submitter processing," or "BCA processing" was performed. Of the 673 reviewed cases, 330 cases (49%) were cases where only lifts were submitted from evidence technicians in the field, 288 cases (43%) were cases were BCA was required to process exhibits, and 55 cases (8%) were cases where the evidence technician did the processing before submitting the exhibit. Roughly speaking then, about half the cases submitted to BCA required no processing, while half the cases required some processing and/or photography.

When we examined the effect of processing by technicians prior to submission, we see in Figure 3, that the lift cases bore identifiable latent prints 77% of the time, while the submission of the exhibit only for BCA processing produced identifiable latent prints 41% of the time. Where the submitter performed processing of the exhibit prior to submission, identifiable latent prints were recovered 67% of the time. One of the explanations for this difference may be that the submitters processed many more items that were not

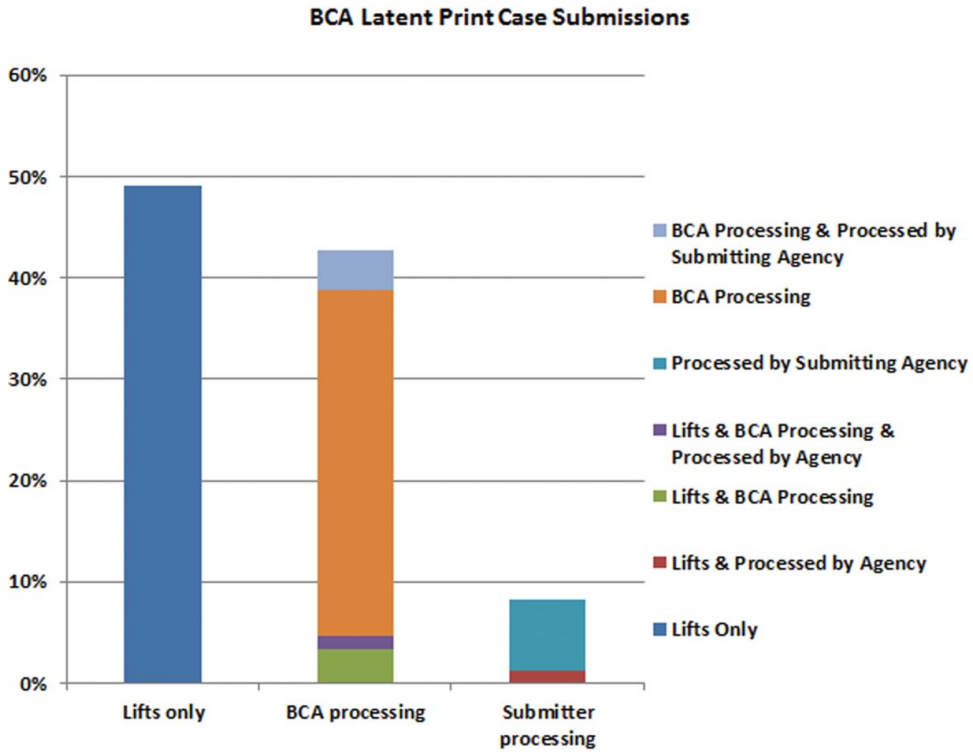


FIGURE 2 Distribution of cases in the 2003/2004 data set by level of pre-processing performed by submitters to BCA.

submitted; they only submitted those items where they observed some apparent ridge detail. The same would be true with respect to lifts. This may demonstrate an efficient selection of exhibits, both for the presence of useable ridge detail, but also for the purpose of choosing to process the exhibit in the first place. In other words, field technicians may be making good choices about what exhibits to process and which to submit.

Another explanation (not mutually exclusive) for the high recovery of identifiable latent prints from pre-processed exhibits is that preservation in the field, or after a relatively short time from the deposition of the latent print, may increase the recovery rates due to the fragility and volatility of latent print residues. While the crime lab may have premier equipment and expertise in the development of latent prints, these

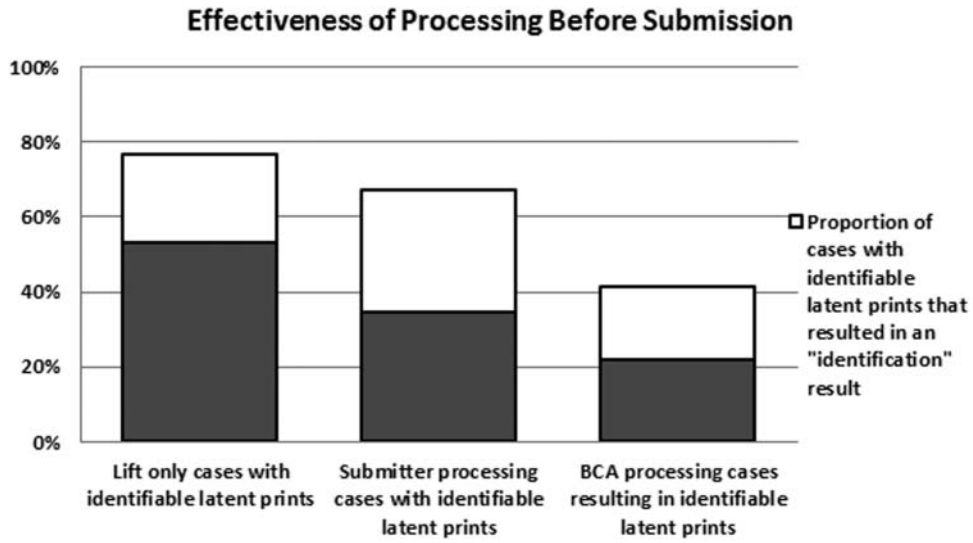


FIGURE 3 The percentage of cases that resulted in identifiable latent prints and the fraction of those cases with identifiable latent prints that resulted in an "identification" decision reported.

advantages may be lost when the evidence sits for several months before being processed due to delays in submission and case backlogs. Lastly there may be some potential loss of evidence during the collection, packaging, and transportation of the unprocessed evidence to the crime laboratory.

Figure 3 also shows the relative proportion of cases with identifiable latent prints which subsequently led to at least one "identification" decision in the case. It can be seen in Figure 3 that while lift cases produced identifiable latent prints 77% of the time, only about one-third (31%) of these cases led to an "identification." In the cases where the BCA processed the item or the submitter processed the item, identifiable latent prints were recovered about half the time (47% and 49% respectively). A possible explanation for this difference is, again, the relevance of the exhibit. In lift cases, lifts may often come from immovable objects in public places or with unrestricted access (doors, counters, windows, tables, vehicles, vending machines, Automatic Bank Teller Machines, etc.). Many individuals without relation to the crime could have touched these surfaces from which the lifts were generated. Whereas the choice to process an exhibit with cyanoacrylate, ninhydrin, etc. may be with an eye towards a very relevant object related to the crime, with limited access to a handful of individuals.

The BCA has four major protocols for processing evidence depending on the type of surface and latent print residue that may be deposited on the substrate. These processing protocols are: 1) non-porous (e.g. glass, plastic, metal, etc.); 2) porous (e.g. papers, checks, cardboard, etc.); 3) adhesive (duct tape, stickers, stamps, etc.); 4) blood processing (enhancement of visible ridge detail deposited with blood matrix). Figure 4 shows that in the 288 cases where BCA processing was required, non-porous processing (N = 206) is the most commonly used processing protocol at BCA, followed by porous processing (N = 51). Some cases (N = 29) required the use of multiple processing techniques. Typically, when porous processing or multiple techniques are required, these cases become more labor intensive and time-consuming. Also, most exhibits where tape is involved require multiple processes (i.e. non-porous and adhesive processing).

In the 2003/2004 data, we collected information about the recovery rates of identifiable latent prints from various non-porous exhibits. We did not look at recovery rates for porous, blood, or adhesive processing cases. We investigated latent print recovery rates for three categories of exhibits: 1) plastic bags, 2) firearms, and 3) ammunition for firearms (see Table 2).

In the 45 cases where plastic bags were submitted, 201 plastic bags exhibits were processed for latent prints. Twenty-six (26) identifiable latent prints were recovered from these 201 plastic bags. This is an average recovery rate of 13% for the plastic bags. It should be noted that the true recovery rate may actually be lower since some of these 26 identifiable latent prints

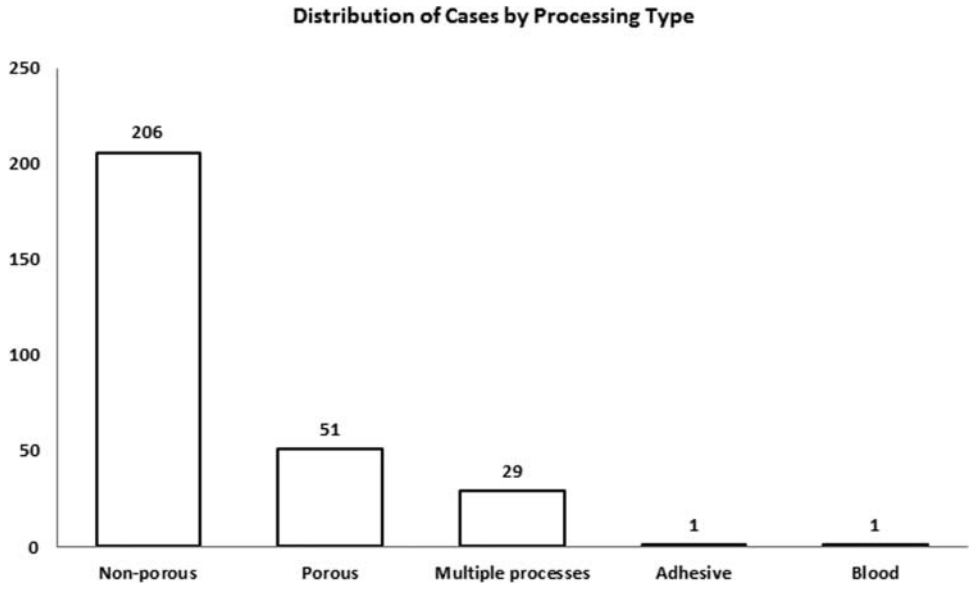


FIGURE 4 The distribution of the cases submitted to BCA (N = 288) for processing in the 2003/2004 data set (N = 673).

TABLE 2 Distribution of Cases and Exhibits that were Processed at BCA in the 2003/2004 Data Set

	Plastic bags	Firearms	Ammunitions
Number of cases with selected exhibit type	45	73	40
Number of exhibits processed	201	104	341
Number of identifiable latent prints	26	14	0
Identifiable latent print recovery rate	13%	13%	0%
Number of "identification" decisions reported	14	5	0

were found on the same bag. In other words, 13% recovery rate is likely an overestimate. In the 73 cases where firearms evidence was submitted, 104 firearms were processed for latent prints. Fourteen (14) identifiable latent prints were recovered from these 104 firearms. This is an average recovery rate of 13%. Again, this may be a slight overestimate if multiple latent prints were found on the same firearm, but these data are similar to other reported sources (0; 2; Pratt 2012). Finally, 40 cases were submitted for the processing of firearms ammunition. In 341 exhibits, no identifiable latent prints were recovered. This exceedingly low probability of success for latent print recovery on ammunition is also noted by the same aforementioned sources.

The low recovery rates from ammunition raises two important points. The first point is that given the low (non-existent) success of latent print processing techniques on ammunition, perhaps these exhibits should be going exclusively for DNA testing. Recovery of DNA from cartridges and cartridge cases, while still low and often involves mixtures or low-quantity DNA (1; Horsman-Hall et al. 2009), is still more successful on average than latent print processing. The second point is that these low recovery rates, in contrast to the constant success of our fictional TV counterparts, is likely contributing to the increased demand for what is referred colloquially by examiners as "negative testimony"—testimony in jury trials to address the question of why no identifiable latent prints were recovered from the exhibit(s).

We explored the sub-classification of plastic bag, firearms, and ammunition exhibits as shown in Tables 3, 4, and 5. The 201 plastic bag exhibits consisted of plastic bags of various types and size (see Figure 5). The 104 firearms exhibits consisted of pistols, revolvers, shotguns, and rifles (see Figure 6). The 341 ammunition exhibits consisted of various caliber fired and unfired ammunition. In the plastic bag category, it can be seen that Ziploc bags and garbage bags were the most successful for latent print recovery. This is likely due to the larger surface area and generally smoother surface of these exhibits. In the firearms category, recovery rates were higher for rifles over shotguns. Revolvers gave the highest recovery rates for all the firearms. Lastly, in the ammunition category, neither the cartridge, nor the cartridge case was a substrate conducive to the development of latent prints. Anecdotally, in the tens of thousands of cartridges and cartridge cases processed at the BCA in the last 30 years, only a handful of identifiable latent prints have been recovered. These tended to be on large caliber rifle or shotgun ammunition. Therefore, these results for ammunition processing are not surprising to us.

TABLE 3 Latent Print Recovery Rates for Plastic Bag Exhibits in the 2003/2004 Data Set

	Plastic bags		
	Ziploc bag	Sandwich bag	Garbage bag
Number of cases with selected exhibit type	30	20	3
Number of exhibits processed	133	65	3
Number of identifiable latent prints	22	2	2
Identifiable latent print recovery rate	17%	3%	67%
Number of "identification" decision reported	12	0	2

TABLE 4 Latent Print Recovery Rates for Firearms Exhibits in the 2003/2004 Data Set

	Firearms			
	Revolver	Pistol	Shotgun	Rifle
Number of cases with selected exhibit type	11	42	16	24
Number of exhibits processed	14	50	14	32
Number of identifiable latent prints	5	3	1	5
Identifiable latent print recovery rate	36%	6%	7%	16%
Number of "identification" decision reported	2	2	0	1

Conclusion Rate Data (2009–2010 cases)

The results in the following sections originate from the 2009/2010 data set. In these data, we primarily explored the distribution of conclusions reported by the analysts. We also explored the potential impact of case information and interaction with investigators. Lastly we identified and explored a subset of cases where a single latent print was recovered and associated with a suspect in the case. These issues were not explored in, and therefore not comparable to, the 2003/2004 data set.

Finger and Palm Print Distribution

As previously noted in Table 1, there were 480 cases (out of 885) cases with identifiable latent prints (see Table 1). In these 480 cases, there were a total of 1,446 identifiable latent prints that were recovered. Table 6 shows the distribution of whether they came from a finger, a palm, or a finger joint (including cases where the anatomical origin cannot be determined). We also investigated if these distributions were dependent on case type, i.e. was the analyst more likely to recover palm prints in a homicide than in a burglary. There was no significant change in the distribution per case type category (crimes against people, property crimes, drugs, weapons, other).

Approximately 1 in 7 recovered identifiable latent prints was a latent palm print. With respect to the rate of identification, latent fingerprints and latent palm prints were being identified at fairly similar rates (41% and 32%, respectively). This is in sharp contrast to the 11% rate for latent finger joints or "unknown" (when the analyst could not state with any certainty if the latent print was from a finger or palm due to the lack of anatomical or orientation focal points to associate with a finger or palm). There are two reasons (not mutually exclusive) for this. The first is that an analyst may have a better chance of finding the latent print "match" if he or she knows where to look. Since many comparisons are still being done manually by the analyst and without the aid of computers, the analyst must have a good idea where to look for the latent print, or search every conceivable area of friction ridge skin in each suspect or victim. The second reason is that these latent prints tend to be from areas of the skin not routinely captured during standard booking procedures and therefore the proper comparable area was not recorded. The exemplars are incomplete and a "match" is impossible.

An important point from these data is that a significant amount of latent print evidence originates from

TABLE 5 Latent Print Recovery Rates for Ammunition Exhibits in the 2003/2004 Data Set. A Cartridge is Unfired Ammunition; A Cartridge Case is the Case from a Fired Cartridge

	Ammunition	
	Cartridge	Cartridge case
Number of cases with selected exhibit type	31	22
Number of exhibits processed	253	88
Number of identifiable latent prints	0	0
Identifiable latent print recovery rate	0%	0%
Number of "identification" decision reported	0	0

FIGURE 5 Examples of various types of plastic bag exhibits. These examples illustrate the style of bags categorized in the 2003/2004 data set as "sandwich bags" (left), "Ziploc bags" (center), and "garbage bags" (right).

palm prints. There are still a number of agencies without the capabilities of searching palm print databases or recording palm prints during booking. There is no doubt that they are missing opportunities to identify suspects in cases. The need for specific palm print comparison training and the need for technology which capitalizes on palm print recording and databases is an absolute necessity.

Rates of Identification

We examined the number of latent prints that were deemed "identifiable" in four broad categories of case types: crimes against people, property crimes, weapons cases, and drugs. Table 7 shows the distribution for these case type categories. From Table 7 it can be seen that about half (54%) of the submitted cases to BCA in the 2009/2010 data set resulted in at least 1 "identifiable" latent print found on the evidence. These "identifiable" latent prints were predominantly found in property crimes and crimes against people (59% and 51% of those cases respectively). In drugs and weapons cases the chance of finding an "identifiable" latent print was significantly lower. This is consistent and explainable with the previously considered 2003/2004 latent print recovery data from drugs and weapons exhibits (see Table 1). It is interesting to also note that crimes against people and drug cases produced the most number of "identifiable" latent prints per case, although these cases represent a smaller fraction of all the cases submitted to BCA. This trend can be observed in Table 7.

FIGURE 6 Examples of various types of firearms exhibits. These examples illustrate the style of firearms categorized in the 2003/2004 data set as "revolver (A)," "pistol (B)," "shotgun (C)," and "rifle (D)."

TABLE 6 The Distribution of Identifiable Latent Prints that Originated from Fingers, Palms, or Finger Joints/Unknown

	Identifiable Latent Prints (N = 1446)		
	Fingers	Palms	Joints/Unknown
Number of identifiable latent prints (% of total)	1124 (78%)	221 (15%)	101 (7%)
Number that were identified (% of identifiable latent prints)	461 (41%)	72 (32%)	11 (11%)

In crimes against people cases, there may be several reasons for observing more "identifiable" latent prints per case. One reason is that a full battery of possible examinations are typically done in these types of cases, and only a single routine process may be employed in property crimes. Thus using all available sequential processes may result in more recovered latent prints. It may also be influenced by the relevance of the evidence collected by specialized and trained crime scene technicians. Another consideration, as suggested by some commentators, is motivation (Charlton, Fraser-Mackenzie, and Dror 2010). This is the notion that forensic analysts motivated by the severity of the crime will be more inclined to include marginal latent prints (thus "pushing the envelope") in a conscious or subconscious drive to aid investigators in serious and violent crimes. This issue will be explored in a later section of the present paper.

One argument against the notion of motivated analysts pushing the envelope to find marginal latent prints in more serious or violent crimes is the fact that Table 7 shows that in 51% of crimes against people identifiable latent prints were recovered. This can be compared to the 59% of property crimes where identifiable latent prints were recovered. One would expect a much higher percentage of identifiable latent prints claimed in crimes against people if the seriousness of the crime was influencing the analysts' decisions for value determinations. This is not to say that it is not occurring in some isolated incidents, but clearly there is not a trend here of rampant bias to include marginal latent prints in the "identifiable" category in these cases.

It is important to also consider, that although there is a small percentage of total cases (6%) with 6 or more identifiable latent prints, these cases tend to be very time consuming. When these cases have multiple suspects and victims against which to compare, a substantial amount of comparison time will be spent by the initial analyst and possibly, a second analyst who will have to verify the conclusions in a case. A more detailed analysis showed that these cases with more than 6 identifiable latent prints were predominantly homicide cases or stalking/harassment cases. Anecdotally, homicide cases tend to produce more exhibits and have more processing, and stalking cases tend to produce large amounts of identifiable latent prints often on a series of letters sent to the victim over time—many of which are handled by several people before finally involving the police.

Investigating further, we explored the rate of identification and exclusion decisions. Table 8 shows the

TABLE 7 Distribution of Identifiable Latent Prints Per Case Type Category

	Number of cases considered	Number of cases with at least 1 identifiable latent print (% of cases considered)	Total number of identifiable latent prints	Average number of identifiable latent prints per case (where there was at least 1 latent print)
Property crimes	655	384 (59%)	1014	2.6
Crimes against people	129	66 (51%)	307	4.6
Drugs	79	26 (33%)	114	4.4
Weapons	22	4 (18%)	11	2.8
Total	885	480 (54%)	1446	3.0

TABLE 8 Rate of "Identification" and "Exclusion" Decisions Sorted by Case Type Category

	Number of identifiable latent prints considered	Number of "identification" decisions	Number of "exclusion" decisions
Property Crimes	1014	370 (36%)	657
Crimes Against People	307	126 (41%)	511
Drugs	114	47 (41%)	69
Weapons	11	1 (9%)	17
Total	1446	544	1254

distribution of these rates for the four case type categories. We see that the rate of identification for property crimes, crimes against people and drugs cases are all approximately the same rates (36%, 41%, and 41%, respectively). This is evidence against the notion that analysts are more "motivated" to make (unwarranted) "identification" decisions in crimes against people because of their need to aid police (Charlton, Fraser-Mackenzie, and Dror 2010). Again, this does not preclude the possibility of occurrence in isolated incidents. Interestingly, the rate of "identification" decisions is exceptionally low and the rate of "exclusion" decisions is quite high in weapons cases, compared to the other case types in Table 8. One possible explanation for this is that police officers who are recovering these weapons (especially from a vehicle or off of the suspect) may not be wearing gloves since the primary purpose of the search may be to render their environment safe. In Minnesota, unfortunately, peace officers, fire and emergency personnel do not have their fingerprints in a non-criminal database (by Minnesota statute). Therefore, a number of these exhibits may have police officer prints on them, without any way of identifying the officer in the case. In Table 8, for all case types, we see that the total number of "exclusion" decisions are significantly greater than (by about 2.5 times) the total number of "identification" decisions. It is important to remember that a latent print can only be "identified" once, but a single latent print in the case can result in one "exclusion" decision ***per considered individual***. Therefore, this imbalance of "identification" versus "exclusion" decisions is not surprising, especially in crimes against people where there are significantly more individuals against which to compare.

Number of Suspects, Victims, and Effectiveness of AFIS

Table 9 shows the number of suspect names provided in each case by the submitting officer in the 2009/2010 data set. It is not surprising that weapons and drugs cases almost always (86% and 96% of the time, respectively) have at least one suspect named. It is also not surprising that these cases commonly have multiple suspects named. Often these cases are requested for latent print analysis when a raid or search of a dwelling or vehicle is performed by law enforcement. When they recover the contraband in the dwelling or vehicle, the parties deny knowledge or ownership of the items. Latent prints are usually requested for the government to prove "ownership" or

TABLE 9 Distribution of the Number of Suspects Provided Per Case Type Category

	Number of cases with no suspect provided (% case type total)	Number of cases with 1 suspect provided (% case type total)	Number of cases with 2 to 5 suspects provided (% case type total)	Number of cases with 6 or more suspects provided (% case type total)	Total number of cases
Crimes against people	28 (22%)	63 (49%)	37 (29%)	1 (<1%)	129
Property crimes	380 (58%)	158 (24%)	113 (17%)	4 (1%)	655
Weapons	3 (14%)	6 (27%)	13 (59%)	0	22
Drugs	3 (4%)	39 (49%)	37 (47%)	0	79
Totals	414 (47%)	266 (30%)	200 (23%)	5 (<1%)	885

at least "knowledge of" through contact established by a latent print identification. In crimes against people, it can be seen that most crimes against people (78%) have at least one suspect named in the case. This may be because the nature of these crimes requires contact between two people. The victims may know the perpetrator or perhaps there is a more intense investigation in these cases because of the severity of these crimes. Just over half (58%) of the property crimes submitted to BCA do not have a suspect named. These crimes are often committed when there are no victims present or witnesses to the crime. These cases will require AFIS searches to generate potential suspects.

AFIS was used in 323 out of the 885 reviewed cases (36%). In the BCA-LPU, AFIS is typically utilized for any unidentified latent prints in a case, but only after they have been compared and possibly identified to the victim/elimination prints or a suspect proffered by the case investigator. Furthermore, the unidentified latent prints must be suitable for an AFIS search. Certain types of latent prints (e.g. finger joints, extreme fingertips, etc.) may be identifiable, but not appropriate for a search in AFIS because these areas of the friction ridge skin are not recorded during a standard booking in Minnesota. In the 323 cases where AFIS was utilized, 99 cases generated new suspects. Eighty-two of the 99 cases (83%) where a new suspect was developed were property crimes; 11 cases (11%) were crimes against people, and 6 cases (6%) were drug cases. No new suspects were developed with AFIS in weapons cases.

In the 323 cases where AFIS was used, a total of 658 latent prints were searched in AFIS. This averages to 2 latent prints per case that were entered into AFIS (median = 1). Eleven cases had AFIS entry of 6 or more latent prints; one homicide case had 56 entries and generated 7 new suspects. The AFIS searches led to the development of 111 new suspects based on identifications made from AFIS resulting in an AFIS hit rate of 17%. This also means that AFIS provided a new suspect in approximately 1 in 3 cases (99 of 323 cases) where it was used, and that BCA generated new suspects using AFIS in approximately 1 in 10 of all cases submitted to BCA (99 of 885 cases).

In Table 10, it was reported that there were 544 "identification" decisions in the 2009/2010 data set. These 544 "identification" decisions are sorted into the number of "identification" decisions reported to a suspect in the case versus a victim/elimination source. It should be noted that in drugs and weapons cases there is rarely a "victim" listed, and officer elimination prints are rarely submitted. The proportion of "identification" decisions to suspect versus victim is nearly equal in crimes against people. In property crimes, a suspect was three times more likely to be identified than a victim/elimination source. This is likely due to the reasons as discussed previously: there tends to be contact between the perpetrator and victim in crimes against people, whereas in property crimes the victim(s) are not present during the commission of the crime.

TABLE 10 Distribution of "Identification" Decisions Attributed to Suspects or Victims/Elimination Sources for the 2009/2010 Data Set

	Number of "identification" decisions (N = 544)	
	Suspects	Victim/Elimination
All cases	396 (73%)	148 (27%)
By case type:		
Crimes against people	70	56
Property crimes	278	92
Drugs	47	0
Weapons	1	0

However, it seems plausible that in property crimes, since these are typically burglary or auto theft cases at BCA, we could be equally (or more) likely to find victim prints on surfaces that the victim routinely touches. Since this was not the case in the 2009/2010 data set, does it have something to do with smart choices made at a crime scene? Is this because information exchanged between the victim and the investigator leads to better choices of the most relevant evidence?

Contextual Information and Interaction with Investigators

Table 11 shows the distribution of cases where the level of interaction between the case analyst and the submitting officer(s) or prosecutor was categorized as "high," "moderate," or "none/minimal" based on criteria previously discussed in *Materials and Methods*. Table 11 also shows the distribution of cases where the level of context information available to the case analyst was categorized as "high," "moderate," or "none/minimal" based on the previously discussed criteria.

TABLE 11 **Distribution of Case Type, Level of Contextual Case Information Supplied to the Analyst, and Level of Interaction Between the Analyst and Investigators**

Level of Interaction	High				Moderate				None/minimal			
Amount of Context Information	High	Mod	None-Minimal	Total	High	Mod	None-Minimal	Total	High	Mod	None-Minimal	Total
Crimes against people	13	0	4	17	10	7	7	24	34	16	38	88
Property Crimes	4	2	8	14	14	11	10	35	141	95	370	606
Drugs	1	1	7	9	0	2	10	12	4	7	47	58
Weapons	0	1	2	3	1	0	0	1	6	1	11	18
Totals for "Level of Interaction"				43				72				770
Totals for "Amount of Context Information"		Number of Cases with High Context Information		228		Number of Cases with Moderate Context Information		143		Number of Cases with No/minimal Context Information		514

Table 11 shows that in 87% of the cases (770 out of 885), there was minimal interaction between the analyst and the investigation. In these cases, evidence was received with a request to process, the analyst performed the examinations, and issued a report, with no communications between the requester and the analyst. It should be recognized, that other agencies may have a routinely different level of interaction with investigators. A smaller police department may have investigators directly handing evidence and interacting with analysts, or possibly the crime scene investigator *is also* the latent print analyst in the case. These data show that most examinations are routine tests with minimal or no interaction between the BCA analysts and investigators.

Table 11 also shows that 58% of cases submitted at BCA have no context/case information provided (514 out of 885), while 26% have a high amount of context/case information provided (228 out of 885) and 16% have a moderate amount of context/case information provided (143 out of 885). Further analysis showed that it was predominantly smaller/rural agencies which were providing more context information and longer, more detailed reports (38% of the time from rural agencies versus 24% from large metropolitan cities).

Two subsets of those data were compared: the cases where there was high context information and high interaction (high context/high interaction; N = 18) versus the cases where there was no context information and no interaction (no context/no interaction; N = 466). The reason for doing so is that it has been asserted that the high context/high interaction cases are essentially where there is the most danger of bias—that the analyst is receiving significant non-domain information and cues from investigators. This is actually a very limited number of cases in the sample (2%). This is in contrast to the 53% of cases with no context information and interaction with investigators.

It can be seen in Table 12 that when comparing no context/no interaction cases against high context/high interaction cases, the most obvious difference is that the high context/high interaction cases produced a disproportionately larger number of "identifiable" latent prints (an average of 6.7 per case versus 1.4 in cases of no context/no interaction). This is explainable given that many of the high context/high interaction are disproportionately crimes against people (and specifically 11 out of 13 are homicides). Homicides, as previously discussed, tend to generate significantly more evidence, and have the highest level of context information and interaction between investigators, prosecutors, and analysts.

Only 25 of the 121 identifiable latent prints resulted in an "identification" decision (a 21% identification rate) in the high context/high interaction cases. It is striking to note that in the no context/no interaction cases, 142 of the 650 identifiable latent prints resulted

TABLE 12 Distribution of "Identification" and "Exclusion" Decisions Sorted by Case Type, Level of Contextual Case Information Supplied to the Analyst, and Level of Interaction Between the Analyst and Investigators

	N	Number of identifiable latent prints	Number of latent prints identified to suspect	Number of exclusion decisions to suspect(s)
No context -No interaction				
All case types	466	650	142	172
Crimes against people	38	44	8	4
Property crimes	370	579	128	151
Drugs	47	22	6	12
Weapons	11	5	0	5
High context – High interaction				
All case types	18	121	25	334
Crimes against people	13	106	20	301
Property crimes	4	7	0	26
Drugs	1	8	5	7
Weapons	0	0	0	0

in an "identification" decision (a 22% identification rate). Essentially there was no difference in the rate of identification between these two subgroups. This is not compelling evidence that analysts are highly motivated to find only evidence to support the police theory and are being influenced by interactions with police and prosecutors (Koppl and Sacks 2013). This is not to say that it could not have happened in any one of these cases, but rather, there is no compelling evidence of such a trend or routine practice.

There was a difference in the rates of exclusions to suspects: there were nearly 3 "exclusion" decisions per latent print for high context/high interaction cases, whereas there was only 1 "exclusion" decision for every 4 latent prints in no context/no interaction cases. Proportionately, there were 12 times as many exclusions of suspects in high context/high interaction cases as there were in no context/no interaction cases. This is likely due to the higher number of suspects against which to compare in homicide cases in the high context/high interaction cases compared to the large number of property crimes, where there is usually no suspect provided about half the time, dominating the no context/no interaction cases.

Another way to look at the above data is that if an agency *was* to decide to shield an analyst from all context information and interaction with the investigators it would not necessarily have a deleterious effect on the number of "identification" decisions. A fair question however is whether the necessary sequential unmasking steps are worth the effort. At an agency like BCA, it would certainly require hiring additional technical staff and changing workflow procedures, writing computer code and creating permissions on who has access to information and how it will be disseminated. If this were done for all sections at the BCA, it would feasibly require at least 5 technically trained staff to manage case information and coordinate cases among bench analysts. This is likely a minimum salary cost (not including benefits) of $250,000 per year. Given current backlogs and a need for faster turn-around time, is this really the highest priority? Those that call for sequential unmasking procedures in all cases have not offered a realistic analysis on the impact on work flow and cost to implement full blinding procedures (Kassin, Dror, and Kukucka 2013). More importantly, no pilot studies have been published showing that testing errors will be decreased with such procedures in place. Before widespread implementation of such procedures, the authors call for research demonstrating that in a complex, high through-put crime laboratory these procedures will have any serious reduction of error. A cost-benefit analysis, with actual data from those who understand the workings of a crime lab, has yet to be offered (5; Kassin, Dror, and Kukucka 2013). Perhaps this money might be better spent on the back-end, limiting which evidence is presented in court and how it may be presented to a jury (for example, using "hot-tubbing" approaches) (Champod and Vuille 2010). Or perhaps this money could be used for expert fees and independent testing to review cases for defense, when there is a dispute of the crime lab's findings. In this vein, Saks,

et al. proposed a *forensic voucher* system (Saks et al. 2001). These *select* cases could then be subjected to a sequential unmasking procedure during an independent review, rather than subject all cases *a priori* to such a labor intensive approach.

Single Latent Print Associations and the Potential for Bias Effects

Given the attention the Brandon Mayfield case has been given, the authors felt it important to investigate the realistic possibilities of the frequency of cases in the 2009/2010 data set that could have "Mayfield-case-like" factors. In the Mayfield case, latent prints recovered from evidence at the scene of a commuter train bombing in Madrid, Spain, on March 11, 2004, were sent to federal agencies around the world. The FBI in the U.S. received these images and searched them in their AFIS database. A single, complex latent print, was erroneously identified to an American named Brandon Mayfield (Stacey 2004). It was the only physical evidence associating Mayfield to the case. The case analysts were exposed to significant contextual information and there were significant interactions between the fingerprint examiners and Spanish officers. However, these interactions between U.S. and Spanish officials and the exposure to extraneous contextual information came *after* the "identification" decision was declared, but the analysts continued to maintain the decision, even in the face of contradictory information. Nonetheless, this case is treated as a poster child for high bias and context effects (National Research Council 2009).

The authors explored how many of the cases with "identification" decisions in the 2009/2010 data set reported a single "identification" decision to a suspect in the case. To be clear, there could have been multiple identifiable latent prints and multiple suspects proffered, but only one of the latent prints in the case was identified to a suspect. Thus the latent print evidence in the case is a single link. This choice has been made because in all of the reported cases of erroneous identifications, it has always been a single erroneous identification decision to an individual. The prevailing theory is that these errors are relatively rare events. The likelihood of one erroneous identification decision being made to a single suspect, and then verified by a second examiner is estimated to be exceedingly low—much less than 0.1% (Ulery et al. 2012). The chance of it happening twice to the same individual with two different latent prints would be significantly smaller. It is the primary reason that SWGFAST and the FBI have both chosen to focus their attention during "blind verification" on single conclusion decisions (Cole 2013, Scientific Working Group on Friction Ridge Analysis Study and Technology [SWGFAST] 2012).

In the 2009/2010 data set, 89 of the 396 "identification" decisions to suspects (see Table 10) were single "identification" decisions to a suspect. When we further examined the assignment of the level of context information and interaction in these 89 cases, we found that only 1 of these cases was "high context/high interaction." In the 885 cases reviewed, a single case had a single identification to a suspect with the analyst being exposed to high level of context information and having a high level of interaction with investigators. It was a homicide case. Figure 7 shows the latent print in this case. Forty-two (42) of the 89 "single ID cases" (47%) had no context information/interaction. The remaining 46 "single ID cases" (52%) had some combination of context information and interaction other than "high/high" or "none/none."

The latent print in Figure 7 shows a relatively non-complex latent palm print. The latent print has a large amount of clear ridge detail, with intermittent areas of distortion. The latent print is in blood and was processed with a dye stain; it exhibits areas of classic blood

FIGURE 7 **The latent print from the only case in the 2009/2010 data set with a single "identification" decision to a suspect and where there is high context/high interaction.**

matrix distortion effects (Langenburg 2008). The authors provided this blood print to five latent print experts, certified by the International Association for Identification (IAI) and asked them to rate the difficulty. All five experts indicated the blood print to be "easy" for comparisons purposes.

It is intriguing that only one case for 12 months of randomly sampled case data met the conditions of "high context/high interaction/single identification to a suspect." Furthermore, the palm print examination in the case is "easy" from the perspective of a fingerprint expert. In fact, based on previous research at BCA (Neumann et al. 2011), the percentage of cases with difficult, marginal latent print examinations is relatively small (<5%). It would appear to be uncommon for a case to have high context, high interaction, a single identification to a suspect, *and also* be of marginal value or a difficult examination. Research to date has shown relatively little error from bias effects for experts and novices when the latent print comparisons are deemed "easy." Errors from bias effects were much more pronounced when the examinations were deemed "difficult" and/or dealt with "exclusion" decisions (Langenburg, Champod, and Wertheim 2009). Again, we make the point, is it necessary to blind all cases when such a small fraction pose any real risk of error? As a more resource friendly option, we could utilize sequential unmasking techniques on this small subset of cases and instances. These cases could be further vetted by identifying them as ideal for review by defense experts, who could then utilize a process of sequential unmasking when reviewing the conclusions of the laboratory.

CONCLUSIONS

The present study and accumulated data sets sampled from four different years at the BCA (2003, 2004, 2009, and 2010) revealed a number of interesting trends. The data are useful for managers to compare laboratory output. They are useful for researchers needing accurate estimates of latent print results from actual casework. They are useful for policy and decision makers to understand the impact that external factors can have on latent print results (e.g., an increase in DNA property crime cases, submission of known or potential suspects, etc.).

We have summarized the major trends observed in the present study as follows:

- From 2003/2004 to 2009/2010, there was an increase in the number (and proportion) of property crime case submissions for latent prints. We theorize this increase may be a "trickle down" effect from increased DNA submissions. Unfortunately for the latent print unit, the personnel and resources have not adjusted accordingly to the increase, thus contributing to the problems of a growing backlog and decreasing morale.
- Just over half of the cases submitted to the BCA-LPU revealed at least one identifiable latent print. Approximately 1 in 4 cases submitted to the BCA-LPU resulted in at least one "identification" result. Approximately 3 in 4 "identification" decisions were to a suspect in the case versus a victim/elimination source. However, approximately half the cases submitted to the BCA-LPU do not have a suspect named by investigators. These cases with no named suspects were predominantly (over 90% of the time) property crimes.
- AFIS was used in about 1 in 3 cases submitted to the BCA-LPU and was used predominantly in property crimes (as noted, due to the lack of provided suspects). A new suspect was generated in about 1 in 3 of the searched cases and had a latent print "hit" rate of 17%. Most cases where a search was required had 1 to 2 latent prints to search in AFIS.
- Approximately 1 in 7 latent prints appeared to originate from a palm (as opposed to a finger or finger joint). This demonstrates the need for palm print databases/exemplars and training on palm prints.
- While only about half of the cases submitted to BCA-LPU had any processing (powder and lift, cyanoacrylate fuming, etc.) done prior to submission, these cases nearly doubled the chance of finding an identifiable latent print. This is an important message to crime scene technicians weighing the risk/benefit of processing the exhibit in the field versus submitting the exhibit to a lab where it may take several months before being processed.
- Non-porous processes (cyanoacrylate fuming followed by dye stain or powder) were the most commonly employed process by the BCA-LPU.
- The recovery rate for identifiable latent prints from plastic bags (submitted mostly in drugs cases) was 13%. The recovery rate for identifiable latent prints from firearms was 13%. No identifiable latent prints were recovered from fired or unfired ammunition in the study.

- The rate of identifiable latent prints that were subsequently identified was approximately the same for property crimes, crimes against people, and drug cases (all around 40%).
- Most cases (87%) submitted to the BCA-LPU have no interaction between the analyst and the investigator in the case. Just over half (58%) of the cases submitted to the BCA-LPU have no case information/context information submitted other than the requested forms that include suspect/victim names, dates of birth, exhibits to be examined, etc. This resulted in half (53%) of the cases having no interaction/no context information, and only 2% of the cases having a high level of interaction/high level of context information exchanged between the forensic analyst and the police investigators.
- The rate of latent print "identification" decisions was the same for identifiable latent prints recovered in cases of no context/no interaction versus cases of high context/high interaction (21% and 22%, respectively). This is not compelling evidence of a trend that forensic analysts are being motivated and influenced by context information or interaction with law enforcement to produce more "identification" decisions to "aid the police."
- Approximately 10% of the BCA-LPU cases resulted in a single latent print "identification" decision to one of the suspects in the case. Half of these cases were classified as no context/no interaction, while only one case in the entire set had a single identification to a suspect in the case under the high context/high interaction condition.

From these findings we draw three conclusions. The first conclusion is that these data were valuable to the BCA-LPU in understanding the basic effectiveness and rate of success for current processes. It is less effective to go to management and say "we need palm print training because we see a lot of palm prints in our cases," versus "palm prints are an integral part of my duties—1 in 7 of the latent prints I examine are palm prints; without training or a database to search, I am not utilizing a large portion of my evidence." Managers and policy makers tend to react to data and dollars versus vague assertions. The data also give the BCA-LPU a baseline performance statistic, so that if we make changes to policy or processes, we will have data against which to compare the effectiveness of the change.

The second conclusion is that, unlike our fictional *CSI* counterparts on television, most case submissions are actually unsuccessful. In half the cases submitted we find no identifiable latent prints, and in the half that we do, only half of ***those*** cases result in an "identification" decision reported by the analyst. Of ***those*** "identification" decisions, three-fourths of the time they are to a suspect in the case. So in effect, only about 1 in 6 cases submitted to the BCA-LPU are returned with what is likely to be a "helpful" result to law enforcement (i.e., a suspect was identified with latent print evidence).

We are hopeful that data in the present paper, and some other similar papers, can be presented by other individuals during testimony, and thus not require an analyst to testify to why latent prints were not found in this case and the absence of identifiable latent prints is common. Especially for a state or federal agency, the travel time and costs can be a resource drain. When waiting time, delays, and continuances are factored in, this can be a serious waste of analyst time and tax dollars. Data such as that reported in this paper can be relayed by the local crime scene technician or an investigator (provided they have some basic forensic experience), thus precluding the need for a lab analyst to appear and give testimony.

The last conclusion is that there was little evidence of a trend for forensic analysts at the BCA-LPU to be biased toward aiding law enforcement from interactions or information exchanged between the fingerprint examiner and police investigators. The fact that the rate of "identification" decisions was identical in the subsets of no context/no interaction and high context/high interaction does not show a tendency for the analyst in those high context/high interaction cases (which tended to be crimes against people, and specifically homicide cases) to push the envelope and either claim more identifiable latent prints or claim more "identification" decisions.

This does not mean that we dispute the inherent dangers of error from bias, nor do we ignore the research that has demonstrated bias effects. We believe that there is usefulness in sequential unmasking or blind verification procedures, but to date, there are no studies that demonstrate such procedures applied to all cases will in effect reduce the number of errors in casework or be cost effective and worth the resources dedicated to instituting a masked workflow. In Langenburg 2012, it was proposed that instituting blind verification

in all cases would lead to more erroneous *exclusions*, and these errors would become a constant drain on resources by constantly performing quality reviews and dealing with corrective action issues.

Based on the data in this study, and still recognizing the obvious concern of error from bias effects, it makes the most sense from a resource standpoint to recommend that if a sequential unmasking approach is to be instituted, then it should be used in the small subset of cases where the effect of bias is most likely to have an impact. For the BCA-LPU, this would represent from as many as 10% of the submitted cases (for *all* cases where there is a single "identification" decision") to as low as 1% (for cases where there is a single *complex* "identification" to a *suspect*). In this way, a much more resource friendly approach could be adopted. BCA-LPU currently has a standard operating procedure (SOP) for "Blind Verification," and this standard captures the essence of the similarly titled SWGFAST standard. The BCA-LPU "Blind Verification" SOP is applied currently exactly as described above, judiciously, when the perceived risk or benefit is sufficient to justify its use.

Limitations and Further Research

The sampled cases were a cross-section of cases for the specified years 2003–2004 and 2009–2010. Personnel, policies, and procedures have changed significantly in the last decade or so in the BCA-LPU. Those cases sampled represented the attitudes and procedures of the day. The cases selected were representative samples of a specific time window in the BCA-LPU. Each case, however, has its own unique set of circumstances, and so while we looked at overall trends in the present study, this does not mean to imply that in one singular case an analyst could have done something different, or been influenced by context information, or made an error, etc. The focus was on general and distinctive trends.

Another limitation is how the cases were categorized and assessed. For example if a case was a burglary where the perpetrator left behind a note bearing racial epithets and threats, is this a property crime or a crime against people? If it was clear, we used the higher potential criminal charge in the case, but often, this information may not be available upon submission so the case is classified as best as possible. When assigning the level of interaction between analyst and investigators or the level of contextual information available in the case, we again, had to make some judgment calls. Typically, we opted for a higher level of interaction/context information if there was any doubt. For example, if the case only had a short note such as "we are looking for subject's prints on the gun," this was designated as a "minimal to no context" case. If the officer wrote (and this would be extraordinary and did not occur in these samples) "we are looking for subject's prints on the gun—we know he did it and he's a bad person who needs to come off the streets," then this would be categorized as a "high context" case even though it is a single, short statement made to the laboratory. While length was a consideration, content of the information was also considered as well. This is where reasonable judgment was exercised, but was subjective nonetheless.

With respect to the effects of context information and interactions with law enforcement, we only compared the two conditions of high context/high interaction versus no context/no interaction. There may be bias effects present in cases that have some combination of context/interaction other than "high context" *and* "high interaction." Given the attention that has been placed on high context and high interaction with law enforcement, this seemed to be a reasonable starting point. Other combinations, can, and should, be explored. Furthermore, we don't know if bias effects from context information are weaker or stronger influences than those influences from interaction with police investigators. Perhaps a "moderate context" case may produce bias effects equivalent to a "high interaction" case. We simply do not know enough about the frequency and impact of bias effects leading to error in forensic casework.

Finally, the data obviously represent the casework, policies, procedures, and personnel of the BCA-LPU. These data may not be representative for agencies of a different size or with different workflows. For some agencies, what constitutes 'a case' may differ dramatically than BCA-LPU. How AFIS searches/cases, technical reviews, cases are documented, exhibits processed, etc. will significantly affect the counts. It is important to carefully consider the BCA-LPU policies, demographics, and workflow that is described in the introduction of this paper before making comparisons to another agency.

The authors envision a few follow-up studies from this. Firstly, one of the authors works in Geneva,

Switzerland. It would be interesting to make some comparisons between the two laboratories. The perception is that the U.S. has so much more crime (specifically violent crime) compared to European countries. It would be interesting to compare data sets between the two laboratories. Secondly, during the study, we identified a list of cases that bore single "identification" and "exclusion" decisions. We could measure the repeatability and reproducibility of those decisions under different context. A set-up similar to Dror, et al (2006) could be employed. Furthermore, we could assign a level of difficulty in advance for each of the decisions and use minutiae counts to predict outcomes based on recent studies (Ulery et al. 2013, Neumann et al. 2013). Lastly, we are interested in more recovery rate data. The 2003/2004 dataset was ten years old and we would prefer to re-examine recovery rates under newer policies and procedures, including the use of Indanedione for porous exhibit processing and the use of digital capture and enhancement.

ACKNOWLEDGMENTS

The authors wish to thank Brenda Hummel, Hamline University. She was the intern who kindly extracted the data from the case information and LIMS. We are thankful for and appreciate the collaborative spirit of the Forensic Laboratory of the Canton Police Geneva, Switzerland to loan the author, Flore Bochet, for a time to perform this research. We wish to also thank the BCA management and BCA-LPU latent print examiners for their willingness to share these data for the benefit of the community. Thank you to the anonymous reviewers and their helpful comments.

REFERENCES

Barnum, C. A. and D. R. Klasey. 1997. Factors Affecting the recovery of latent prints on firearms. *J. Forensic Identification* 47(2): 141–147.

Champod, C. and J. Vuille. 2010. Preuve scientifique en Europe – Admissibilité, appréciation et égalité des armes. Strasbourg, Germany: Conseil de l'Europe - Bureau du Comité européen pour les problèmes criminels (CDPC).

Charlton, D., P. A. F. Fraser-Mackenzie, and I. E. Dror. 2010. Emotional experiences and motivating factors associated with fingerprint analysis. *J. Forensic Sci.* 55(2): 385–393.

Cole, S. A. 2006. The prevalence and potential causes of wrongful conviction by fingerprint evidence. *Golden Gate Univ. Law Rev.* 37(1): 39–105.

Cole, S. A. 2013. Implementing counter-measures against confirmation bias in forensic science. *J. Appl. Res. Memory Cognit.* 2(1): 61–62.

Cook, R., I. W. Evett, G. Jackson, P. J. Jones, and J. A. Lambert. 1998. A model for case assessment and interpretation. *Sci. Justice* 38(3): 151–156.

Cook, R., I. W. Evett, G. Jackson, P. J. Jones, and J. A. Lambert. 1999. Case pre-assessment and review in a two-way transfer case. *Sci. Justice* 39(2): 103–111.

Dhingsa, R., A. Qayyum, F. V. Coakley, Y. Lu, K. D. Jones, M. G. Swanson, P. R. Carroll, H. Hricak, and J. Kurhanewicz. 2004. Prostate cancer localization with endorectal MR imaging and MR spectroscopic imaging: effect of clinical data on reader accuracy. *Radiology* 230(1): 215–220.

Dieltjes, P., R. Mieremet, S. Zuniga, T. Kraaijenbrink, J. Pijpe, and P. de Knijff. 2011. A sensitive method to extract DNA from biological traces present on ammunition for the purpose of genetic profiling. *Int. J. Legal Med.* 125(4): 597–602.

Dror, I. E. 2013. The ambition to be scientific: Human expert performance and objectivity. *Sci. Justice* 53(2):81–82.

Dror, I. E., and D. Charlton. 2006. Why experts make errors. *J. Forensic Identif.* 56(4): 600–616.

Dror, I. E., D. Charlton, and A. E. Péron. 2005. Contextual information renders fingerprint experts vulnerable to making erroneous identifications. *Appl. Cognit. Psychol.* 19(6): 799–809.

Dror, I. E., and G. Hampikian. 2011. Subjectivity and bias in forensic DNA mixture interpretation. *Sci. Justice* 51(4): 204–208.

Durose, M. R. 2008. *Census of publicly funded forensic crime laboratories, 2005*. Washington, DC: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Haber, Lyn, and R. N. Haber. 2008. Scientific validation of fingerprint evidence under Daubert. *Law, Probability and Risk* 7(2): *Sci. Justice* 87–109.

Horsman–Hall, K. M., Y. Orihuela, S. L. Karczynski, A. L. Davis, J. D. Ban, and S. A. Greenspoon. 2009. Development of STR profiles from firearms and fired cartridge cases. *Forensic Sci. Int. Genet.* 3(4): 242–250.

Johnson, S. 2010. Development of latent prints on firearms evidence. *J. Forensic Identif.* 60(2):148–151.

Kassin, S. M., I. E. Dror, and J. Kukucka. 2013. The forensic confirmation bias: Problems, perspectives, and proposed solutions. *J. Appl. Res. Memory Cognit.* 2(1): 42–52.

Kelty, S. F., R. Julian, and A. Ross. 2013. Dismantling the justice silos: Avoiding the pitfalls and reaping the benefits of information-sharing between forensic science, medicine and law. *Forensic Sci. Int.* 230(1–3): 8–15.

Koppl, R., and M. Sacks. 2013. The criminal justice system creates incentives for false convictions. *Criminal Justice Ethics* 32(2): 126–162.

Koppl, R. 2005. How to improve forensic science. *Eur. J. Law Econ.* 20(3): 255–286.

Krane, D. E., S. Ford, J. R. Gilder, K. Inman, A. Jamieson, R. Koppl, I. L. Kornfield, D. M. Risinger, N. Rudin, M. S. Taylor, and W. C. Thompson. 2008. Sequential unmasking: A Means of minimizing observer effects in forensic dna interpretation. *J. Forensic Sci.* 53(4): 1006–1007.

Langenburg, G. 2008. Deposition of bloody friction ridge impressions. *J. Forensic Identif.* 58(3): 355–389.

Langenburg, G. 2012. *A critical analysis and review of the ACE-V process*. Lausanne, France: Ecole des Sciences Criminelles (ESC)-Institut de Police Scientifique (IPS), University of Lausanne.

Langenburg, G., and C. Champod. 2011. The GYRO System: A recommended approach to more transparent documentation. *J. Forensic Identif.* 61(4): 373–384.

Langenburg, G., C. Champod, and P. Wertheim. 2009. Testing for Potential contextual bias effects during the verification stage of the ACE-V methodology when conducting fingerprint comparisons. *J. Forensic Sci.* 54(3): 571–582.

Loy, C.T., and L. Irwig. 2004. Accuracy of diagnostic tests read with and without clinical information: A systematic review. *J. Am. Med. Assoc.* 292(13): 1602–1609.

Maldonado, B. 2012. Study on developoing latent fingerprints on firearm evidence. *J. Forensic Identif.* 62(5): 425–429.

Margot, P. 2011. Forensic science on trial–What is the law of the land? *Aust. J. Forensic Sci.* 43(2–3): 89–103.

Mnookin, J. L. 2010. The courts, the NAS, and the future of forensic science. *Brooklyn Law Rev.* 75(4):1209–1276.

National Research Council. 2009. *Strengthening forensic science in the United States: A path forward*. Washington, D.C.: The National Academies Press.

Neumann, C., C. Champod, M. Yoo, T. Genessay, and G. Langenburg. 2013. *Improving the understanding and the reliability of the concept of "sufficiency" in friction ridge examination*. Washington, DC: U.S. Department of Justice.

Neumann, C., I. Mateos-Garcia, G. Langenburg, M. Schwartz, M. Koolen, and J. Kostroski. 2011. Operational benefits and challenges of the use of fingerprint statistical models: A field study. *Forensic Sci. Int.* 212(1–3):32–46.

Office of the Inspector General (OIG). 2006. *A review of the FBI's handling of the Brandon Mayfield case*. Washington, DC: U.S. Department of Justice.

Peterson, J. L., M. J. Hickman, K. J. Strom, and D. J. Johnson. 2013. Effect of forensic evidence on criminal justice case processing. *J. Forensic Sci.* 58(Suppl 1): S78–90.

Peterson, J. L., and P. Markham. 1995. Crime laboratory proficiency testing results, 1978–1991, II: Resolving questions of common origin. *J. Forensic Sci.* 40(6): 1009–1029.

Potchen, E. J., T. G. Cooper, A. E. Sierra, G. R. Aben, M. J. Potchen, M. G. Potter, and J. E. Siebert. 2000. Measuring performance in chest radiography. *Radiology* 217:456–459.

Potchen, E. J., J. W. Gard, P. Lazar, P. Lahaie, and M. Andary. 1979. The effect of clinical history data on chest film interpretation: direction or distraction. *Invest. Radiol.* 14:404.

Pratt, A. 2012. Fingerprints and firearms. *J. Forensic Identi.* 62(3): 234–242.

Roberts, P., and C. Willmore. 1993. *The role of forensic evidence in criminal proceedings, Royal Commission on Criminal Justice Research Study No. 11*. London, UK: HMSO.

Saks, M., L. Constantine, M. Dolezal, J. Garcia, G. Horton, T. Leavell, M. Levin, J. Muntz, R. Pastor, L. Rivera, J. Stewart, F. Strumpf, C. Titus, and H. VanderHaar. 2001. Model prevention and remedy of Erroneous Convictions Act. *Arizona State Law J.* 33:665–718.

Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST). *Document #14 Standard for the application of blind verification of friction ridge examinations* (11/14/12 ver. 2.0) 2012. http://www.swgfast.org/Documents.html, accessed June 16, 2014.

Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST). *Document #10 Standards for examining friction ridge impressions and resulting conclusions* (04/27/13 ver. 2.0) 2013. http://www.swgfast.org/Documents.html, accessed June 16, 2014.

Stacey, R. B. 2004. A report on the erroneous fingerprint individualization in the Madrid train bombing case. *J. Forensic Identi.* 54(6): 706–718.

Thornton, J. I. 2010. Letter to the editor—-a rejection of "working blind" as a cure for contextual bias. *J. Forensic Sci.* 55(6):1663.

Ulery, B. T., R. A. Hicklin, J. Buscaglia, and M. A. Roberts. 2012. Repeatability and reproducibility of decisions by latent fingerprint examiners. *PLoS One* 7(3): e32800. doi: 10.1371/journal.pone.0032800.

Ulery, B. T., R. A. Hicklin, G. I. Kiebuzinski, M. A. Roberts, and J. Buscaglia. 2013. Understanding the sufficiency of information for latent fingerprint value determinations. *Forensic Sci. Int.* 230(1–3): 99–106.

U.S. Census Bureau. *Annual estimates of the population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2012*. Population Division 2012. http://www.census.gov/popest/data/national/totals/2012/index.html (accessed July 25, 2013).

Wells, Gary L., M. M. Wilford, and L. Smalarz. 2013. Forensic science testing: The forensic filler-control method for controlling contextual bias, estimating error rates, and calibrating analysts' reports. *J. Appl. Res. Memory Cog.* 2(1): 53–55.

Forensic Science Policy & Management: An International Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ufpm20

The Journey or the Destination? Disentangling Process and Outcome in Forensic Identification

Jeff Kukucka[a]
[a] Towson University, Towson, Maryland
Published online: 08 Dec 2014.

To cite this article: Jeff Kukucka (2014) The Journey or the Destination? Disentangling Process and Outcome in Forensic Identification, Forensic Science Policy & Management: An International Journal, 5:3-4, 112-114, DOI: 10.1080/19409044.2014.966928

To link to this article: http://dx.doi.org/10.1080/19409044.2014.966928

Forensic Science Policy & Management, 5(3–4):112–114, 2014

ISSN: 1940-9044 print / 1940-9036 online
DOI: 10.1080/19409044.2014.966928

The Journey or the Destination? Disentangling Process and Outcome in Forensic Identification

Jeff Kukucka
Towson University, Towson, Maryland

Recently, this journal published an archival study by Langenburg, Bochet, and Ford (2014) that explored the impact of two sources of contextual bias—namely, exposure to case information and interaction with investigators—on the identification decisions of certified latent print examiners. This paper is commendable as an effort to examine the effects of bias on real-world casework. However, it is also limited by several methodological oversights and ultimately perpetuates a common misconception about contextual bias. Below, I briefly summarize this study before elaborating my concerns with its conclusions.

The study included a sample of 885 criminal cases in which law enforcement had solicited the help of examiners from an accredited latent fingerprint laboratory. First, the authors categorized each case as either high, moderate, or low in (a) the degree of interaction between the examiner and investigators, and (b) the amount of case information available to the examiner. Then, they compared identification decisions from 466 cases that were deemed "low" in both dimensions against 18 cases that were deemed "high" in both, and found that the rates of identification were nearly identical (22% and 21%, respectively). From this, the authors concluded that the aforementioned contextual factors had no appreciable biasing effect on examiners' identification decisions.

This conclusion is misleading insofar as it is derived solely from the raw number of identification decisions, and not from any index of their validity, origin, or strength. As others have noted (e.g., Dror 2009), we must be careful to distinguish between the decision-making *process* and the decision *outcome*. Even if bias fails to change the latter, it may nonetheless affect the former. For example, if the forensic evidence suggests a certain conclusion and contextual factors encourage the same conclusion, the examiner's decision will not change, but his or her confidence in this decision may increase as a result. Unfortunately, the data from this study tell us nothing about the process by which examiners arrived at their decisions, and thus it remains unknown whether and how context impacted them.

Even if we ignore the importance of process as the authors have done, their conclusion with respect to outcomes is also unsound. Their inference that context had no effect on identification decisions is based on a comparison of cases with low versus high degrees of bias. This assumes that bias affects decisions in only one direction, which is a dubious assumption. Instead, contextual factors may bias examiners toward identification in some cases, and bias them toward

Address correspondence to Jeff Kukucka, Towson University, Towson, MD 21252. E-mail: jkukucka@towson.edu

non-identification in others. Collapsing the cases together and counting the overall rate of identification decisions will not reveal biases operating in offsetting directions, and may show–as their study did–that the overall rate is unaffected.

Moreover, the authors' decision to compare only those cases that were deemed "low" versus "high" in *both* contextual factors was puzzling. There is no reason to believe that bias arises only when both factors are present at high levels. (Indeed, some, e.g., Whitman and Koppl (2010), have argued that bias exists even when there is zero communication with investigators.) Even at "moderate" levels, each factor may be a sufficient–but not necessary–condition to produce bias. The authors agree that "other combinations can and should be explored" and that the effects of the two contextual factors may differ in magnitude (see p. 35), but they stop short of providing any such exploration. One wonders why, with a provocative hypothesis and access to troves of data (i.e., 885 cases), the authors were content to focus on a comparison group that utilized only 18 (2.03%) of the cases at their disposal.

Perhaps this is a moot point, given that the manner in which the study categorized cases as high, moderate, or low in bias was blatantly unscientific. The authors openly admit that their categorization scheme was "subjective" and relied on "a judgment call of the researchers" (see p. 35; p. 18), which inspires little confidence in its validity. As such, these dubious categorizations inspire little confidence in the conclusions that are later derived from them. Instead, multiple independent raters should have made these judgments, so as to permit the calculation and reporting of inter-rater reliability. It is rather ironic that the authors tout the value of having identification decisions verified by a second examiner, but do not bother to adopt an analogous practice in their own research.

In any event, the archival nature of their data precludes any conclusions with respect to causality. To properly test whether bias affects decision outcomes, decisions made in the presence of biasing factors must be compared against those of a control group that makes the same decisions in a context known to be completely devoid of biasing factors. This critical control condition is missing in this study, yet the authors conclude with certainty that bias had no effect on decision outcomes.

In contrast, decades of psychological research unequivocally show that perceptual judgments across many domains–including forensic science–are sensitive to context (see Kassin, Dror, and Kukucka 2013). This should not be taken to suggest that forensic examiners who are susceptible to bias are unskilled or careless; rather, contextual bias is an inherent and unconscious feature of human psychology. For example, in one study (Dror and Charlton 2006), experienced fingerprint examiners unknowingly changed 17% of their own prior identification decisions after being given case information that implied the guilt or innocence of the suspect. Notably, given that Dror and Charlton integrated these judgments into examiners' quotidian casework, their findings cannot easily be dismissed as the product of "contrived research" (Langenburg, Bochet, and Ford 2014, p. 16).

To safeguard against bias, medical researchers have long demanded the use of double-blind placebo-controlled studies in which both doctors and patients are deliberately kept uninformed as to the experimental conditions. Similarly, psychology researchers strive to keep experimenters blind to a study's hypotheses when possible (see Rosenthal 1966). To similarly minimize the risk of bias among forensic examiners, I join others in supporting the adoption of *sequential unmasking* protocols (see Krane et al. 2008), which shield examiners from irrelevant case information until the critical stages of the decision-making process are complete. The authors of the current study recognize the value of sequential unmasking, but also raise two oft-heard objections.

First, they argue that its universal implementation would be costly–both financially (see also Charlton, 2013) and in terms of efficiency–and instead propose that it be applied only to cases where the risk of bias is high. However, they fail to clarify how and by whom such "high-risk" cases would be identified, and how this screening process would prove less costly. On the contrary, encouraging reports from forensic laboratories that have adopted sequential unmasking as standard practice (Found & Ganas 2013; Stoel, Dror, & Miller 2014) suggest that the protocol has been neither onerous nor expensive to implement, and has yielded more benefits than costs.

Second, the authors speculate that, "Case information could help the analyst make more accurate,

efficient, and informed decisions" (p. 16; see also Elaad 2013]. This argument is fundamentally misguided. As others have explained, e.g., Page, Taylor, and Blenkin (2012), the proper role of a forensic examiner is to produce a judgment that is independent and circumscribed to the forensic evidence at hand. Exposure to extraneous case information compromises the independence of this judgment, making it unclear whether an accurate judgment is the product of the examiner's unique expertise or of contextual happenstance. When examiners are shielded from case information, legal fact-finders can rest assured that their judgments are the product of the forensic evidence at hand, and nothing else (Dror, Kassin, and Kukucka 2013).

Simply put, contextual influences can unwittingly lead forensic examiners to the right decision, but for the wrong reasons. By focusing exclusively on the *outcomes* of their analyses, the current study neglects the fact that bias can likewise distort the *process* of the analysis, even if it fails to change the outcome. In light of this, as well as the methodological issues enumerated above, this study offers little information about the true scope of the problem. In order to best direct our efforts to combat contextual bias, it will be crucial to develop a more nuanced understanding of how context impacts both the outcome and process of forensic identification.

REFERENCES

Charlton, D. 2013. Standards to avoid bias in fingerprint examination: Are such standards doomed to be based on fiscal expediency? *Journal of Applied Research in Memory and Cognition* 2:71–72.

Dror, I. E. 2009. On proper research and understanding of the interplay between bias and decision outcomes. *Forensic Science International* 191:e17–e18.

Dror, I. E., and D. Charlton. 2006. Why experts make errors. *Journal of Forensic Identification* 56:600–616.

Dror, I. E., S. M. Kassin, and J. Kukucka. 2013. New application of psychology to law: Improving forensic evidence and expert witness contributions. *Journal of Applied Research in Memory and Cognition*, 2:78–81.

Elaad, E. 2013. Psychological contamination in forensic decisions. *Journal of Applied Research in Memory and Cognition* 2:76–77.

Found, B., and J. Ganas. 2013. The management of domain irrelevant context information in forensic handwriting examination casework. *Science and Justice* 53:154–158.

Kassin, S. M., I. E. Dror, and J. Kukucka. 2013. The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition* 2: 42–52.

Krane, D. E., S. Ford, J. R. Gilder, K. Inman, A. Jamieson, R. Koppl, I. L. Kornfield, et al. 2008. Sequential unmasking: A means of minimizing observer effects in forensic DNA interpretation. *Journal of Forensic Sciences* 53:1006–1007.

Langenburg, G., F. Bochet, and S. Ford. 2014. A report of statistics from latent print casework. *Forensic Science Policy & Management* 5:15–37.

Page, M., J. Taylor, and M. Blenkin. 2012. Context effects and observer bias: Implications for forensic odontology. *Journal of Forensic Sciences* 57:108–112.

Rosenthal, R. 1966. *Experimenter effects in behavioral research*. New York, NY: Appleton-Century-Crofts.

Stoel, R. D., I. E. Dror, and L. S. Miller. 2014. Bias among forensic document examiners: Still a need for procedural change. *Australian Journal of Forensic Sciences* 46:91–97.

Whitman, G., and R. Koppl. 2010. Rational bias in forensic science. *Law, Probability, & Risk* 9:69–90.

Do Observer Effects Matter? A Comment on Langenburg, Bochet, and Ford

Journal:	*Forensic Science Policy & Management: An International Journal*
Manuscript ID:	UFPM-2014-0014.R1
Manuscript Type:	Article
Date Submitted by the Author:	30-Nov-2014
Complete List of Authors:	Koppl, Roger; Syracuse University, Charlton, David; Surrey and Sussex Forensic Services, Kornfield, Irving; University of Maine, Krane, Dan; Wright State University, Risinger, Michael; Seton Hall University, Robertson, Christopher; University of Arizona, Saks, Michael; Arizona State University, Thompson, William; University of California Irvine,
Keywords:	fingerprints < analytical sections, process improvement, quality issues, error rates < methods and practices, bias < methods and practices

Do Observer Effects Matter? A Comment on Langenburg, Bochet, and Ford*

ABSTRACT

We identify methodological problems in Langenburg et al. (2014), which undermine its conclusions about the size of the observer effect problem and the importance of sequential unmasking as a solution. The scoring method of Langenburg et al. (2014) appears to be subjective. The classification of cases is not congruent with the three keys to observer effects in forensic science: the analyst's state of expectation, the analyst's state of desire, and the degree of ambiguity in the evidence being examined. Nor does the paper adequately support its claim, "[I]t has been asserted that the high context/high interaction cases are essentially where there is the most danger of bias." While the paper tends to minimize concern over observer effects, the evidence in it seems to support the view that fingerprint analysts look to contextual information to help them make decisions.

Do Observer Effect Matter?

Langenburg et al. (2014) have provided a service by presenting data on latent print work at the Minnesota Bureau of Criminal Apprehension. They uncover potentially useful facts such as the rate at which latents were recovered from plastic bags. Thus, the paper succeeds in its stated goal of providing "casework statistics" (p. 16) from Minnesota's Bureau of Criminal Apprehension Latent Print Unit (BCA-LPU). Additionally, "A portion of the . . . paper was dedicated the exploration of possible bias effects from significant interaction between the forensic analyst and the case investigator, or from analyst exposure to contextual information about the case" (p. 16). Langenburg et al. drew several conclusions in this later part of their paper, including "there is usefulness in sequential unmasking." We commend this affirmation of a role for sequential unmasking, even though Langenburg et al. see the potential as more limited than we do.

In spite of this point of agreement, we believe that the portion of the paper on bias was flawed and cannot be used to infer that the risk of observer effects is lower in fingerprint analysis than in other expert domains. The study is of limited value, therefore, in judging the proper scope for sequential unmasking or other measures that are meant to limit bias or its bad effects. The paper suffers from some basic methodological flaws and, consequently, it may not be possible to draw policy-relevant inferences from its analysis.

Langenburg et al. distinguish between "interactions" and "contextual information" as sources of bias. They use a two-dimensional classification in which cases are rated on the "level of interaction" of the analyst with case investigators and on the "amount" of domain-irrelevant context information the analyst was exposed to. The level of interaction might be "high," "moderate," or "none/minimal," and the amount of domain-irrelevant context information might be "high," "moderate," or "none/minimal." Langenburg et al. seem to think that the risk of a bias-induced error grows as we move from none/minimal to high on either dimension. Thus, they

reason, a bias-induced error is most likely in cases with high interaction and high context information, and error is least likely in cases with none/minimal interaction and none/minimal context information. They say, "Two subsets of those data were compared: the cases where there was high context information and high interaction (high context/high interaction; N D 18) versus the cases where there was no context information and no interaction (no context/no interaction; N D 466). The reason for doing so is that it has been asserted that the high context/ high interaction cases are essentially where there is the most danger of bias—that the analyst is receiving significant non-domain information and cues from investigators" (p. 30).

As a bottom line, the authors report that 142 of 650 (22%) "no context/no interaction" cases resulted in an individualization, whereas 25 of 121 (21%) "high-context / high interaction" cases resulted in an individualization (pp. 30-31). Langenburg et al. infer, "Essentially, there was no difference in the rate of identification between these two subgroups." But the 95% confidence interval for the first rate minus the second rate is about -7% to 8%. (Our calculation assumes that we may model the underlying processes as independent Bernoulli trials.) While, they could correctly point out that this result shows that we cannot rule out the null hypothesis of equal rates between the two groups, their evidence does not rule out substantial observation effects. Yet, there are more fundamental problems. First, there is a methodological question of scoring. Although the authors acknowledge that the scoring of context and interaction was a "judgment call," they do not report whether raters were blinded to the dependent variables under study when they rendered those judgments. Nor do the authors report any measures of inter-rater or intra-rater reliability for these assessments. It is thus hard to know if these measures have any validity, even if the underlying concepts were clear.

Second, there is a problem of "confounding." As Vandenbroucke (2002) explains, the word "confounding" is generally "used for one particular form of the confusion of two effects: the confusion due to extraneous causes, i.e., other factors that really do influence" the processes under study (p. 219). By the authors' own account, the paper's "high context/high interaction" cases have a disproportionate number of homicides, while the "no context/no interaction" cases have a disproportionate number of property crimes. To compare their individualization rates thus seems an apples and oranges comparison.

One might, perhaps, use multivariate regressions to control for such confounds. Or one might attempt some sort of matching or subset analysis. We wonder, however, whether the dataset used is big enough to allow the fruitful use of such techniques. Even then, one might worry about an unavoidable risk of surveys of the type made by Langenburg et al.: omitted variables that correlate with the observed variables and thus skew the analysis.

Such considerations support the view that controlled experiments may be the best way to get at the issues in question. There have been such experiments using trained, working, professional fingerprint examiners (Dror and Charlton 2006, Dror, Charlton, and Peron 2006). These studies

support the view that fingerprint examiners, like all other humans, are subject to observer effects. The effect sizes in these studies are large enough to make bias by domain-irrelevant information a source of concern for anyone interested in avoiding false convictions.

Langenburg et al. find that the rate of identification is about the same in the cases classified as high context/high interaction and those classified as no context/no interaction. Theysay, "This is not compelling evidence that analysts are highly motivated to find only evidence to support the police theory and are being influenced by interactions with police and prosecutors" (p. 31). But by their own account, as we have noted, the cases dubbed "high context/high interaction" include a disproportionate number of homicides and those dubbed "no context/no interaction" include a disproportionate number of property crimes. They say,

> There was a difference in the rates of exclusions to suspects: there were nearly 3 "exclusion" decisions per latent print for high context/high interaction cases, whereas there was only 1 "exclusion" decision for every 4 latent prints in no context/no interaction cases. Proportionately, there were 12 times as many exclusions of suspects in high context/high interaction cases as there were in no context/no interaction cases. This is likely due to the higher number of suspects against which to compare in homicide cases in the high context/high interaction cases compared to the large number of property crimes, where there is usually no suspect pro- vided about half the time, dominating the no context/ no interaction cases (p. 31).

Thus, Langenburg et al. seem to have found that in cases such as murder we usually have multiple suspects and a larger numbers of latents to consider. In those cases, they report, examiners pay lots of attention to the case file and have frequent interactions with police investigators. By contrast, in the typical property crime, there are no suspects and only a few latents to consider. In the first class more latents are judged to be "identifiable" than in the second class. We also get more exclusions in the first class than in the second class. Of the latents judged useable in each class the ratio of identifications is about the same.

All of this is supposed to support the view that bias is somehow a smaller problem than advocates of sequential unmasking believe. But their own evidence says that when examiners have a greater chance of making an identification that might later be shown to be incorrect, they are more likely to interact with police investigators and to acquire case information. This result suggests to us that fingerprint analysts look to contextual information to help them make decisions.

Although the foregoing is sufficient to understand why the Langenburg et al. paper is not contrary to the emerging consensus that observer effects are a real and substantial problem, it is also important to address several foundational methodological and conceptual issues for the sake of future research and to advance the literature on observer effects in forensic science. Fundamentally, we think that classification scheme is inadequate for their stated purpose. It does not adequately reflect the three keys to observer effects in forensic science: the analyst's

state of expectation, the analyst's state of desire, and the degree of ambiguity in the evidence being examined. Krane et al. (2008) say, "Observer effects are rooted in the universal human tendency to interpret data in a manner consistent with one's expectations. This tendency is particularly likely to distort the results of a scientific test when the underlying data are ambiguous and the scientist is exposed to domain-irrelevant information that engages emotions or desires" (p. 1006).

Risinger et al. (2002, p. 12) explain, "At the most general level, observer effects are errors of apprehension, recording, recall, computation, or interpretation that result from some trait or state of the observer." The relevant "state of the observer" may be a state of expectation. Domain-irrelevant information can create an expectation that a given pair of known and unknown prints have a common source or that they do not have a common source. The relevant "state of the observer" may be a state of desire. As Risinger et al. (2002, p. 24) note, "[W]here an observer has strong motivation to see something, perhaps a motivation springing from hope or anger, reinforced by role-defined desires, that something has an increased likelihood of being 'seen.'" Risinger et al. (2002) also note the importance of ambiguity. "Of course, where the evidence is clear, the cognitive biases, which operate best on ambiguity, can be overridden. Conversely, observer effects are most potent where ambiguity is greatest, when an observer's judgment is most likely to succumb to expectation, subjective preference, or external utility" (Risinger et al. 2002, p. 16). Whitman and Koppl (2010) have a model of Bayesian decision-making in which ambiguity increases the chance of a bias-induced error. Expectation, desire ("subjective preference or external utility") and ambiguity are the three key factors producing observer effects in forensic-science decision-making.

Langeburg et al. do not provide citations that support their statement, "it has been asserted that the high context/high interaction cases are essentially where there is the most danger of bias" (p. 30, emphasis added). When first drawing the distinction between interactions and context information they cite three works authored or co-authored by Itiel Dror (Kassin, Dror, and Kukucka 2013; Dror 2013; Dror and Hampikian 2011). But the word "interaction" appears in only one of the cited works, and then only in the phrase "social-interactional context," which was used once, in the description of someone else's study of confirmation bias (Kassim, Dror, and Kukucka, p. 44). At least one important article, however, does say that "interactions" between examiners and investigators create a risk of observer effects. Risinger et al. (2002) note the danger of such interactions and once use that word in that connection (p. 37). In elaborating on the problem, Risinger et al. quote Evan Hodge.

> [The examiner] gave in to investigative pressure. We all do this (give in to investigative pressure) to one extent or another. A hot case comes in, the investigators want to wait, want to look over your shoulder, want to see the ident, help you shoot the gun, etc. Do you take shortcuts? Do the words "the commissioner, or the director, or the captain wants to know right now" affect you? Of course they do, don't kid yourself (Hodge 1988, p. 292 as quoted in Risinger et al. 2002, p. 38).

Hodge's article has been cited elsewhere in the literature on observer effects in forensic science (Koppl 2005, p. 261; Kelly and Wearne 1998, p. 17). It is therefore true that "interactions" between examiners and investigators has been an object of concern. But this concern needs to be understood in the context of the scientific literature on observer effects. The root concern is not "interaction," per se, but the ambiguity of the evidence and an analyst's states of expectation and desire.

An interaction between an analyst and an investigator may produce a change in the analyst's state of expectation, her state of desire, or both. In this sense, the category "interaction" is a mélange of expectation and desire. An interaction may produce only a change in expectation if the analyst is already motivated to support the police theory. It may produce only a change in desire if the analyst, for example, feels the sort of social pressure Hodge (1988) warned of. Finally, of course, an interaction may produce a change in both the analyst's state of desire and her state of expectation. Similarly, "context" as defined by Langenburg et al. may change an analyst's state of expectation, state of desire, or both.

We are not aware of anything in the literature on observer effects in forensic science that asserts that the combination of "high context" with "high interaction" is "essentially where there is the most danger of bias" (p. 30). Rather, the key factors are expectation, desire, and ambiguity – not interaction and context. As we have seen, the sequential unmasking letter of Krane et al. (2008) does suggest that a combination of "expectation" and "desire" may be more likely to generate a bias-induced error than either in isolation. Thus, this combination has been viewed with particular concern in the literature on observer effects in forensic science. But if "context" and "interaction" as defined by Langenburg et al. are both mélanges of "expectation" and "desire" as defined in the literature on observer effects, then it seems doubtful what causal significance should be imputed to the combination of "context" and "interaction."
Langenburg et al. measure "interaction" principally by the number of communications between an analyst and an investigator. To be categorized as "high interaction" the record must reveal "at least 3 phone calls, at least 3 email exchanges, or attendance at the crime scene" (p. 18). While these factors may be generally correlated with an analyst's state of desire, state of expectation, or both, it must also be considered that a single intense interaction could easily have a greater effect than countless insubstantial ones.

Langenburg et al. measure "context" by the number of "case details" to which the analyst was exposed. But the number of details does not have to be high to induce a high state of expectation in an analyst. Some of the language used by Langenburg et al. suggests that they considered not only the number of details, but also whether the details given were "significant" (p. 19). Unfortunately, context was uniformly "high" in the one example they give to illustrate the difference between high and low context cases. They say,

> Typically, we opted for a higher level of interaction/context information if there was any doubt. For example, if the case only had a short note such as "we are looking for subject's prints on the gun," this was designated as a "minimal to no context" case. If the officer wrote (and this would be extraordinary and did not occur in these samples) "we are looking for subject's prints on the gun— we know he did it and he's a bad person who needs to come off the streets," then this would be categorized as a "high context" case even though it is a single, short statement made to the laboratory (p. 22).

In a private communication, Langenburg has explained to us that police investigators must include an "evidence submission form" with all forensic evidence sent to the crime lab. This form includes the names and birth dates of both victims and any suspects. If there is no suspect, as when drugs are found at the side of the road, the lab may not take the case in. Analysts will generally be aware of the information on the submission form because they are required tocheck it for possible errors. Langenburg has explained to us that these completed forms were not generally considered a source of potentially biasing information in the study. An egregious remark in the context box could cause the case to be classified differently, but not remarks indicating that results are needed by a certain date or to look for latent prints first on this object, then on the other, and so on. In our experience, the information routinely supplied in evidence submission forms will often, indeed usually, contain potentially biasing information. Indeed, in the very case used by Langenburg as an exemplar of "minimal to no context" the analyst has been told, as it were, the "right" answer. Certainly, the known and unknown prints must be considered together at some point. But the principles behind sequential unmasking as articulated by Krane et al. (2008), suggest that the crime-scene latent be submitted for characterization before any known prints are available to the analyst. Analysts should first determine whether the detail observed is sufficient to make the crime scene item a potential source of useful information before any known prints are available to her. (This proposal addresses the initial "analysis" stage of the ACE-V method of fingerprint examination.)
The general literature on observer effects (reviewed in Risinger et al. 2002) as well as the literature specific to fingerprint analysis (Dror and Charlton 2006, Dror, Charlton, and Peron 2006) together suggest that the simultaneous presentation of the latent and one known print creates a risk of error through observer effects. Thus, the example of "minimal to no context" given by Langenburg et al. is, instead, a case in which there is an appreciable risk of bias. As Whitman and Koppl (2010, p.70) have noted, "The authorities—police and prosecutors—implicitly convey information to forensic examiners by their very decision to submit samples for testing." Saying the suspect is a bad person is not always necessary to create a state of expectation in an analyst.

Let us consider the analysts' state of desire. The fingerprint examiners of the BCA-LPU are employed in a law-enforcement agency. This alone has the potential to create a desire to help law enforcement officers, and this desire, in turn, has the potential to bias decision-making. As the National Academy of Sciences has noted, "Forensic scientists who sit administratively in law enforcement agencies or prosecutors' offices, or who are hired by those units, are subject to a

general risk of bias" (NAS p. 6-2). The administrative position of fingerprint examiners in the BCA-LPU is invariant across all cases studied by Langenburg et al. It is thus possible, even likely, that these examiners generally had a state of desire capable of producing observer effects. It seems questionable, then, whether the analyst's state of desire is highly correlated with either "context" or "interaction" as defined by Langenburg et al., even though one or more case details or pressure from an investigator could enhance an analyst's state of desire (Charlton et al. 2010).

Let us now consider the analysts' state of expectation. If the analysts are typically presented the known and unknown prints together, then they will typically have at least some degree of expectation that one or more latent prints have the same sourceas a print from the "subject." Thus, the analyst's state of expectation, like her state of desire, may not be highly correlated with either "context" or "interaction" as defined by Langenburg et al., even though one or more case details or an interaction with an investigator could enhance an analyst's state of expectation.

Importantly, the classification employed by Langenburg et al. does not appear to consider in any way the ambiguity of the latent prints submitted for evaluation. Dror et al. (2005) distinguish the "bottom up" information given by the latent or other evidence from domain-irrelevant information, which they dub "top down." They say, "weakening the bottom-up information may allow the top-down component more room to influence the process" (p. 803). When an evidence sample is unambiguous, context information is less likely to induce an error.

In the end, fingerprint analysts must decide when to declare an individualization, when to declare an exclusion, and when to declare that no reliable judgment can be made. Presumably, analysts want to moderate what we might call "reversal risk," the risk that a decision will later be determined to have been mistaken. Charlton et al. (2010) conducted a survey of fingerprint examiners and concluded in part, "[T]here was an expression of fear and consequence in making an erroneous match" (p. 391). They also found "a desire to avoid ambiguity" (p. 390). Domain-irrelevant information may help them reduce subjective doubt about which decision has the lowest reversal risk. If fingerprint examiners are decision makers who are similar to decision makers in other areas, including other expert domains, then the potential of domain-irrelevant information to help resolve ambiguity and subjective doubt may lead them to more energetically seek out and respond to domain-irrelevant information when reversal risk is greater. From this point of view, the differences Langenburg et al. found between "high context/high interaction" cases and "no context/no interaction" cases seem consistent with the view that domain-irrelevant information may be creating observer effects in the BCA-LPU.

REFERENCES

Charlton, D., P. Fraser-Mackenzie, and I. E. Dror. 2010. Emotional experiences and motivating factors associated with fingerprint analysis. Journal of Forensics Sciences, 55 (2): 383-393.

Dror, I. E. 2013. The ambition to be scientific: Human expert performance and objectivity. Sci. Justice 53(2):81–82.

Dror, I. E. and D. Charlton. 2006. Why experts make errors. Journal of Forensic Identification 56: 600–616.

Dror, I. E., D. Charlton, and A. Peron. 2006. Contextual information renders experts vulnerable to making erroneous identifications. Forensic Science International, 156, 174–178. http://dx.doi.org/10.1016/j.forsciint.2005.10.017

Dror, I. E., and G. Hampikian. 2011. Subjectivity and bias in forensic DNA mixture interpretation. Sci. Justice 51(4): 204–208.

Hodge, E. 1988. Guarding against error. Ass'n Firearms & Toolmark Examiners' J. 20: 290-293.

Kassin, S. M., I. E. Dror, and J. Kukucka. 2013. The forensic confirmation bias: Problems, perspectives, and proposed solutions. J. Appl. Res. Memory Cognit. 2(1): 42–52.

Kelly, J. F. and P. Wearne. 1998. Tainting Evidence: Inside the Scandals at the FBI Crime Lab, New York: The Free Press.

Langenburg, G., F. Bochet, and S. Ford. 2014. A report of statistics from latent print casework. Forensic Science Policy & Management, 5(1-2): 15-37.

Risinger, M., M. J. Saks, W. C. Thompson, and R. Rosenthal. 2002. The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. California Law Review, 90(1): 1–56.

Vandenbroucke, J. P. 2004. The history of counfounding. Soz Praventiv Med, 47(4): 216-224.

Whitman, D. G. and R. Koppl. 2010. Rational bias in forensicscience. Law, Prob. & Risk, 9(1): 69-90.

* We thank two anonymous referees for helpful comments.

ANDREW SULNER

Andrew Sulner is a board certified forensic document examiner and attorney who earned a Master of Science degree in Forensic Science and a Juris Doctorate degree (with Honors) from George Washington University. The Sulner name is associated with three generations of document examiners, and Andrew Sulner has over 30 years of experience in examining questioned and disputed documents on behalf of major law firms, banks, insurance companies and financial institutions, as well as federal and state law enforcement and regulatory agencies. Mr. Sulner, who is also a certified fraud examiner and former state prosecutor, has been consulted nationally and internationally as an expert in determining the authenticity of documents, and his testimony as a forensic document examiner has been favorably cited in numerous federal and state court decisions. Mr. Sulner's state of the art forensic document laboratory is located in the heart of New York City.

In addition to being a Diplomate of the Board of Forensic Document Examiners (BFDE), Mr. Sulner is a Fellow of the American Academy of Forensic Sciences (AAFS) and a member of: the Association of Forensic Document Examiners (AFDE); the Association of Certified Fraud Examiners (ACFE); the National District Attorneys Association (NDAA); the American Bar Association (ABA); the American Society of Testing and Materials International (ASTM); and the New York, Florida, California and District of Columbia Bars. Mr. Sulner is the immediate past Chair of the Jurisprudence Section of the AAFS, and currently serves as Vice-President of the BFDE and Chair of the BFDE Ethics Committee; he is a past President of the Society of Medical Jurisprudence and Forensic Sciences. Mr. Sulner continues to serve the forensic science community as an active member of ASTM's Committee E-30 on Forensic Sciences, and as a member of the Editorial Board of the *Journal of Forensic Document Examination*.

Mr. Sulner has authored professional publications and presentations on the subject of forensic document examination, and he is a frequent speaker and lecturer at universities and continuing education seminars sponsored by forensic science, legal, and other professional membership associations throughout the United States. He has also appeared as a forensic document examiner on many television shows, including ABC's "20/20" and news shows aired by CBS, NBC, FoxTV, and others.

Handwriting: Cognitive Bias

Andrew Sulner

Published Online: 11 DEC 2014

DOI: 10.1002/9780470061589.fsa1120

Book Title

Wiley Encyclopedia of Forensic Science

Additional Information (Show All)

How to Cite | Author Information | Publication History

How to Cite

Sulner, A. 2014. Handwriting: Cognitive Bias. Wiley Encyclopedia of Forensic Science. 1–15.

Abstract

Although the forensic science community has long been cognizant of the need to avoid physical contamination of evidence, it has been reluctant to acknowledge the possibility of mental contamination of evidence in the form of cognitively biased forensic evaluations. As highlighted in the 2009 National Academies of Science report, *Strengthening Forensic Science in the U.S: a Path Forward* , empirical research in the fields of behavioral science and information obtained from reviews of forensic practitioner errors in several high-profile cases have clearly established the adverse impact that contextual and motivational biases can have on human judgment and the accuracy of forensic evaluations of evidence. Nevertheless, far too many forensic handwriting experts still steadfastly believe that proper training and experience somehow shield them from the biasing influences that have been proven to impact the accuracy of visual observations and decision making on the part of human beings in all other walks of life. Understanding the various sources of bias and learning how to limit or minimize their influence is essential for improving the accuracy of decisions made by forensic handwriting experts and the reliability of their expert testimony.

In this article, Andrew Sulner, a third-generation board certified forensic document examiner and former state prosecutor, explains and illustrates how the opinions and expert testimony of individuals performing comparative analyses of signatures and handwriting are susceptible to biasing influences that can improperly taint and sway their decision-making process and the manner in which they testify as experts. Actual case histories are used to demonstrate how cognitive or motivational bias can contribute to erroneous findings and/or disingenuous testimony on the part of experienced and presumptively well-trained forensic document examiners. Individual debiasing techniques and institutional context management and evidence testing protocols for minimizing examiner bias in forensic handwriting investigations are also discussed.

Keywords: confirmation bias; context effects; debiasing techniques; examiner bias; Fischhof Method; flawed forensics; forensic bias; forensic document examination; handwriting comparisons; signature comparisons

Author Contact Address: Forensic Document Examinations, LLC
220 East 57th Street, Suite 200, New York, NY 10022
Email: andysulner@aol.com

Sulner, A., (2014). Handwriting: Cognitive Bias, *Wiley Encyclopedia of Forensic Science*, Jamieson, A., Moenssens, A. (eds). John Wiley & Sons Ltd., Chichester, UK, pp 1-15

Handwriting: Cognitive Bias

Introduction

Many disciplines in forensic science require experts to make subjective judgments about whether two things are sufficiently similar to conclude that both originate from the same source. Although the forensic science community has long been cognizant of the need to avoid physical contamination of evidence, it has been reluctant to acknowledge the possibility of mental contamination of evidence in the form of cognitively biased forensic evaluations. It preferred to operate under the belief that proper training and experience somehow shield forensic examiners from the biasing influences that have been proven to impact the accuracy of visual observations and decision making on the part of human beings in all other walks of life. This naive and ill-founded belief, debunked by a vast body of empirical research in the fields of behavioral science as well as data obtained from DNA exoneration cases and reviews of forensic practitioner errors in several high-profile cases, was singled out in the 2009 National Academies of Science report, *Strengthening Forensic Science in the U.S*: *a Path Forward* (NAS Report) [1], which emphasized the adverse impact that contextual and motivational biases can have on human judgment and the accuracy of forensic evaluations of evidence.

Susceptibility to cognitive bias is not a character flaw; it results from imperfections inherent in human perception and reasoning, and as such, cannot be eliminated by sheer force of will. The human mind is capable of *unconsciously* leading even an honest and well-intentioned individual to act in a manner that is inconsistent with his or her best judgment. Research studies and case histories have demonstrated that even for well-trained and experienced experts, perceptual distortion, inaccurate judgment, or illogical interpretation of evidence can result from a variety of biasing influences.

This article illustrates how the opinions and expert testimony of forensic document examiners performing comparative analyses of signatures and handwriting are susceptible to biasing influences that can improperly taint and sway their decision-making process and the manner in which they testify as experts. Understanding the various sources of bias and learning how to limit or minimize their influence are essential for improving the accuracy of decisions made by forensic handwriting experts and the reliability of their expert testimony.

Essential Requirements for Performing Reliable Forensic Handwriting Examinations

Suitability for Comparison and Presence of Sufficient Discriminating Writing Features

Handwriting can take the form of connected writing, as in cursive script or signatures, or disconnected writing, as in hand lettering (hand printing) and the writing of numerals and symbols. The discriminating features of writing include elements of style (e.g., letter formations, spatial and proportional relationships between letters and words, and formatting features) and execution (e.g., speed and fluidity of writing movements). It is the totality (combination) of the discriminating, habitual writing habits that forensic document examiners compare and evaluate in cases involving handwriting identification and/or signature verification. Essentially, the pictorial, structural, and line quality features that are perceived to characterize two sets of writing specimens are independently assessed and then compared *inter se* to determine whether they are sufficiently similar to support a conclusion of common authorship or sufficiently dissimilar to indicate that different writers produced the two sets of writings.

At the outset of any handwriting investigation, the examiner makes judgments about whether the questioned writing is sufficiently devoid of distortion or disguise to render it suitable for comparison purposes, and whether it contains enough distinguishing features to support a decision regarding source of origin. Such judgments are discretionary and examiner dependent. Once the questioned writing is adjudged to be suitable for comparison purposes, the examiner then evaluates the quantity and quality of the submitted exemplar (known) writing to assess its suitability and adequacy for comparison purposes.

To determine whether a questioned signature is genuine, a trained forensic handwriting expert focuses

on the intricate details that make up the component (structural) parts of the signature and the relative speed and fluency (rhythm) with which those details are executed. An attempt to duplicate the signature of another person based on a known sample or "model" of that person's signature is referred to as a *forgery by simulation* or *simulated forgery*. In so doing, the forger attempts to duplicate the normal and natural writing habits and abilities of another while simultaneously discarding his or her own customary writing habits and abilities.

Adequacy of Exemplars Used for Comparison Purposes

Obtaining a sufficient number of samples of an individual's normal writing is an essential requirement in investigating whether such individual authored a questioned or disputed handwritten item; these samples are termed *exemplars*. The exemplars must be sufficient in quantity to provide a sound basis for evaluating and ascertaining the natural range of variation found within the subject individual's handwriting or signature pattern. Variations found within the same person's writing or signature pattern are often referred to as "*intra*-writer" *differences*, whereas "*inter*-writer" differences refer to *dissimilarities* that are attributable to another writer. In any case involving questioned writings or disputed signatures, the critical task for the forensic document examiner is to ascertain whether apparent differences are *intra*-writer differences indicative of common authorship, or *inter*-writer differences evidencing different writers.

Ideally, the exemplars should be written as close as possible to the alleged date(s) of preparation of the questioned writing(s). A principal source of error in disputed signature cases is when the handwriting expert bases an opinion of forgery on exemplar signatures of a remote date, an inadequate amount of exemplar signatures, or exemplar signatures that are "cherry picked" by a disclaiming signatory in an attempt to provide spurious support for an unmeritorious claim of forgery. In many signature comparison cases, on obtaining a truly representative sampling of the disclaiming party's signature pattern, what at first glance were perceived as "apparent differences" are oftentimes demonstrated and proven to be "normal variations" within the same person's signature pattern (*intra*-writer differences), and hence *prima facie* proof of genuineness.

Objectivity in the Analysis and Interpretation of Evidence

Objectivity is essential to the integrity and accuracy of any forensic handwriting investigation. Unfortunately, a handwriting expert's neutrality and objectivity can be compromised by domain-irrelevant contextual (background) information and motivational factors, causing the truth-seeking goal to be eclipsed by an outcome-oriented goal.

Sources of Cognitive Bias That Can Unduly Influence the Outcome of Forensic Handwriting Examinations

Contextual bias is the most common form of cognitive bias encountered in forensics. It occurs when potentially biasing background information that is irrelevant to the discipline-specific task assigned to an examiner (e.g., examining and comparing handwriting) is conveyed to the examiner before the examiner has completed the task and reached a conclusion. As noted by Miller [2], it is not uncommon for forensic document examiners to be "briefed" about the background of the case surrounding the document(s) being submitted to them for forensic handwriting analysis. Such extraneous information usually suggests the outcome preferred or desired by the party requesting the analysis, and consequently, has the potential to unduly influence and distort the examiner's visual perception and evaluation of the handwriting evidence submitted. In the law enforcement or criminal justice setting, potentially biasing information usually concerns the crime itself, the criminal background of the suspect, or knowledge of a confession or some other form of physical or testimonial evidence linking the suspect to the crime.

The idea that document examiners should be insulated from all information about an investigation except necessary, domain-specific information is not novel. William E. Hagan's 1894 treatise on the examination of disputed handwriting and signatures contained the following commentary highlighting the need to keep document examiners "blinded" from such biasing influences:

> "... the examiner must depend wholly upon what is seen [in the forensic examination], leaving out of consideration all suggestions or hints from interested parties; and if possible it best subserves the conditions of fair examination that the expert should not know the interest which the party employing him to make the examination has in the result. Where the expert has no knowledge of the moral evidence or aspects of the case in which signatures are a matter of context, there is nothing to mislead him, or to influence the forming of an opinion; and while knowing of the case as presented by one side of the context might or might not shade the opinion formulated, yet it is better that the latter be based entirely on what the writing itself shows, and nothing else." [3]

Motivational bias on the part of forensic experts can be attributed to a variety of factors, and research on motivated reasoning has shown that an individual's reasoning processes are more readily biased when the individual is motivated by goals other than accuracy [4]. Wharton provides the following commentary regarding motivational bias on the part of handwriting experts:

> "It is well known that in cases of peculiar difficulty, when the difference, if there be any, between two handwritings is only noticeable by perceptions, the most sensitive experts, no matter how conscientious, often take unconsciously such a bias from the party employing them as to give to their judgment the almost infinitely slight impulse that turns the scale; nor is it strange that, in an instrument so delicate, aberrations from its true course should be produced by attractions or repulsions otherwise unappreciable. If an expert could be absolutely secluded from such extraneous influences, his judgment might be depended on at least for impartiality. This, however, is impracticable. A jury is bound, therefore, to accept the opinion of an expert as to handwriting, even when uncontradicted, as an argument rather than a proof; and to make allowance for all the disturbing influences by which the judgment of the expert may be moved." [5]

Wharton's view of when the judgments of handwriting experts are most vulnerable to bias is confirmed by psychological research indicating that forensic practitioners are less likely to be swayed by potentially biasing influences when the evidence is clear-cut and unambiguous [6]. Simply put, it is far more difficult to rationalize a desired outcome in the face of very strong if not irrefutable evidence to the contrary.

Contextual and motivational influences can produce confirmation bias, the tendency to seek out and interpret evidence in ways that support or confirm pre-existing beliefs and desires. Conflicts between truth-seeking goals and outcome-oriented goals are often fueled by the adversarial nature of the legal process itself [7]. The 2009 NAS Report [8] and research studies [9] indicate that forensic practitioners assigned to evaluate evidence may be motivated to see their side of a case prevail, which can lead them to endorse a biased view of the evidence that is consistent with their adoption of an adversarial outcome-oriented role instead of an objective truth-seeking one. Moreover, "tough-on-crime" attitudes prevalent within the law enforcement community tend to foster confirmation biases that leave prosecutors, investigators, and forensic specialists in crime laboratories more inclined to prioritize the value of obtaining a conviction of the accused over the countervailing priority of protecting the accused from a wrongful conviction [10–12]. These potentially biasing influences render handwriting and other forensic experts vulnerable to making erroneous decisions about the evidence they evaluate.

Absent the implementation of practical bias-minimizing procedures when evidence is submitted to and evaluated by handwriting experts in civil or criminal cases, it is unlikely to expect such experts to be kept "blinded" from domain-irrelevant information or other potential biasing influences. Some of the recommended changes that have been proposed for minimizing bias in handwriting investigations are discussed later in this article.

Observer Effects: How Examiner Bias Can Unduly Influence Forensic Handwriting Expertise

Observer effects refer to the ways in which an examiner's perception and interpretation of evidence can be influenced by the examiner's preconceived beliefs and motives, or by the surrounding context, which can include background information conveyed to the examiner as well as the evidence itself, the latter being a phenomenon often overlooked. Examiner bias in forensic handwriting investigations can influence

how examinations and comparisons are performed, the visual perceptions and observations of the examiner, the findings and opinions drawn from evaluating and comparing questioned and known writings, and the manner in which the examiner testifies in court. Some of the mechanisms and mental processes by which such cognitive bias can operate are discussed in the following sections.

Selective Exposure: Choosing Which Evidence to Examine

One way for a handwriting expert to arrive at a particular conclusion or outcome is to choose which evidence to examine for the purpose of testing the hypothesis under consideration. Handwriting experts who engage in selective exposure practices "shield" themselves and others from discordant evidence by using only handwriting or signature exemplars that support the favored outcome and ignoring, withholding and/or disregarding those exemplars that contradict or refute the favored outcome. In some instances, such practices may be more attributable to a lack of ethics than the influence of bias.

Selective Scrutiny: Selectively Evaluating Evidence in a Manner That Favors a Particular Outcome

A handwriting expert's selective scrutiny of evidence occurs when the expert searches *only* for evidence that will confirm the expert's favored outcome. An example would be where a handwriting expert's favored outcome is common authorship and in the course of examining the evidence, the expert's attention is disproportionately focused on looking for similarities in writing features between the questioned and exemplar writing, thereby failing to meaningfully search for or recognize the presence of differences in writing features between the two sets of writings. As discussed later, the Fischhof Method of upside-down writing comparison developed by this author's grandfather in the early 1900s can help to minimize the risk that differences between two sets of writings will be overlooked in cases of particular difficulty or ambiguity.

Overlooking Differences in Writing Features Due to Observer Effects from the Evidence Itself ("Familiarity Heuristic")

Much of the literature in the emerging field of cognitive forensics has focused on observer effects resulting from extraneous (domain-irrelevant) contextual information, neglecting the effect that the evidence itself can have on the observer's visual perceptions of that evidence. Although cognitive psychologists have long been aware that familiarity can cause oversight of unusual events or situations, this heuristic (which I have labeled the "familiarity heuristic") has been largely overlooked as a factor contributing to observer effects in forensic handwriting examinations and comparisons [13].[a]

Julius Fischhof, a pioneer in the field of questioned documents and Eastern Europe's leading handwriting expert in the 19th century,[b] recognized that in the context of text-based handwriting, the familiarity of letters or words can *unconsciously* contribute to the failure on the part of an examiner to observe or recognize salient writing features, most notably differences between two sets of text-based signatures or handwritten items that appear to be very similar. Fischhof discovered that by comparing such questioned and known signatures or handwriting upside down, the examiner is not subconsciously influenced by reading individual letters or words and has a more objective view of writing features [14]. In essence, the Fischhof Method of upside-down comparison offers the examiner a means of avoiding undesirable observer effects from the very thing being observed – the handwriting – by preventing the ocular distraction that results from following written characters or words that are readily familiar to the observer. It serves to minimize the cognitive "noise" and "interference" resulting from the familiarity heuristic associated with observations of text-based handwriting by altering the handwritten image into something that is unrecognizable (illegible), thereby tricking the brain into thinking it is seeing an unfamiliar image. By providing a totally different visual perspective of the very same evidence, the Fischhof Method can afford forensic document examiners the type of visual feedback that can help them avoid overlooking perceptible differences in relevant writing features that might impact the accuracy of their judgments about handwriting. It is akin to

the professional proofreader's strategy of reading material backward when checking for misspellings in order to avoid becoming distracted by content issues.

In a 1989-New York State Surrogate Court decision [15] involving conflicting expert testimony about the validity of a signature appearing on a shareholders agreement between two brothers, the Court commented favorably on the Fischhof Method:

> "The petitioner's expert, a well-known authority and author in the field of handwriting analysis, concluded that the signature of Walter Last on the shareholder's agreement was a forgery. Her testimony included a detailed analysis of the subject signature with a comparison to known exemplars of the decedent's signature. She employed an "upside-down" technique in which a known and a questioned signature are compared after they are inverted. Since there is a natural tendency to read words instead of noting variations in characters, this method allegedly gives the examiner a truer basis for comparison. Employing photographic enlargements of known signatures and the questioned signature, and acknowledging that no two signatures of the same person are exactly alike, she emphasized differences in both primary and secondary characteristics and opined that the questioned signature was not that of the decedent.
> There was an attempt to show, both by testimony that the Last brothers signed each other's signature, and by noting certain characteristics in Bert's signature, that the questioned signature "is more identical to the characteristics in Bert Last's handwriting ... than with Walter Last's signature". The court, after being advised that the petitioner did not intend to show that Bert had committed the forgery, ruled the testimony irrelevant and barred further questioning along these lines.
> The respondent countered with another expert, a trained examiner of questioned documents. He described his method of examining the questioned signature and comparing it with a series of known signatures of the decedent. The expert considered such features as skill, slant, speed, spacing proportions, relative size, and upper case letter versus lower-case comparisons. Pen stops, hesitations, tremors and possible tracing were also taken into account. Pictorial aspects and design forms were reviewed, particularly as they applied to variation (no two signatures of a person are exactly alike). On the basis of these tests, this expert concluded that the questioned signature is that of the decedent. When asked why the questioned signature appeared to have a break between two letters, he said the lack of a "connecting stroke" was insignificant, attributing it to a normal variation. Under extensive cross-examination, he explained apparent inconsistencies in the signatures, such as hooks, straight lines and spaces. He found all fell within the parameters of variation contemplated in multiple, one-author signatures.
> The expert testimony offered by the petitioner, while lacking in certain respects, was more convincing than that presented by the respondent. ***The analysis conducted by the petitioner's expert***, ***particularly the "upside-down" comparison***, ***was credible and persuasive***. The explanation offered by the respondent's expert was insufficient to eliminate glaring differences between the signatures, particularly as regards spacing. The normal variation present in everyone's signature does not account for the divergence in primary characteristics, as cogently explained by the petitioner's expert.'' (Emphasis added) [15]

Selective Stopping ("Rush to Judgment" Mindset)

Selective Stopping occurs when an investigation prematurely terminates further inquiries after having found some evidence to support a favored outcome but before adequate consideration was given to alternative hypotheses or the existence and availability of evidence that would tend to refute the favored outcome. This "rush to judgment" mindset, a byproduct of confirmation bias, has contributed to flawed FBI investigations in several high-profile cases, such as the wrongful arrests and subsequent exonerations of Richard Jewell in connection with the 1996 bombing of Atlanta's Centennial Olympic Park that killed 1 individual and injured 117 others, and Brandon Mayfield in connection with the 2004 Madrid train bombings that killed 191 individuals and wounded 1800.

Selective Reevaluation of Evidence and/or Revision of Findings

More often than not, domain-irrelevant background information about a given case is conveyed to an examiner at the time the evidence is initially submitted for analysis. Sometimes, the information is obtained afterward, as when the examiner learns that his/her findings are inconsistent with test results obtained from forensic analysis of other

items of evidence in the case, or from analysis of the very same evidence by a different analyst. Cross-communication of findings from analysis of the same or other evidence can unduly influence the objectivity of handwriting experts, and it has been raised as a possible source of error in many cases involving handwriting identifications, especially where the disclosure of such information has prompted the examiner to refine or change the initial conclusion after "reevaluating" the very same evidence. In response to the 2009 NAS Report and revelations that cognitive bias contributed to laboratory and practitioner errors in some high-profile criminal cases, the FBI laboratory has reportedly discontinued its long-standing practice of allowing forensic examiners in one discipline unit to know the findings of forensic examiners in another discipline unit and to confer with one another in the event of conflicting results.

The Impact of Examiner Bias: Flawed Opinions and/or Disingenuous Testimony

The following two case studies demonstrate how cognitive or motivational bias can contribute to erroneous findings and/or disingenuous testimony on the part of experienced and presumptively well-trained forensic document examiners.

Case Study 1: Questioned Signatures

Felder v. Storobin, 100 A.D.3d 11, 953 N.Y.S.2d 602 (N.Y. App. Div. 2012), involved an appeal from a trial court's decision to dismiss a proceeding which Felder commenced against Storobin to invalidate a petition designating Storobin as a candidate for election to the New York State Senate. The New York appellate court described the factual issues and expert testimony before the trial court (Supreme Court) as follows:

> "Felder alleged that five signatures witnessed by Storobin were actually forged: those of Anatoliy Smolyanskiy, Edith Garcia, Arnaldo Garcia, Carina Tretyakov, and Lyudmila Tretyakov. [The] handwriting expert called as a witness by Felder, testified that each of these five signatures was forged, based upon his comparison of the designating petition with the voter registration records maintained by the Board of Elections. He described the differences in signatures as great and glaring. With respect to four of the signatories, the exemplar signatures from the Board of Elections were 28 years old, 20 years old, 19 years old, and 12 years old, respectively. [The expert] conceded in his testimony that a person's signature may change with time and age. Felder did not call as witnesses any of the voters in question, and did not produce comparative signature evidence more recent than that set forth in the records obtained from the Board of Elections.
>
> Storobin called Smolyanskiy, the fifth signatory, as a witness. Smolyanskiy identified his signature on the designating petition, and recalled signing his name to it in the presence of Storobin and another person. Storobin testified that he personally obtained the signatures at issue. The Supreme Court credited Smolyanskiy's and Storobin's testimony as to Smolyanskiy's signature.
>
> As to the remaining four signatures, the Supreme Court found [the handwriting expert's] testimony insufficient to meet the burden of proof for fraud, particularly in light of, inter alia, the significant gaps in time between the dates of the voters' exemplar signatures from the Board of Elections and the signatures on the designating petition. The Supreme Court also found the testimony of Storobin to be credible.'' [16]

In affirming the lower court's findings and decision to dismiss the case, the appellate court noted

> "[The handwriting expert's] testimony that Smolyanskiy's signature was forged, followed by Smolyanskiy's testimony that the designating petition had, in fact, been signed by him, eviscerated [his] credibility as an expert witness on the issue of the authenticity of Smolyanskiy's signature, and allowed the Supreme Court to find [his] testimony to be "unconvincing and questionable, at best" as to the remaining designating signatures as well. Again, we defer to the Supreme Court's assessment that [the handwriting expert's] testimony was not credible." [16]

The handwriting expert who testified in this case was employed for 30 years with a state police crime laboratory that allowed him to accept private sector civil casework; he was board certified by the American Board of Forensic Document Examiners (ABFDE) and a long-standing member of the Questioned Document Section of the American Academy of Forensic Sciences (AAFS) and the

American Society of Questioned Document Examiners (ASQDE). He clearly possessed the requisite education, training, and experience to know that basing an opinion of forgery on perceived differences appearing in signatures written many years before the questioned signatures (more than two decades in one instance) violated one of the basic tenets and technical standards of forensic signature examination and comparison.

This civil case illustrates how selective stopping can unduly influence the decision making and testimony of a handwriting expert. The only appropriate decision for the handwriting expert in this case would have been to discontinue comparison and express no opinion until more contemporaneous exemplars were obtained.

Case Study 2: Questioned Handwriting (Disguised Hand Printing)

Adams v. Weber, 2005 Extra LEXIS 216 (Circuit Court, Fifth Judicial District, SD, 2005), involved an action brought by Samuel D. Adams a/k/a Dale S. White pursuant to an application for a Writ of Habeas Corpus. Adams claimed that his court-appointed defense attorney (Brankin) provided ineffective assistance of counsel when representing him in connection with a 2001 criminal case.

As described in the South Dakota Circuit Court's Memorandum Decision, the underlying criminal prosecution arose out of an incident that occurred while Gayle Wanous (Wanous) was working alone in her flower shop. A native American man entered the shop, identified himself as Sam Adams, and said that he is on his lunch break from Dakota Connection and wants to buy some flowers for his girl friend. As the customer started to write out an enclosure card at the counter, he asked Wanous to add something to his order. When Wanous went to a back workroom to get something, she was struck from behind on the head. When she awoke, she had no recollection of what had happened. After cleaning the blood from her hands and head, she immediately called the police on noticing that her cash drawer was open and all the cash and checks had been taken. Shortly after Chief Flannery and Sergeant Fisher of the Sisseton Police Department arrived at the scene, Wanous was taken by ambulance to a hospital where she remained for 5 days after her injuries were discovered to include a 4-inch cut to the back of her head, a fractured skull, and a concussion. Four days after being discharged from the hospital, Wanous provided the police with her initial statement regarding the incident, and several days later (2 weeks after the incident), Wanous identified Adams from a photo lineup as the person who was in her store at the time she was attacked. Defendant Sam Adams was subsequently prosecuted for aggravated assault, first-degree robbery, and first-degree burglary. A jury found Adams guilty of all three counts and he was sentenced to 25 years in prison [17].

The only physical evidence recovered from the scene that might link the defendant to the crime was the small enclosure card found on the counter that contained the hand printed phrase "To Karen From Sam". No other physical evidence was recovered to connect Adams to the crime scene – no fingerprints, no blood, no DNA [17].

In granting Sam Adams' petition for a Writ of Habeas Corpus,[c] the Circuit Court determined that the defense attorney's laziness and complete incompetence undermined every aspect of Adams' defense. The Court cited numerous instances of gross ineptitude on the part of the indigent defendant's court-appointed attorney, with perhaps the most damaging one being his total lack of preparation concerning handwriting analysis [18]. With respect to the trial testimony of the State's handwriting expert, the Court noted:

> "The State's case against Adams was largely circumstantial in nature. The State did not produce any witnesses to the alleged crimes, other than the victim. The Sisseton Police Department did not collect any evidence that directly tied Adams to the scene of the crime. No fingerprints, weapon, or money was ever recovered. A major piece of physical evidence presented to the jury was the enclosure card left on Wanous' counter. As related earlier, it was inscribed with the words "To Karen, From Sam." However, the handwriting on the card was unnatural stick writing rather than normal printing. The State alleged this card was written by Adams. The defense maintained Adams did not write the card, and believed Sergeant Fisher forged it as evidence against Adams.
> The State employed a forensic document examiner . . . as an expert to analyze the handwriting on the card. Brankin never attempted to employ his own handwriting expert, and never educated himself on the area of handwriting analysis. Approximately two weeks before the start of trial he stipulated to

> the [State's use of its handwriting expert] without fully reviewing the contents of her report.
> Prior to issuing her report, the [State's expert] received writing exemplars of both Adams and Fisher that were analyzed against the card. [The expert] stated that she received an inconclusive result when she analyzed Adams' handwriting, but that she could conclusively rule out Sergeant Fisher as the author. At trial, [the State's expert] testified to her inconclusive finding in regards to Adams, but went on to detail similarities between Adams handwriting and the card. Brankin allowed her to give lengthy testimony on the issue without challenging her conclusions or prompting her to detail the similarities between the two.'' [19]

[The State's expert] testified that she had been employed as a forensic document examiner with the Minneapolis Police Department since 1978, and had also been accepting private sector civil casework assignments since 1988. Her professional training included a 4-year apprenticeship with the Questioned Document unit of the Indiana State Police, FBI and US Secret Service training courses, and attendance at symposiums and workshops sponsored by professional membership organizations in the field. She was a member of the Questioned Document Section of the AAFS and the Midwestern Association of Forensic Scientists (MAFS). Although not board certified, she had testified approximately 190 times in local, state, and federal courts [20].

At the Habeas hearing, Adams presented testimony from three handwriting experts: Allan Keown, Vickie Willard, and Pat Girouard. All three experts, two of whom (Willard and Girouard) were Diplomates of the Board of Forensic Document Examiners (BFDE),[d] echoed essentially the same concerns about the impropriety of comparing two sets of writings not suitable for comparison, and [the State expert's] bias in overstating an inconclusive opinion and providing disingenuous testimony. Each opined that [the State's expert] was allowed to offer improper opinions that contravened the technical standards of handwriting analysis, and that Adams' defense attorney was not familiar with those standards and wholly unprepared to meaningfully challenge the admissibility of [her] opinions or to impeach and discredit her testimony. The two technical standards at issue were American Society for Testing and Materials (ASTM) standards, one establishing best practices for performing handwriting examinations and the other defining the standard terminology used by forensic document examiners in expressing conclusions. Willard pointed out that as an active member of ASTM Subcommittee E30.02 on Questioned Documents, [the State's expert] had actually participated in writing and developing the two ASTM standards at issue [21].

The technical deficiencies of the handwriting opinions and biased nature of the trial testimony presented by [the State's expert] are specifically described in the following sections.

Incomparability of Writing Features: the Significance of the Questioned Hand Printing Being Unnatural Stick Printing and the Exemplars Being Natural Printing. [The State's expert] testified that the enclosure card found at the crime scene was "written in unnatural stick printing ... as a means of disguise." Both Sam Adams' exemplars and Sgt. Fisher's exemplars were admittedly written in natural printing, and [the State's expert] made no request to obtain additional exemplar writing from either subject.

In comparing questioned writing consisting of unnatural stick printing with exemplars comprising only natural printing, [the State's expert] departed from the standard methodology and recognized best practices set forth in ASTM Standard E2290-03, *Standard Guide for Examination of Handwritten Items* (ASTM Standard E2290), which provided as follows:

> § 7.6.1: If [the questioned writing] is not natural writing, or ... the available questioned writing is not suitable for comparison, discontinue these procedures and report accordingly.
> § 7.9.1: If [the known writing] is not natural writing, or ... the available questioned writing is not suitable for comparison, discontinue these procedures and report accordingly.
> § 7.11.1: If the bodies of writing are not comparable, discontinue comparison and request comparable known writing, if appropriate. [22]

Distorting an Inconclusive Opinion in a Manner That Favors a Particular Outcome. [The expert's] conclusion as to whether Sam Adams wrote the unnatural stick printing on the enclosure card (marked at trial as "Exhibit 4") was stated to be "inconclusive", as indicated by the following excerpt from the official transcript of her direct testimony:

> Q: When you did your comparison of Exhibit 4 to the items related to Sam Adams' handwriting, what was your conclusion from the comparison?

A: ***My conclusion was that I was inconclusive***. There were both similarities to Mr. Adams' writing, as well as characteristics that were not found in the sample that I had of this writer. So based on the combination of what I had and what I did not have, I determined that with what was submitted that, actually, ***a conclusion could not be rendered in one direction or another***. (Emphasis added) [23]

ASTM Standard E1658-96, *Standard Terminology for Expressing Conclusions of Forensic Document Examiners* (ASTM Standard E1658), recommends and defines several terms that forensic document examiners should use to express the level of confidence associated with their opinion(s); it provides a standardized framework for understanding the true meaning of the level of confidence associated with an opinion or conclusion expressed by a forensic document examiner. As defined in ASTM Standard E1658, the terms *inconclusive* and *indeterminable* are synonymous and represent "the zero point of the confidence scale"; these terms are "used when there are significant limiting factors, such as disguise in the questioned and/or known writing or a lack of comparable writing, and the examiner does not have a leaning one way or the other." [24]

Once [the State's expert] expressed an inconclusive opinion as to whether Sam Adams wrote the card and testified that no conclusion could be reached one way or the other, the only proper and accurate statement that she could make was that Sam Adams "*cannot be eliminated or identified* as the writer". However, [she] chose to embellish her testimony with an inaccurate and misleading statement designed to favor a particular outcome, as reflected in the following exchange during her direct testimony:

Q: Ms. [*name omitted*], you're not saying that Sam did not write this card?
A: No. ***You cannot eliminate him as a writer***, no.
Q: But you're not saying that he did?
A: ***Neither can you identify him positively as the writer***, no. (Emphasis added) [25]

[The expert's] "gratuitous" inclusion of the word "positively" in testifying that Sam Adams *cannot be eliminated or positively identified* as the writer clearly manifested bias in favor of the prosecution. This overstatement was highly prejudicial to the defendant because it wrongfully implied a "near-match", i.e., that the defendant *can be identified as the writer*, *but just not positively*. This form of disingenuous testimony is not uncommon in criminal cases involving prosecution handwriting experts who appear to be unduly influenced or motivated to testify in a manner that suggests support for the inculpatory hypothesis even when the evidence itself favors neither the inculpatory nor exculpatory hypothesis.

Misinterpreting Evidence or Providing Exaggerated Testimony in Order to Support a Favored Outcome. The defense maintained that Adams did not write the card, and suggested that Officer Fisher fabricated it as evidence against Adams [26]. The following excerpt of [the expert's] trial testimony concerns the results of her examination and comparison of the unnatural stick printing on the enclosure card with the naturally written hand printing exemplars of Officer Fisher:

Q: What was your ***conclusion from the comparison of Officer Fisher's known handwriting*** to the questioned document?
A: My conclusion was that it was ***highly probable that he was not the writer of the questioned material***. (Emphasis added) [27]

ASTM Standard E1658, *supra*, defines "highly probable" as meaning that "the evidence is very persuasive" and "the examiner is virtually certain" of the conclusion (opinion) expressed [28]. Hence, [the State's expert] concluded that Officer Fisher could be eliminated *with virtual certainty* as the writer of the unnatural stick printing appearing on the enclosure card.

As noted earlier, no conclusion could be rendered in one direction or another regarding authorship of the questioned writing because the unnatural stick printing appearing on the card was not suitable for comparison with the naturally written exemplars available for both Adams and Officer Fisher. Accordingly, an inconclusive opinion was warranted with respect to whether Officer Fisher wrote the enclosure card for the same reason the State's handwriting expert reached an inconclusive opinion with respect to whether defendant Adams wrote the card – unnatural stick printing cannot be compared to natural printing.

The only way Officer Fisher could have properly been eliminated as the writer of the enclosure card was by evidence showing that his writing skills were

so impaired as to have made it impossible for him to produce the unnatural stick printing at issue. Such evidence being absent, the inclination on the part of the prosecution's handwriting expert to disassociate the disguised hand printing on the enclosure card from the natural hand printing of Officer Fisher clearly reflects a biased conclusion derived from an illogical interpretation of evidence, presumably resulting from the prosecution expert's adoption of an adversarial role in which the outcome-oriented goal trumped the truth-seeking goal.

Proposed Solutions for Minimizing Examiner Bias in Handwriting Investigations

Forensic document examiners and others in the forensic science community have historically dismissed cognitive bias as a factor contributing to examiner errors in casework, insisting that such errors are caused by incompetence or dishonesty rather than domain-irrelevant contextual or motivational influences. Consequently, there has been a long-standing reluctance on the part of the forensic community at large to acknowledge the need to develop internal procedures and strategies designed to minimize the likelihood of having the objectivity of forensic decision making compromised by potentially biasing influences. However, there now exists a substantial body of empirical research reported in peer-reviewed scientific and legal journals and presented at professional conferences that clearly establishes the susceptibility of handwriting, fingerprint, and other pattern recognition experts to having the results of their examinations and comparisons cognitively contaminated and unduly influenced by domain-irrelevant contextual information and motivational bias [13, 28–55]. With more and more stakeholders recognizing and understanding the insidious manner in which cognitive contaminants can be toxic to one's neutrality, proposals and recommendations for minimizing examiner bias are now receiving considerably more attention within the forensic science, legal, and academic communities, as reflected in some of the more recent peer-reviewed publications and presentations addressing this topic [29, 30, 35–44, 47, 50–53].

In the case of forensic handwriting investigations, almost all of the procedures and strategies that can be used to minimize examiner bias involve either implementing examiner debiasing techniques or restructuring institutional context management and evidence testing protocols, as briefly summarized in the following sections.

Debiasing Techniques for Examiners

Considering the "Oppositional Hypothesis" First. As domain-irrelevant information invariably enters the scene through the mouths of lawyers or clients intent on convincing the handwriting expert of the merits of their claim(s), healthy skepticism on the part of the expert goes a long way toward ensuring neutrality in the analysis and evaluation of handwriting evidence. In considering the oppositional hypothesis first, an examiner approaches the investigation with the mindset of having been hired by the adverse or oppositional party. In this way, the examiner is forced to consider the least favored hypothesis first and elaborate on the reasons for rejecting it. Only then does the examiner consider the most favored hypothesis.

Considering Alternative Possibilities and Hypotheses (Playing the Role of "Devil's Advocate"). Considering all plausible alternative possibilities before deciding on a particular one is essential to the integrity of any type of investigation. Promoting a "devil's advocate" mindset in which thinking "outside the box" is encouraged should therefore be prioritized in the training and continuing education of all forensic experts, as contrarian and critical thinking skills are needed in order to be able to both generate and properly evaluate plausible alternative hypotheses. This is particularly important in the case of handwriting, as its physical appearance can be significantly affected by a variety of environmental and motivational factors (e.g., awkward writing position, the influence of drug or alcohol intoxication/withdrawal, the import of the document itself, and deliberate attempt at disguise).

It has also been suggested that examiners should not be allowed to *summarily* dismiss alternative possibilities and hypotheses, and that any refutations should be accompanied by documentation that describes in detail the reasons for rejecting each alternative possibility.

Using the Fischhof Method to Compare Text-based Writings That Appear Very Similar and as a "Self-Review" of One's Initial Observations. The Fischhof Method of upside-down comparison described earlier in this article can be used as a possible safeguard against overlooking differences in salient writing features whenever a handwriting expert is confronted with two sets of text-based writings that appear quite similar. This method of comparison can also afford an examiner a "fresh new look" at the evidence, enabling observations from the initial analysis to be measured against observations derived from inverted image comparisons of the very same evidence. Optimally, such a self-conducted review should take place at a time well after the initial handwriting analysis so as to reduce the likelihood of any "interference" produced by recall of observations made during the initial analysis.

Institutional Context Management Protocols and Procedures

Separating the Crime Laboratory or Evidence Analysis Function from the Police and Prosecutorial Functions. The 2009 NAS Report recommended separating the crime laboratory function from any law enforcement department or agency, theorizing that a truly autonomous crime laboratory would mitigate, if not remove, the institutional pressures placed on crime laboratory analysts to produce results that favor the police or prosecution theory of the case, and would foster a more neutral mindset that prioritizes the truth-seeking goal. Houck has outlined the potential difficulties of this approach [39].

Using Sequential Unmasking Procedures and Case Managers. Context management protocols involve shielding the examiner from domain-irrelevant information and employing "sequential unmasking" procedures to control the order (sequence) in which domain-relevant but potentially biasing information is "unmasked" and disclosed to the examiner. Ideally, the examiner is kept as blind as possible for as long as possible, and remains unaware of domain-irrelevant information until all examinations and tests are completed.

Optimally, a case manager who is privy to all the facts of the case is responsible for determining what evidence to test and for evaluating and interpreting the test results in the context of the case, for example, assessing whether the test results support an inculpatory or exculpatory hypothesis. The case manager should also possess, or have access to, relevant subject matter expertise, as difficult decisions may need to be made about what information is domain relevant and when and how such information should be obtained and disclosed to the examiner.

The strict protocol for sequential unmasking requires that after looking at the questioned item(s) and before looking at any exemplar(s), the examiner must determine (and make a written record of) the specific distinguishing features that the examiner would rely on in deciding whether to associate or disassociate the questioned item(s) with the exemplar(s). This procedural requirement is deemed a necessary safeguard against target shifting, in which knowledge about features contained in the exemplar(s) influences the examiner's interpretation of the questioned item(s) and the examiner's decision about which features are relevant and irrelevant for comparison purposes [13, 40, 43, 46, 47].

Although sequential unmasking procedures can be implemented with relative ease, most, if not all, forensic laboratories in the United States have not done so for handwriting investigations. This may be due to the fact that the method by which a handwriting expert selects the salient writing features to be used for comparison purposes is subjective and examiner-dependent, there being no standardized best-practice protocol for how such feature selection should be made, let alone documented. However, the Document Examination Unit of the Victoria Police Services Department in Australia most recently embarked on a pilot study using a modified version of the sequential unmasking protocol for handwriting cases. The reader is referred to Found and Ganas [38] for a detailed description of the Australian protocol and how their sequential unmasking procedures were modified so that only the essential information required for performing the requisite handwriting examinations and comparisons is available to the examiner.

Using Exemplar (Evidence) Lineups and Blind Evidence Submission Protocols. In investigations seeking to determine whether a handwritten item can be attributed to a particular source writer, the forensic document examiner is often presented with the questioned item(s) and *only* the suspected (targeted)

writer's reference item(s), i.e., handwriting exemplars. Some commentators from the legal and scientific communities have criticized the suggestiveness inherent in such a procedure, arguing that exemplar lineups should be used for handwriting identifications and other types of evidential source attributions in much the same way that photo lineups are used for eyewitness identifications [13, 29, 40, 43, 46, 47]. As the same deficiencies that make a photo lineup unduly suggestive make an exemplar lineup unduly suggestive, both types of lineups require presenting similar-looking "fillers" ("foils") to the observer (the handwriting expert or the eyewitness). Thus, in blind exemplar lineups, the examiner would receive the handwritten item(s) in question along with an array of similar-looking handwriting exemplars from a group of anonymous individuals, including the suspected writer, and the examiner would receive no information or cognitive cues that might unduly influence the examiner to reach a particular outcome. To ensure that the analyst receives no improper cues from the person(s) tasked with submitting the evidence to be analyzed, it has been suggested that exemplar lineups be double blind, meaning that both the analyst and the individual(s) submitting the evidence or arranging the exemplar lineup(s) not know the identity of the suspect or the preferred outcome [13, 29].

In theory, every forensic pattern recognition discipline that requires comparisons between unknown (questioned) items and known reference items (exemplars) in order to determine the source of the unknown item(s) can benefit from the use of exemplar lineups, as the presence of a large number of "fillers" resembling the questioned item would arguably enhance the reliability of any ensuing identification or source attribution. In practice, however, obtaining handwriting "fillers" sufficiently similar to the questioned writing is far more difficult than obtaining suitable "fillers" for photo lineups, and such exemplar lineups may be of little usefulness in instances where the questioned writing displays several highly distinctive, individualizing writing features. In addition, double-blind lineups would seem to be far more feasible in handwriting investigations undertaken by public sector or institutional forensic laboratories where examiners can work with case managers possessing discipline-specific (handwriting) expertise and "blind" evidence lineup administrators than by private sector examiners who work as solo practitioners and receive their casework assignments directly from lawyers or clients, oftentimes accompanied by cues that indicate the desired outcome.

The use of single-blind exemplar lineups in handwriting cases is not a new development. For example, in investigating the source of anonymous handwritten letters emanating from a limited population of possible writers, experienced forensic document examiners routinely insist on using exemplar lineups and being kept blinded to which of the exemplars belongs to the suspected writer until such time as the examiner has completed all examinations and reached a decision.[e] However, the usefulness of such lineups varies inversely with the distinctiveness of the handwriting features observed in the questioned writing(s); that is, the more distinctive the writing features in the questioned item, the more difficult it will be to find similar-looking "fillers," and hence, the less useful the "fillers" will be.

Single-blind exemplar lineups pose a real test that a handwriting expert can conspicuously fail, for even if the expert has been exposed to biasing influences, the expert would still not know which of the exemplars actually came from the suspected writer. The double-blind exemplar lineup method offers the kind of proficiency testing that can produce objective, meaningful data regarding individual examiner error rates, which may account for continued resistance within the relevant community to the use of such evidence lineups.

Conclusion

Empirical research and case reviews establish that contextual and motivational biases influence the decisions, opinions, and testimony of handwriting experts. These biasing influences consciously or unconsciously lead to errors, which account for the NAS Report's recommendation that standard operating procedures be implemented to minimize, to the greatest extent reasonably possible, potential bias and sources of human error in forensic practice [56]. Forensic document examiners must focus attention on how best to deal with the problem of cognitive bias and its impact on handwriting expertise. In order for this to happen, they need to be convinced that cognitive bias cannot be eliminated by sheer willpower.

Experts are human, and as such, they will remain susceptible to having the results of their examinations and comparisons influenced by domain-irrelevant

contextual information and motivational forces. Changes in evidence examination procedures are necessary to ensure that examiners are shielded from extraneous cognitive contaminants. More research efforts in the emerging field of cognitive forensics are needed to study the effect of potentially biasing influences on forensic judgments about evidence and ways to reduce or eliminate some of those influences.

End Notes

[a.] The flexibility of the human cognitive system permits us to "tune" ourselves to perceive some things and ignore other things, usually so automatically and seamlessly that we rarely realize we are doing it. This tuning process results in "selective attention" to information. See Ref 13, *infra*, at 15.

[b.] Fischhof, whose forensic document laboratory was located in Budapest Hungary, was one of a handful of handwriting experts consulted in connection with France's infamous Dreyfus Affair and earned his reputation when he opined that Alfred Dreyfus did not write the memorandum ("bordereaux") that was wrongfully attributed to him by several biased prosecution handwriting experts, some of questionable credentials. The Dreyfus Affair in 19th century France was one of the most notorious cases involving biased opinions offered by handwriting experts. The case revolved around a handwritten memorandum (the infamous *bordereaux*) containing details of France's secret military plans and weaponry that was delivered to an agent of the German government. Amidst a national and political climate of anti-German and anti-Jewish sentiment, Captain Alfred Dreyfus, the French army's only Jewish officer, was charged with treason and convicted in a sham trial involving flawed and biased testimony on the part of experts claiming handwriting expertise. The wrongful conviction and exile of Dreyfus to Devil's Island drew worldwide attention and outrage, prompting Emile Zola, France's most respected writer, to pen his famous expose, *J'Accuse*. Postconviction investigations revealed that the "bordereaux" was written by Major Ferdinand Esterhazy, the son of a French General who was an illegitimate member of the aristocratic Esterhazy family of Hungary.

[c.] After winning his court-ordered release from the South Dakota State Penitentiary, Sam Adams was never retried by the State.

[d.] See http://www.bfde.org. The BFDE was the first forensic document examiners certification board to be accredited by The Forensic Specialties Accreditation Board (FSAB), which accredits forensic specialty boards in the United States that certify practitioners (specialists) in various forensic disciplines. Only one other forensic document examiners' certification board has been accredited since, the ABFDE.

[e.] The failure to use exemplar lineups may also have contributed to the erroneous handwriting opinions offered in France's infamous Dreyfus Affair, discussed *supra* at note 12. Nearly a century later, France is once again confronted with the exemplar lineup issue in a murder case known as the "Gregory Affair". The case involves the October 1984 kidnapping of 4-year-old Gregory Villemin and a series of anonymous handwritten poison-pen letters that were sent to the family. After the boy's body was discovered, a witness incriminated an uncle, who was indicted after a handwriting expert identified him as the author of the anonymous notes. The witness later recanted, but not before Gregory's father killed the uncle to avenge his son's murder. In a subsequent blind exemplar lineup procedure in which handwriting exemplars from all members of the family were examined and compared to the anonymous notes by a second expert, Gregory's mother, Christine, was identified as the author, and she was indicted. Christine served almost 8 years in prison for Gregory's murder before being cleared on appeal. *See* Bernstein, *All of France is asking*: *Who killed petit Gregory*?, N.Y. Times, July 16, 1985; and International Herald Tribune, July 18, 1985, at 2, col.3. In 2008, an appellate court ordered the case to be reopened in the hope that new advances in forensic science can shed light on DNA evidence. France still awaits the final chapter of this unsolved murder mystery that has seen two defendants indicted for the same murder on the basis of being identified by two different handwriting experts as the sole author of the anonymous notes.

References

[1] National Academy of Forensic Sciences (2009). *Strengthening Forensic Science in the United States: a path forward*, National Academies Press, Washington DC.

[2] Miller, L.S. (1984). Bias among forensic document examiners: a need for procedural changes, *Journal of Police Science and Administration* **12**(4), 407–411.

[3] Hagan, W.E. (1984). *A Treatise on Disputed Handwriting and the Determination of Genuine From Forged Signatures*, Banks & Brothers, New York, p. 82.

[4] See Simon, D. (2012). *In Doubt: the Psychology of the Criminal Justice Process*, Harvard University Press, Cambridge, p. 25, citing Kunda, Z. (1990). The case for motivated reasoning, *Psychological Bulletin* **108**, 480–498, p. 480.

[5] *Wharton on Evidence* (3rd Ed.) §722, p. 711.

[6] See Simon, D. (2012). *supra* note 4, p. 25, citing Ask, K. & Granhag, P.A. (2008). The "elasticity" of criminal evidence: a moderator of investigative bias, *Applied Cognitive Psychology*, **22**, 1245–1259; and Dror, I.E. & Charlton, D. (2006). Why experts make errors, *Journal of Forensic Identification* **56**(4), 600–616.

[7] See Mnookin, J.L. (2008). Expert evidence, partisanship and epistemic competence, *Brooklyn Law Review* **73**, 1009–1033.

[8] NAS Report (2009), *supra* note 1, pp. 122–124.

[9] See Simon, D. (2012). *supra* note 4, p. 26, citing Simon, D., Stenstrom, D. & Read, S.J. (2008). On the objectivity of investigations: an experiment. Paper presented at Conference for Emprical Studies, Cornell law School. September 9–10.

[10] See, e.g., Perrot, S.B. & Taylor, D.M. (1995). Attitudinal differences between police constables and their supervisors: potential influences of personality, work environment, and occupational role, *Criminal Justice and Behavior* **22**, 326–339.

[11] Wortley, R.K. & Homel, R.J. (1995). Police prejudice as a function of training and outgroup contact: a longitudinal investigation, *Law and Human Behavior* **19**, 305–317.

[12] Ask, K. & Granhag, P.A. (2005). Motivational sources of confirmation bias in criminal investigations: the need for cognitive closure, *Journal of Investigative Psychology and Offender Profiling* **2**, 43–63.

[13] Risinger, D.M., Saks, M.J., Thompson, W.C. & Rosenthal, R. (2002). The Daubert/Kumho implications of observer effects in forensic science: hidden problems of expectation and suggestion, *California Law Review* **90**, 1–56.

[14] Sulner, H.F. (1966). *Disputed Documents: New Methods for Examining Questioned Documents*, Oceana Publications, New York, p. 48; Fischhof, J. (1927). *New Method of Comparing Handwriting*, City Printing Office, Szeged, Hungary.

[15] *In re Last*, 1989 NY Misc. LEXIS 933, 1989 WL 1783230 (New York State Surrogate's Court, Westchester County, 1989), pp. 2–3.

[16] *Felder v. Storobin*, 2012 953 N.Y.S.2d 602, at 606–607, 609.

[17] *Adams v. Weber*, 2005 Extra LEXIS 216 (Circuit Court, Fifth Judicial District, SD, 2005), at pp. 1–6, 30.

[18] Defense attorney Brankin failed to request, let alone consult with, an expert in forensic document examination, and stipulated to the State's use of a handwriting expert two weeks before the start of the trial. "Brankin did not review [the State handwriting expert's] report prior to trial, and had no knowledge of the technical standards used in handwriting analysis ... The skills and diligence Brankin exhibited in regards to this issue fell well outside the objective standard of reasonableness expected from competent trial counsel." *Adams v. Weber*, *supra* note 17, at pp. 49–50.

[19] *Adams v. Weber*, *supra* note 17, at pp. 30–31.

[20] Transcript of October 17, 2001 trial testimony of the State's handwriting expert in *State of South Dakota v. Sam Adams*, CR 01–61, pp. 381–383.

[21] Affidavit of Vickie L. Willard sworn to March 11, 2004.

[22] ASTM Standard E2290-03 (2003), *Standard Guide for Examination of Handwritten Items*, American Standards for Testing and Materials International, p. 3.

[23] Transcript of October 17, 2001 trial testimony of the State's handwriting expert in *State of South Dakota v. Sam Adams*, CR 01–61, p. 392.

[24] ASTM Standard E1658-96 (1996), *Standard Terminology for Expressing Conclusions of Forensic Document Examiners*, American Standards for Testing and Materials International, §4.1, p. 2.

[25] Transcript of October 17, 2001 trial testimony of the State's handwriting expert in *State of South Dakota v. Sam Adams*, CR 01–61, p. 409.

[26] The Circuit Court pointed out inconsistencies in the trial testimony of the two police officers that responded to the crime scene. "Chief Flannery and the victim both testified the victim did not give the officers a physical description of the perpetrator. Sergeant Fisher stated the victim did provide a description. The two officers testified inconsistently as to who arrived before whom at the Treasure Chest. They both testified that the other was the officer to actually pick up the enclosure card from Wanous' counter. They both stated that they considered the other officer to be in charge of the investigation." *Adams v. Weber*, *supra* note 17, at p. 38.

[27] Transcript of October 17, 2001 trial testimony of the State's handwriting expert in *State of South Dakota v. Sam Adams*, CR 01–61, at p. 415.

[28] ASTM Standard E1658-96 (1996), *Standard Terminology for Expressing Conclusions of Forensic Document Examiners*, American Standards for Testing and Materials International, §4.1, p. 1.

[29] Canter, D., Hammond, L. & Youngs, D. (2013). Cognitive bias in line-up identifications: the impact of administrative knowledge, *Science and Justice* **53**, 83–88.

[30] Cole, S.A. (2013). Implementing counter-measures against confirmation bias in forensic science, *Journal of Applied Research in Memory and Cognition* **2**(1), 46–47.

[31] Dror, I.E., Charlton, D. & Peron, A. (2006). Contextual information renders experts vulnerable to making erroneous identifications, *Forensic Science International* **156**, 74–78.

[32] Dror, I.E. & Charlton, D. (2006). Why experts make errors, *Journal of Forensic Identification* **56**(4), 600–616.

[33] Dror, I.E. (2009). On proper research and understanding of the interplay between bias and decision outcomes, *Forensic Science International* **191**, 17–18.
[34] Dror, I.E. & Cole, S.A. (2010). The vision in "blind" justice: Expert perception, judgment, and visual cognition in forensic pattern recognition, *Psychonomic Bulletin & Review* **17**(2), 161–167.
[35] Dror, I. (2012). Cognitive forensics and experimental research about bias in forensic casework, *Science and Justice* **52**, 128–130.
[36] Dror, I.E., Kassin, S.M. & Kukucka, J. (2013). New application of psychology to law: improving forensic evidence and expert witness contributions, *Journal of Applied Research in Memory and Cognition* **2**, 78–81.
[37] Dror, I.E. (2013). Practical solutions to cognitive and human factor challenges in forensic science, *Forensic Science Policy & Management* **4**(3–4), 1–9.
[38] Found, B. & Ganas, J. (2013). The management of domain irrelevant context information in forensic handwriting examination casework, *Science and Justice* **53**, 154–158.
[39] Houck, M.M. (2014). Striving to inhibit bias in criminal justice forensics: the good, the bad, and the ugly of running an independent crime lab. Presented in 2014 AAFS *Bias in Forensics* Workshop, *infra* note 52.
[40] Inman, K. & Rudin, N. (2013). Sequential unmasking: Minimizing observer effects in forensic science, in *Encyclopedia of Forensic Sciences*, 2nd Edition, Vol. 3, J.A. Siegel & P.J. Saukko, eds, Academic Press, Waltham, pp. 542–548.
[41] Kassin, S.M. (2012). Why confessions trump innocence, *American Psychologist* **67**(6), 431–445.
[42] Kassin, S.M., Dror, I.E. & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives and proposed solutions, *Journal of Applied Research in Memory and Cognition* **2**, 42–52.
[43] Krane, D.E., Ford, S., Gilder, J., Inman, K., Jamieson, A., Koppl, R., Kornfeld, I., Risinger, D.M., Rudin, N., Taylor, M.S. & Thompson, W.C. (2008). Sequential unmasking: a means of minimizing observer effects in forensic DNA interpretations (Letter to the Editor), *Journal of Forensic Sciences* **53**(4), 1006–1007.
[44] Kukucka, J. & Kassin, S.M. (2013). Do confessions taint perceptions of handwriting evidence? An empirical test of the forensic confirmation bias, *Law and Human Behavior* **2**, 1–13.
[45] Lord, C.G., Lepper, M.R. & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment, *Journal of Personality and Social Psychology* **47**, 1231–1243.
[46] Risinger, D.M., Denbeaux, M.P. & Saks, M.J. (1989). Exorcism of ignorance as a proxy for rational knowledge: The lessons of handwriting identification "expertise", *University of Pennsylvania Law Review* **137**, 731–792.
[47] Risinger, D.M. (2014). The need for sequential unmasking - with some thoughts on how to undermine proffered expert testimony infected with the precursors of bias. Presented in 2014 AAFS *Bias in Forensics* Workshop, *infra* note 57.
[48] Saks, M.J., Risinger, D.M., Rosenthal, R. & Thompson, W.C. (2003). Context effects in forensic science: a review and application of the science of science to crime laboratory practice in the United States, *Science and Justice* **43**(2), 77–90.
[49] Schiffer, B. & Champod, C. (2007). The potential (negative) influence of observational biases at the analysis stage of fingerprint individualization, *Forensic Science International* **167**, 116–120.
[50] Simon, D. (2012). In Doubt*:* The Psychology of the Criminal Justice Process, Harvard University Press, Cambridge.
[51] Stoel, R.D., Dror, I.E. & Miller, L.S. (2013). Bias among forensic document examiners: still a need for procedural changes, *Australian Journal of Forensic Sciences* **46**(1), 91–97.
[52] Sulner, A. (Chair) and Scheck, B.C. (co-Chair), 2014 Workshop on *Bias in Forensics – Examining the Sources and Impacts of Bias on Perceptual and Cognitive Judgments Made by Forensic Experts*, *Strategies for Excluding or Impeaching Expert Testimony Tainted by Bias*, *and Proposed Solutions for Minimizing or Inhibiting Biasing Influences*, presented at 66th Annual Meeting of the American Academy of Forensic Sciences (AAFS), February 17, 2014, Seattle, Washington.
[53] Sulner, A. (2014). Examining Sources of Bias and Illustrating their Impact on Handwriting Opinions and Testimony of Forensic Document Examiners. Presented in 2014 AAFS *Bias in Forensics* Workshop, *supra* note 52.
[54] Thompson, W.C. (2009). Observer effects, context effects and confirmation bias in forensic science. In Jamieson, A. (Editor) and Moenssens, A. (co-Editor) *Wiley Encyclopedia of Forensic Science*, John Wiley & Sons, Ltd, Chichester, pp 1575–1579.
[55] Thompson, W.C. (2011). What role should investigative facts play in the evaluation of scientific evidence?, *Australian Journal of Forensic Sciences* **43**, 123–134.
[56] NAS Report (2009), *supra* note 1, Recommendation 5.

Related Articles

Interpretation: Observer Effects; **Handwriting and Signatures, Interpretation of Comparison Results**

ANDREW SULNER

In the Matter of Certain Disciplinary Charges Preferred by DANIEL P. GUIDO, Commissioner of the Westchester County Department of Public Safety, Charging Party, -against- Police Officer WILLIAM P. SHAUGHNESSY, Charged Party

[NO NUMBER IN ORIGINAL]

WESTCHESTER COUNTY DEPARTMENT OF PUBLIC SAFETY

1983 Extra LEXIS 3

November 18, 1983, Decided

[*1] JOHN D. RYAN, Hearing Officer.

JOHN D. RYAN

REPORT AND RECOMMENDATION OF HEARING OFFICER

Pursuant to Resolution # 5-1982 dated September 29, 1982, and the accompanying letter dated November 24, 1932, the above entitled proceeding was referred to the undersigned for the purpose of conducting a Hearing upon Departmental Disciplinary Charges preferred against Police Officer William P. Shaughnessy, by Daniel P. Guido, as Commissioner of the Westchester County Department of Public Safety. The Hearing Officer was directed to cause stenographic transcripts of the testimony to be taken and after review and analysis thereof to transmit a Report for review and decision by the Commissioner and Members of the Police Advisory Board.

The undersigned, in accordance with such authority and directive, did conduct Public Hearings in said matter and submits the following Report and Recommendation.

APPEARANCES AT HEARING

The parties appeared by counsel at each stage of the proceedings.

Charging Party: (hereinafter County) Samuel S. Yasgur, Esq., County Attorney (by Antoinette McCarthy, Esq., Assistant County Attorney).

Charged Party: (hereinafter Respondent) Grae & Rose (by Arthur [*2] Grae, Esq.).

Stenographers: Ms. Tammey M. Pastor

Ms. Donna DeSerio

Ms. Amy E. Sikora

SUMMARY OF PLEADINGS

Nature of Pleading	Date	Submitted by
Charges & Specifications (# 1-3)	9/30/82	County

SUMMARY OF PLEADINGS

Nature of Pleading	Date	Submitted by
Answer, Demand for Bill of Particulars, Demand for Discovery	10/1/82	Respondent
Demand for Designation of Hearing Officer, Demand for Speedy Hearing	10/1/82	Respondent
Bill of Particulars & Discovery	10/7/82	County
Omnibus Motion	10/13/82	Respondent
Charges & Specifications (# 4)	12/9/82	County
Answer, Affirmative Defenses, Demand for Bill of Particulars, Demand for Discovery	12/10/82	Respondent
Bill of Particulars & Discovery	12/16/82	County
Affirmation in Opposition to Omnibus Motion	12/20/82	County
Memorandum of Law (Supporting Admissibility of Polygraph Evidence)	12/21/82	County
Memorandum of Law (In Opposition to Polygraph Evidence)	12/21/82	Respondent

SUMMARY OF CHARGES

Respondent was charged with four (4) specifications of misconduct in two (2) sets of charges dated September 30, 1982, and December 9, 1982. In essence, it was alleged that Respondent with the intent to defraud, harass, annoy, and alarm others, falsely and fraudulently [*3] completed approximately ten (10) mail subscription forms for various magazines and book clubs. The said forms, it was alleged, were completed in the names of Stephen Fischer and James Fleming (two of Respondent's superior officers), as well as one Josie Fleming (Mr. Fleming's spouse).

It was further alleged that Respondent was part of a conspiracy, the actions of which resulted in the completion of approximately two hundred (200) such mail subscription forms. As a result, hundreds of books, magazines and similar items were sent to the above-mentioned individuals as well as one John Castle, another police officer. Far more serious was the allegation that the conspirators (one of which Shaughnessy was alleged to be) cut the lug-nuts on Fleming's personal automobile, which resulted in the wheels falling off the automobile.

The activities indicated were alleged to have persisted throughout the year of 1981 (i.e., between February and December), and were apparently in retribution for the involvement of the targets in the so-called "cooping investigation".

STATEMENT OF APPLICABLE LAW

The procedures concerning a Disciplinary Proceeding of this type are set forth in Section 75 Civil [*4] Service Law.

All procedures followed herein have been taken in accordance with the mandate and requirements of said Law.

It is further to be noted that this proceeding, being in the nature of an Administrative Hearing, "the burden of proving misconduct shall be upon the person alleging the same".

The sanctions that may be imposed are solely within the province of the Commissioner (Charging Party) upon a "guilty" finding after a Hearing, as set forth in Section 75 # 3, Civil Service Law.

The "burden of proof" in Disciplinary Proceedings must be established by substantial evidence - which is such relevant, credible, probative and logical evidence that persuades a fair and detached finder of the fact to reach his conclusions based upon consideration of the entire Record. In this respect it differs from the usual rule in civil actions which requires "proof by a preponderance of the credible evidence" (For principle involved; see: *Gramatan Avenue Associates v. State Division of Human Rights, 45 N.Y. 2d 176,*

Pell v. Board of Education, 34 N.Y. 2d 222.

In Displinary Proceedings, the charges must not be based on frivolous, vague [*5] or trivial matters, nor on whim, caprice or subterfuge. They generally pertain to matters, actions and procedures of an employee indicating neglect of duty; inadequate performance or failure to comply with rules and regulations governing the employment.

(*People ex rel Van Tine v. Purdy, 221 N.Y. 396*
People ex rel Long v. Whitney, 143 App. Div. 17
127 Supp. 554

Griffin v. Thompson, 202 N.Y. 104
Mc Millan v. Morganthau, 146 Misc. 588
263 Supp. 568).

The Hearing Officer, having observed and heard the various witnesses, is in a better position to evaluate their testimony and weigh the credibility, relevancy, and sufficiency thereof. Where a conflict in the evidence exists, the Hearing Officer may accept that version of the testimony offered, supporting his findings.

(*Goddeau v. Levitt, 56 App. Div. 2d 681,* 391 Supp. 2d 745.

Nolan v. Comptroller, 59 App. Div. 2d 799, 398 Supp. 2d 771).

HEARING DATES

Public Hearings were held before me with respect to these charges on:

December 21, [*6] 1982
January 26, 1983
February 11, 1983
February 14, 1983
February 24, 1983
February 25, 1983
March 30, 1983
April 22, 1983
May 16, 1983
June 30, 1983
July 6, 1983
July 11, 1983
July 13, 1983

July 21, 1983

A certified stenographic transcript of each Hearing date, consisting of a total of sixteen hundred eight (1608) pages, is forwarded with this report. Due to the extraordinary length of the transcript, reference will be made throughout this report to the particular page in the transcript in which testimony was received. This procedure will assist the reader in locating a particular point of interest.

EXHIBIT SUMMARY

There were two hundred thirty-seven (237) exhibits and sub-exhibits marked by the parties in this particular hearing. Of that total two hundred Twelve (212) exhibits and sub-exhibits were received in evidence. There were an additional two (2) exhibits marked and received by the hearing officer.

One hundred sixty (160) exhibits and sub-exhibits were marked by the County. Of these, one hundred fifty-eight (158) were received in evidence, with two (2) marked for identification only. The Respondent marked seventy-seven (77) exhibits and sub-exhibits. Of these, fifty-four [*7] (54) were received in evidence and twenty-three (23) were marked for identification purposes only.

A complete exhibit and sub-exhibit list with respect to each party is annexed hereto and made a part of this report.

The key exhibits are as follows:

Item	Number
10 Fraudulent mail subscription forms	County 19-28
Handwriting exemplars of Respondent	County 29-47
Polygraph Chart by County's expert	County 53
Reports of County's handwriting expert	County 55-56 & 68
Enlargement by County's handwriting expert	County 57
Additional exemplars of Respondent re: Fleming & Castle	Respondent R-S
Polygraph charts by Respondent's expert	Respondent I-L
Enlargements by Respondent's handwriting expert	Respondent U-W Y-Z AA-II

Witness List

By the County

Name	Topic	Date of Testimony
Lt. Stephen Fischer	a victim of harassment	1/26/83
Lt. James Fleming	a victim of harassment	2/14/83 2/25/83

Item	Number	
Andrew B. Heberer	polygraph expert	2/11/83 2/24/83
Carl J. Raichle	handwriting expert	6/30/83 7/6/83 7/11/83
	By the Respondent	
Dr. Barry Kaufman	polygraph expert	5/16/83
Terry A. Loftus	postal inspector who assisted in the investigation	7/11/83
Andrew Sulner	handwriting expert	7/13/83 7/21/83
P.O. Garrett Morrison	a reluctant witness suspected to be one of the responsible parties	7/13/83
P.O. William P. Shaughnessy	Respondent	7/21/83

[*8]

SUMMARY OF TESTIMONY

Lt. Stephen Fischer - witness for the County, testified substantially as follows:

He has been a police officer with the Westchester County Department of Public Safety for twenty (20) years. He has been a lieutenant for approximately twelve (12) years and is currently assigned as a desk officer at police headquarters in Hawthorne. In January, 1980, Lt. Fischer was placed in charge of internal affairs and served in that capacity until January, 1981. At that time, the theretofore clandestine investigation of police officers sleeping on the job (so-called "cooping investigation") became a matter of public knowledge. Although Fischer was not involved in this investigation, the apparent belief of his involvement due to his position in internal affairs led to his receiving unsolicited mailings in the nature of newspapers, magazines and books. Lt. Fischer began receiving these items in February, 1981, at his home, former office, as well as the Hawthorne headquarters. He received approximately ninety-five (95) unsolicited mailings throughout the course of the year. As each unsolicited item was received, it was turned over to the then Sargeant (now Lieutenant) [*9] James Fleming, the officer in charge of cancellation.

Lieutenant Fischer was shown eight (8) mail subscription order forms (Co. Ex. 19-26) which Shaughnessy was alleged to have authored. Fischer also testified that he did not sign or give anyone else permission or authority to sign those particular order forms.

Fischer further testified that his initial reaction of annoyance evolved into a deep concern over possible damage to his credit rating. He also stated that in December, 1981, his personal mail had been re-directed to the State of Alaska for approximately three (3) weeks. Understandably, his reaction was intense annoyance.

Lieutenant Fischer has known the Respondent officer for approximately ten (10) years. Respondent had been assigned to Fischer's platoon on a number of occasions. Lieutenant Fischer evaluated the Respondent (p.161) as an above average, good worker who has never been the subject of a civilian complaint. He further stated that Shaughnessy got along well with his co-workers and was not the type of officer who needed constant supervision. Although Respondent has been the recipient of constructive criticism by Lieutenant Fischer, he received it well. Fischer believed [*10] his relationship with Shaughnessy was a good one and without animosity (p.163, 173).

There was, however, one incident during the summer of 1980 in which Lieutenant Fischer was asked by Commissioner Delaney to investigate a so-called "moonlighting" position held by Shaughnessy. This evening job had been approved by the then Deputy Commissioner Fulgenzi. Fischer testified that his investigation resulted in Shaughnessy being directed to terminate this position. Shaughessy fully cooperated with Fischer during this investigation and was upset over being forced to discontinue this employment. His animosity, however, was never directed at Fischer.

Lt. James Fleming - witness for the County, testified substantially as follows:

That for the past ten (10) years he has been employed as a police officer with the Westchester County Department of Public Safety and has been a Lieutenant since January 17, 1983. He is currently the commanding officer of the Internal Affairs Unit.

In October, 1980, he was assigned to Staff Services and participated in the "cooping investigation" which resulted in charges being brought against twenty-two (22) officers.

Apparently as a result, commencing in February, [*11] 1981, he began receiving unsolicited mailings. He received approximately seventy-eight (78) mailings at his home and office in his own name and in that of his wife. Lieutenant Fleming further testified to an incident occurring on January 13, 1981 (p.326) in which he experienced a blow-out of his right front tire just prior to his entering onto the Taconic State Parkway. Thereafter, at the gas station, he noticed that none of his four tires would accept air and that four valve stems had been snapped. On April 3, 1981, while on the Taconic Parkway, he observed a snapping sound (p.328) from his left rear tire. He had to drive onto the grass to avoid an accident. The incident occurred because the lug-nuts on the vehicle had been cut.

Fleming described his reaction throughout this long ordeal(p.329) as extreme annoyance, fear for his credit rating, fear for his personal safety and for that of his family.

In the spring of 1981, while at the office of his personal attorney, Lieutenant Fleming had a conversation with the Respondent (who apparently was at this office on unrelated business). A conversation occurred between the two: (p.337) and Shaughnessy expressed an opinion that the mailings [*12] would never stop.

Lieutenant Fleming further testified with regard to his involvement in the cancellation of over two hundred (200) unsolicited mailings (p.340) as well as the procedure utilized with respect to forwarding evidence to the postal authority's crime laboratory (p.346).

On cross examination, Fleming indicated his awareness of a good deal of animosity (p.391) due to his involvement in the investigations. He was receiving the "silent treatment" (p.392) from many other officers. He indicated that he and the Respondent had been friendly in the past (p.401). They shared the same car pool to weapons school (p.440). He classified his relationship with Shaughnessy as a "working relationship", not a social one. (p.409).

Lieutenant Fleming also testified that he has known postal inspector Loftus for approximately five (5) years (p.437). They have a friendly working relationship which in the past had been primarily concerned with stolen welfare checks.

Fleming became aware of the indictment and plea of guilty of Police Officer Robert Duncan in this matter (p.442) and sent portions of Police Officer Garrett Morrison's personnel file containing handwriting to the County's handwriting [*13] expert. However, (p.445), this handwriting was never compared to the questioned documents (mail subscription forms alleged authored by Shaughnessy) (Co. Ex 19-28). Fleming testified that he sent these examples of Morrison's writing because the Respondent had claimed that his writing and Morrison's were similar (p.446-7). The comparison between Morrison's writing and the questioned items was not done because Shaughnessy had already been identified as the author of those items (p.448).

Andrew B. Heberer - witness for the County, County's polygraph expert, testified substantially as follows:

Mr. Heberer's exposure to the polygraph began in 1961. While a police officer in Nassau County, he was assigned to the polygraph section. There he studied for ten (10) weeks under the supervision of two (2) other officers who six (6) months prior had completed a polygraph course. The said ten (10) week period was essentially on the job training during which his administration at polygraph examinations was supervised by the two officers. Thereafter, he performed unsupervised tests for the Nassau County Police Department until 1971. In 1971, he became employed by Industrial Security Analsy, a [*14] private corporation. He began private practice, initially with partners and in 1975 solo. Since 1961, he has administered approximately twenty thousand (20,000) polygraph examinations. He has performed tests for

the Suffolk County District Attorney's Office, The Legal Aid Society, The United States Navy and Air Force. On three (3) occasions he has testified in court with respect to polygraph examinations he conducted. Although he is a member of the New York State Polygraph Association and The Police Polygraph Association (p.214), no qualifying examination is required for either association (p.278).

Mr. Heberer testified that on September 8, 1982, he conducted a polygraph examination of the Respondent. The particular machine utilized was known as a Stoelting Deceptograph (p.221) and the particular technique employed was known as the "known lie technique" (p.263). Basically, with this technique, the subject is asked a short series of irrelevant, relevant and control questions (Co. Ex 52).

The control questions are the so-called "known lies". The machine produces a chart called a polygram (p.231 Co. Ex. 53). The machine measured breathing (pneumo), blood pressure, heart beat (cardio), [*15] and galvanic skin responses (GSR) (p.221). Mr. Heberer described the interpretation of the polygram as a visiual comparison of the reactions indicated on the polygram. If a subject shows more reaction to a control question (i.e., known lie), than a relevant questions, he is being truthful. If he shows more reaction to a relevant questions than a control, he is practicing deception.

Mr. Heberer compared the reactions of the Respondent with regard to the relevant and control questions and was of the opinion that deception was shown on the polygram with respect to Shaughnessy's responses to relevant questions (p.604). He explained that the test cannot state that the Respondent actually completed and/or mailed the subscription forms in question, but can state, by analysis, that deception was shown (p.631).

Carl J. Raichle - witness for the County, County's handwriting expert, testified substantially as follows:

Mr. Raichle is a document analyst employed by the U.S. Postal Service Crime Laboratory in New York City (p.859). He was a police office for fifteen (15) years, was assigned to the New York City crime laboratory in 1973, and has been with the Postal Service since 1977. [*16] He has received a B.S. in Police Science from the John Jay College of Criminal Justice. Mr. Raichle was given intensive training by experienced document examiners while at the New York City Crime lab. He also attended F.B.I. Special Scientific Training School in Virginia. Raichle is a member of numerous forensic science organizations and has been certified by the American Board of Forensic Document Examiners. Furthermore, he has testified in federal and state courts throughout the northeastern portion of the United States, and has been qualified in court as an expert approximately thirty (30) times.

In December, 1981, he examined the questioned documents (Co. Ex. 19-28), and compared those to samples of the Respondent's handwriting consisting of Shaughnessy's memo book and personnel folder (Co. Ex. 46-47) (see p.870). He initially concluded that Respondent "probably wrote" exhibits 19-24 and 26. Mr. Raichle further concluded that "it was conceivable that Shaughnessy wrote 25, 27 and 28 (p.873). These findings were made part of Raichle's initial report on this case dated February 8, 1982 (Co. Ex. 55). Thereafter, Respondent requested and received certain handwriting exemplars of [*17] Shaughnessy which were taken on March 19, 1982 (Co. Ex. 29-45). The questioned documents, memo book, and personnel folder were re-submitted to Mr. Raichle along with the new exemplars. As a result of the availability of additional samples of Respondent's handwriting (p.885), Raichle was now able to render an opinion that Shaughnessy was the author of each of the questioned documents, and issued a report (Co. Ex. 56) to that effect dated June 10, 1982, (p.887).

In addition, on May 9, 1983, additional unsolicited mail subscription forms (Co. Ex. 58-65, 69 and 70) were presented to Raichle for his analysis. While authorship of these particular questioned documents did not constitute part of the original charges against Shaughnessy (specifications 1-3), they were admitted on the issue of conspiracy (p.909) and were ruled admissible on two (2) theories (p.919).

Mr. Raichle concluded, after an examination of these items, that at least four (4) other authors were involved in the drafting of those particular documents, and that Shaughnessy wrote exhibits 60 (p.934), 69 (p.947) 70C, 70D, and 70E (p.948). These were, in other words, unsolicited mail subscription forms additional to the original [*18] ten (10) charged (i.e., Ex. 19-28). Mr. Raichle then rendered a report (Co. Ex. 68) to that effect, dated June 21, 1983.

Terry A. Loftus - called as a witness for Respondent, appeared under subpoena, and testified as follows:

He has been a postal inspector for twelve (12) years and investigates crimes involving the mails (p.1089). On March 19, 1982, he met with the Respondent at the Westchester County District Attorney's Office for the purpose of taking handwriting exemplars (p.1090) to assist Mr. Raichle in his analysis. He directed Shaughnessy to write out cer-

tain names, addresses, zip codes, and to complete sample mail subscription forms. However, he did not direct Shaughnessy as to the proper spelling of any names (p.1093). These sample mail subscription forms were completed in the names of Stephen Fischer, James Fleming, Josie Fleming and John Castle. These constituted all the handwriting exemplars of Shaughnessy.

The exemplars signed in the name of Fischer were submitted to Mr. Raichle to assist in his analysis. The exemplars of Respondent signed in the name "Fleming" (Respondent Ex. R) and in the name of "Castle" (Respondent Ex. S), were never sent by Loftus to Mr. [*19] Raichle for his analysis, but were kept in Loftus' file until just prior to his testimony (p.1094, 1096, 1099).

Loftus insisted that he never made a visual comparison (p.1000) between the "Fleming" and "Castle" series of exemplars (Respondent Ex. R-S), and the questioned documents (Co. Ex. 19-28); not did he make a comparison between the exemplars submitted to Mr. Raichle (Co. Ex. 29-45) and the other exemplars. He indicated that his failure to submit the "Fleming" and "Castle" series of exemplars was due to the fact that the Respondent was not a primary target on Fleming (p.1110); he was a suspect only on Fischer. (note- Shaughnessy was initially charged with writing two (2) questioned documents in the name Fleming - Co. Ex. 27-28). Loftus went on to add that he did not deliberately withhold evidence (p.1113). He maintained this position even though he later saw a report by Raichle naming Shaughnessy as the writer of two (2) items in the name Fleming (p.1116). Furthermore, while he later submitted items to the lab in the name of Castle and Fleming (p.1129) it never crossed his mind to send the exemplars of Shaughnessy taken in the names "Castle" and "Fleming" to the lab (p.1122).
[*20]
P.O. Garrett T. Morrison - called as a witness for Respondent, appeared under subpoena, and testified substantially as follows:

That he is employed by the Westchester County Police, and has been so employed for ten (10) years. (p.1390). In 1981-82 he was assigned to the communications room in headquarters and in such position held access to the records and reports of various police officers (p.1390). He was shown each of the original questioned documents (Co. Ex. 19-28) and denied authorship of each (p.1391).

He was then offered the opportunity by Respondent's attorney to give handwriting exemplars, and expressed a desire to speak to his attorney before doing so (p.1393). Morrison later appeared (p.1342) by counsel, exercised his rights pursuant to the Fifth Amendment to the United States Constitution and refused to give exemplars without a court order.

Dr. Barry Kaufman - called as a witness for Respondent Respondent's polygraph expert, testified substantially as follows:

That he is the executive vice-president of Fargo Overland Corp., New York City, is in charge of their polygraph department and is its chief examiner (p.677). Fargo is involved in the business of security [*21] investigation and lie detection. Kaufman has a B.A. in political science from CUNY, an M.A. in Criminal Justice from C.W. Post, and a PhD in public administration from the City University of Los Angelas. He is a graduate of the Backster School of Lie Detection, San Diego (p.678), the Dektor School of Lie Detection and Counter Intelligence Savannah, and the F.B.I. Criminal Investigations School. He is a certified and licensed (in Vermont) polygraph examiner (p.680). Dr. Kaufman has conducted over ten thousand (10,000) polygraph examinations (p.723). He established the truth verification program and taught as an assistant professor at the New York Institute of Technology, Old Westbury (p.681). Dr. Kaufman is the author of various articles(p.682) and has testified several times in court (p.683). He described the polygraph machine (p.683), what it measures (p.684), as well as the numerical scoring technique employed by his firm (p.684).

On February 8, 1983, he conducted a pre-test interview of the Respondent (p.686), formulated a set of appropriate questions (Respondent Ex. O) (p.693), and conducted polygraph examinations of the Respondent. These examinations resulted in polygrams (Respondent [*22] Ex. I, J, K, L) which were then interpreted by Dr. Kaufman. Kaufman was of the opinion that Shaughnessy testified truthfully when he denied knowledge of, or involvement in, an accident concerning Lt. Fleming's automobile. Furthermore, with respect to Shaughnessy's involvement in authoring or sending the questioned documents, Kaufman's opinion was that the Respondent was testifying truthfully when he denied any involvement (p.759).

Dr. Kaufman described the proper calibration (p.699) of the machine, the significance of formulating the questions properly (p.724), as well as the procedure which should be utilized in the proper administration of a polygraph examination. He described the test given to Respondent as a"direct involvement" test, the most precise available (p.737). Dr. Kaufman also extensively criticized the procedures utilized by Mr. Heberer in his administration of his polygraph ex-

amination to the Respondent. He categorized this test as completely invalid (p.785-7), and indicated his belief that Mr. Heberer's machine was not working properly (p.798).

Andrew Sulner - called as a witness for Respondent, Respondent's handwriting expert, testified substantially as follows: [*23]

That he is a forensic document examiner (p.1224) with offices in New York City. He has received a B.A. from Queens College, J.D., from the National Law Center, George Washington University, as well as an M.A. in forensic science from the same university. He was the first person in the United States to have received his J.D. and M.A. (in forensic science) concurrently. Mr. Sulner is the partner and son of Hanna Sulner who is apparently a world renouned and heavily published document examiner. He has been a forensic document examiner for sixteen (16) years. His firm has been retained by most of the major law firms in New York City (p.1231), and he has worked together on cases with such legal notables as Edward Bennett Williams (p.1238). Mr. Sulner's firm has also been employed by the various District Attorneys' offices in the area, the United States Attorneys Office, Treasury Department, United Nations, Office of Court Administration (New York), the Attorney General's Office (New York) and the American Broadcasting Company (p.1232). He is a member of numerous bar and forensic science associations (p.1242), and has lectured widely throughout the United States (p.1229).

Mr. Sulner was [*24] initially contacted by Respondent's attorney in October, 1982, and was asked to examine the questioned documents (Cp. Ex. 19-28), conduct an analysis and render an opinion as to whether there was any evidence to suggest Shaughnessy wrote these items (p.1250). He discussed the distinctions between "natural variations" and "divergent characteristics" (p.1256). He demonstrated the principle that similarity does not mean identity (p.1261), and that identification is based upon the formation of unconscious writing habits unique to each particular author. Mr. Sulner discussed and demonstrated numerous dissimilarities between the questioned documents and the writing of Shaughnessy (See analysis section of this report for a detailed review). He prepared numerous charts and exhibits (Respondent Ex. U through II) to assist his presentation.

Mr. Sulner's opinion was that there was absolutely no evidence that Shaughnessy wrote any of the questioned documents (p.1254). He further stated there was absolutely no evidence to suggest Shaughnessy wrote any of the uncharged mail subscription forms, authorship of which was attributed to him by Mr. Raichle (i.e., Co. Ex. 60, 69M, 70C, 70D, 70E [*25] - admitted as proof of specification # 4, conspiracy). Mr. Sulner testified that County exhibits 61-64 were written by P.O. Robert Duncan (Duncan has been criminally convicted for his involvement), and that two (2) of the questioned documents (Co. Ex. 27-28) appear to have been written by Duncan (p.1374 and p.1382). He went on to indicate numerous significant, unconscious similarities between the remaining questioned documents (Co. Ex. 19-26), and the handwriting of P.O. Garrett Morrison (p.1355 and p.1368). Mr. Sulner, due to lack of sufficient available exemplars from Morrison, would not positively identify Morrison as the author of these items (p.1355). However, he did testify that due to the number and probative weight to be assigned to the numerous similarities, any indication of authorship is on Morrison(p.1372). Shaughnessy, he testified, is, beyond all doubt, not the author (p. 1372).

William Shaughnessy - the Respondent, testified on his own behalf substantially as follows:

That he has been a police officer ten and one-half (10 1/2) years and has, prior to these occasions, never been the subject of disciplinary charges (p.1552). He is married, the father of two (2) [*26] children, and a graduate of Westchester Community College with an A.A. in Police Science. He served for three (3) years in the United States Army with tours of duty in Germany and Viet Nam. Mr. Shaughnessy received several military citations including a Bronze Star (p.1555). He has received five (5) citations for exceptional police service as well as numerous meritorious duty awards (p.1556).

Mr. Shaughnessy was shown the questioned documents (Co. Ex. 19-28), as well as the additional mail subscription forms admitted on the conspiracy charge (Co. Ex. 69 and 70 CDE). He denied being the author of mail subscription forms (p.1557) and denied being part of any conspiracy as alleged in the specifications herein.

In regard to the accidental meeting with Fleming at the office of their common attorney, Mr. Shaughnessy catagorized the conversation as friendly and somewhat extensive. It lasted nearly fifteen (15) minutes. He stated it was common knowledge, at that time, that Fleming was receiving unsolicited mail. During part of their conversation, Shaughnessy expressed his feeling that these mailings were probably not going to stop due to the anger of numerous officers over the manner in which [*27] the "cooping investigation" was conducted (p.1564). He indicated there were wide discussions among the officers on this topic; however, he never overheard anyone speaking of revenge (p.1574).

ANALYSIS OF THE TESTIMONY

A detailed review of this entire transcript reveals no eye-witness account of Respondent's involvement in any of the activities which formulate the basis for the charges herein. Furthermore, there were no admissions made by Respondent which inculpate him in any way. The closest evidence of admission were the slightly differing accounts of the conversation between Fleming and Shaughnessy, which occurred at the office of their common attorney. The distinctions consisted more of inflections of voice than of substance. Even examined in a light most favorable to the County, the wording, used by Fleming to relate Shaughnessy's conversation, amounts to no more than an opinion than the mailings would persist longer than Fleming realized. Since these unsolicited mailings were a matter of common knowledge among police personnel, the relating of such an opinion, without more, does not constitute an inculpatory statement.

The testimony of P.O. Morrison is accorded little, if [*28] any, value. He exercised his rights to have an attorney present and refused to execute exemplars without a court order, on Fifth Amendment grounds, which he has an absolute right to take. The exercise of such rights before this hearing officer in no way constitutes evidence of any kind or an implication of involvement. Accordingly, the exercise of these rights, before me, was not considered in any manner in this decision.

Although I find highly suspect the explanation of Mr. Loftus pertaining to his failure to submit the "Castle" and "Fleming" exemplars to Raichle, the evidence is not sufficient to establish a deliberate withholding of exculpatory material. The County was without knowledge and in no way responsible for Mr. Loftus' actions. Furthermore, since these exemplars could have been utilized in an analysis of only two (2) of the questioned documents, the failure to transmit these items to Raichle is of little consequence to the ultimate determination of these charges.

As a result, this case turns on a proper analysis of the expert testimony submitted by each party. The County's experts indicated Respondent was deceptive in the polygraph and that he authored the questioned documents. [*29] Respondent's experts say otherwise, My review of the significant testimony and its analysis is as follows:

The Polygraph

The proper administration of a polygraph examination, to simplify, requires the proper formulation of questions, a properly functioning machine, the following of certain procedures by a trained operator, and an effective reliable method of interpreting and comparing the involuntary reactions to relevant and control questions. It is essential when utilizing the "known lie" technique to obtain a "no" response to the control question (p.742). Unless this occurs, one cannot compare the reaction to the known lie to a lie on a relevant question. Furthermore, there should be a sufficient time-spacing between questions to allow the subject to react or relieve (p.690). Fifteen (15) to twenty (20) seconds is ideal. In following this procedure, you avoid a reaction to former questions while asking the latter (p.860).

A review of the testimony of Dr. Kaufman (Respondent's polygraph expert) reveals that he is not only highly experienced, but has been extensively and formerly trained in the administration of the polygraph. His formulation of the questions, (p.694) especially [*30] the control questions (p.696), was done in accordance with excepted procedures. He utilized a Stoelting Electronic Instrument (p.697), manufactured in 1981, which produces clear charts (Respon. Ex. I-L). Furthermore, the pneumo, cardio and GSR tracings were calibrated to assure the machine was functioning properly on each tracing (p.699). The test was conducted in a sterile environment, as is the recommended procedure (p.701). Dr. Kaufman used a numerical system of grading in which you actually measure the reaction in each parameter using a plus (+) and minus(-) procedure (p.743). A score on the positive side is an indication of truthfulness with a "definite truthful" at plus twelve (+12) or higher. The reverse is true for a negative score (p.743). You must utilize a numerical procedure for your test to be used in Court (p.744). Furthermore, Dr. Kaufman manually adjusted his machine after each reaction (p.748). While this is not absolutely required, leaving a machine on automatic restricts the free flow of the pens (reaction indicators).

The reaction of Shaughnessy in the tests was strongest in the control questions. He showed no reaction to the relevant questions (p.742 and 748). [*31] His score on the questions concerning the lug-nut incident was a plus thirteen (+13) and on the questions pertaining to the authoring and mailing of the questioned documents, a plus fourteen (+14) (p.759). Both results indicate he was "definitely truthful" in his responses.

Examination of the polygrams resulting from Dr. Kaufman's tests were carefully examined. While this hearing officer is, of course, not an expert in the polygraph, it clearly appears that there was a greater reaction to the control questions (those numbered in the 40's) as opposed to the relevant (those numbered in the 30's). At any rate, Dr. Kaufman was a highly credible witness.

Mr. Heberer (County's Expert) also has experience in the administration of the polygraph. While has has administered an estimated twenty thousand (20,000) tests over the course of twenty-two (22) years, quantity is not always synonomous with quality. Although he has never been formerly trained in the administration of the polygraph, it does not appear that extensive training is required to administer a proper test. However, a test conducted in a slipshod manner, lacking precise formulation of questions, interpreted without a scientific [*32] numerical scoring system, which proceeded in the face of obvious malfunction in his equipment is accorded no weight by this hearing officer.

Mr. Heberer used a Stoelting Deceptograph which was several years old, not the most advanced machine available; it does not produce an electronically enhanced tracing. However, the machine, when properly utilized, is accurate. There are methods of calibrating each tracing on his machine prior to conducting a test of subject. The GSR tracing can be calibrated by bringing the two heads together (p.293) and pressing the "IK" button. This was not done. There is an additional GSR check on Heberer's machine known as the "5K" button. Mr. Heberer not only testified he has never used this check, he doesn't even known how it works (p.294). There is a procedure which can be utilized to calibrate the pneumo (breathing tracing) (p.295). You can adjust this tracing by use of a centering knob. Although Mr. Heberer indicated that false readings were possible when calibration is not done (p.297), he did not deem it necessary to calibrate this tracing.

In the administration of the test, when attaching the cardio cuff, the examiner should massage the cuff to release [*33] any air bubbles and wait two (2) minutes before proceeding. This was not done (p.299). This resulted in a cardio tracing (the bottom tracing on Mr. Heberer's chart) which can barely be seen, let alone interpreted. While Mr. Heberer concedes that a dichrotic notch systolic stroke are needed in a cardio tracing (p.462), the dichrotic notch was not visible, nor was there a clear systolic stroke on the cardio tracing in his first test. (Note - it is the proper procedure in the industry to render an opinion on the basis of at least two (2) tests). Yet, Heberer insisted this tracing could be interpreted, although admittedly, with great difficulty (p.463).

It is the approved method in the industry to manually adjust the GSR after each response (p.453). While it is not required, and one can leave the tracing on "automatic centering", doing so restricts the free flow of the pens which in turn decrease the observable responses (p.466 and 470). Obviously, this makes interpretation more diffcult. Mr. Heberer chose automatic centering (p.464).

Although Mr. Heberer indicated that it would be beneficial to have obtained a pneumo tracing of three quarter (3/4) inch to one (1) inch for interpretation, [*34] the tracing on his upper pneumo indicator is only one-sixteenth (1/16) inch to one-eighth (1/8) inch (p.472). Since Heberer used a double pneumo indicator, this defect, in and of itself, would not be fatal.

However, the most grievious defect in Mr. Heberer's procedure was the fact that he obtained and utilized a "yes" response to one (1) of his two (2) control questions (p.475). As indicated above, it is absolutely essential in the "known lie" test to obtain a "no" response to the control questions. This "no" response is the "known lie" and the parameter against which a lie on relevant questions is measured. Utilizing a "yes" response to formulate an opinion here is illogical, without scientific foundation, and completely contrary to the theory upon which the "known lie" polygraph technique is based.

The numerous other defects in Mr. Heberer's procedure, such as lack of precision in the formulation of his questions, transference, lack of sterile environment, inadequate time between questions, lack of numerical scoring and proper chart markings, will not be discussed at length herein.

Dr. Kaufman was of the opinion that Heberer's first test and his second test (p.785 and 787) are invalid. [*35] I agree.

It is my opinion that Mr. Shaughnessy passed the only valid polygraph examination taken.

Handwriting Analysis

In handwriting analysis, identification occurs when the expert, after accounting for natural variations (p.1256), can show unconscious writing characteristics present in the "questioned" and "known" writings of an individual, plus no significant differences between the two (p.1256). Similarity does not mean identity (p.1261). Numerous similarities in the writing of many individuals are often based upon class characteristics. These find their origin in the common initial instruction received by most of us in penmanship and related courses. Natural variations in the writing of the same individual occur consistently (p.1256). The average person simply does not write in the exact same manner twice. Once the search for and location of natural variations has been completed, analysis centers on assigning the proper weight to similarities and dissimilarities. If these characteristics are unconscious and extremely unique, a single chacteristic can

result in either identification or exclusion of a subject. Although there were many similarities and dissimilarities, [*36] neither the known or questioned writings revealed a single chacteristic that unique. Each expert agreed that the formulation of the letters,unconscious characteristics, and rare combinations are entitled to great weight. Misspellings and improper use of upper and lower case letters were entitled to less weight, and similarities with class characteristics were entitled to little, if any weight.

In the remainder of this section, I will review, in detail, the significant portions of the testimony of each handwriting expert. This was, without a doubt, the most significant and probative part of the case.

Mr. Carl Raichle - testifying on behalf of the County, (p.896), and identifying Shaughnessy as the author of the questioned writings, demonstrated many similarities between the known writings of Shaughnessy and the questioned writings.

The similarities (p.896), along with comments as to weight (p.1209) of each characteristic, are as follows:

1) use of lower case "e" and capital "R" in the word Fischer. He felt this combination was very conspicuous (p.1029).

2) The capital "L" in the middle of the word police - Although there was a slight curve in the upper part of the [*37] L in the questioned writing, both known and questioned were formed by a downstroke and slight retracing, and the letter was formed above the line. This was rare in the middle of a word (p.1182) and more significant because it only appeared in the word "Police" (p.1185). Raichle felt the formation, casing and use of this letter was so unique as to be very significant (p.1198).

3) "T" crossed on its staff - This was not unique and was a class characteristic (p.1030).

4) Captial "S" was flat on top and had a bow on bottom - (There was extensive comment by Respondent's expert).

5) "O" began and ended at 11:00 - This was also a class characteristic, and not that unique (p.1036).

6) "i" was a short letter - This factor was not unique or fundamental (p.1067). However, he did notice an important difference here. Shaughnessy always dots his "i's", and dots them very close to the stem. Whereas the questioned writings, if dotted, were dotted far to the right and high above the stem (p. 1193).

7) The "c e" combination was similar.

8) The "h" had a downstroke and retracing - This, however, was not unique (p.1039).

9) The pen drag in the "e" to capital "R" in the word "Fischer" [*38] he felt was a significant similarity.

10)The "R" was similar in that a straight, slanted line left the bowl - This, however, was not unique and many people ended their "R's" in this fashion (p.1069). He also noticed differences in the capital "R". In the known writing the originating stroke began outside the upright stroke, and in the questioned, the upright stroke met the finishing stroke (p.1069).

11) The "N" and "F" were three (3) stroke letters - This, however was a class characteristic and not unique (p.1042).

Therefore, to summarize, the most significant similarities observed by Mr. Raichle were the capital "L in Police", the "lower case e and capital R combination in the word "Fischer", and the "pen drag from the lower case e to capital R in the word Fischer".

Mr. Raichle also testified that dissimilarites are very important (p.1010). While there is a general rule that one dissimilarity is enough for exclusion, Raichle doesn't totally agree (p.1012). The difference, he felt, must be significant and unexplained (p.1014). Along with the differences noted in the dots of the "I" and the capital "R", Mr. Raichle later acknowledged the following further dissimilarities [*39] between Shaughnessy's writing and the questioned documents:

1) In the questioned writings the name "Fischer" was consistently misspelled "Fisher". While this was a factor (p.1179), it was not significant enough to exclude Shaughnessy (p.1183).

2) In the word "Stephen" the "P", and in the word "Hawthorne" the "R" and "A" were lower-case in the known and upper-case in the questioned. While this was a factor of some weight (p.1194 and 1198), formation was more important than case (p.1196). As a result, the similarities in the "L" in "Police" were entitled to more weight (p.1198). However, he did indicate that it would be more significant if writings of Shaughnessy that predate the litigation showed this difference (p.1205). They do.

3) Shaughnessy always puts periods after the "N" and "Y" in the abbreviation for "New York". The questioned writing never had periods. Raichle felt this was a significant factor (p.1201) but not enough to exclude Shaughnessy as the author (p.1201).

4) In the known writings of Shaughnessy, the end of the numerical "2", ends straight, whereas in the questioned it has a curl at the end. This, Raichle felt, had some significance (p.1203).

5) The middle [*40] portion of the capital "H" in the known writings slanted downward or was horizontal; it never slanted upward. In the questioned writings, this portion of the letter always slanted upward. While this also was a significant dissimilarity, it was not significant enough to exclude Shaughnessy as the author (p.1207).

6) In the numerical "5", the top portion is always horizontal in the known writings and never in the questioned. This, Mr. Raichle felt, was a factor.

7) The use of the "T" by Shaughnessy indicates that he always crosses it on the stem in the middle of a word and at the top when the "T" ends a word. In the questioned documents the "T" was crossed in the middle. Mr. Raichle felt it was very significant, rare and unusual for an individual to make two (2) different types of "T's", use them interchangeably and always cross at the top at he end of a word (p.1214). However, this also apparently was not significant enough to exclude Shaughnessy.

8) In the known writings, the letter "U" always had a tail at the end, whereas in the questioned writings there was no tail and the "U" was formed like a horseshoe. This, Mr. Raichle felt, was the most significant difference between [*41] the known and questioned writings (p.1209). However, he felt, it was not entitled to as much weight as the similarities in the "L" in police. The similar case, formation and appearance of this "L" letter above the base line was of critical significance.

It is of some interest to note, at this point, that although limited samples were available, P.O. Garrett Morrison also formulated his capital "L" in a similar manner and above the base line (p.1528 and p.1368).

Andrew Sulner - testifying as a handwriting expert on behalf of the Respondent, indicated numerous other significant dissimilarites between the writings of P.O. Shaughnessy and the author of the questioned documents. These distinctions were readily observable when reviewing either the exemplars of Shaughnessy or writings that pre-dated the case. Sulner stressed the importance of unconscious writing habits, unique to the particular author. Similarity itself does not mean identity (p.1261) and is a class characteristic (p.1265).

While the misspelling (sic "Fisher") of the name "Fischer" occurred throughout the questioned writings, it is not, in and of itself, sufficient to exlcude the Respondent (Shaughnessy [*42] always spelled it correctly even in pre-litigation writings). Misspellings can be imposed directly (p.1323). However, when you see this occur, you must examine closely for the unconscious characteristics.

The significant distinctions between the known writing of Shaughnessy and the author of the questioned documents were as follows:

1) The bowl of the "S" in Respondent's writing was always away from the vertical access line. In the questioned writings, the beginning stroke and bowl are touching or very close to the vertical access line. This is a highly significant, unconscious writing distinction also observable in the pre-existing writings of Shaughnessy (p.1291-6).

2) There is consistent misalignment in the "p h" combination by the author of the questioned documents. The "P" is always higher. In Respondent's writing they are aligned. This is an extremely significant distinction (p.1322).

3) In the formation of the "U", the questioned writings, as indicated above, have no terminal stroke, whereas, the known writings always have such a stroke. This, Mr. Sulner felt, was highly significant and indictive of the two (2) authors. It is an unconscious characteristic (p.1300).
[*43]
4) Also highly characteristic and subconscious (p.1303) is the pronounced terminal ending stroke by Shaughnessy in the formation of his lower-case "e" at the end of a word (p.1302). In the questioned writings the terminal stroke of the "e" was brief, abrupt and not extended (p.1300).

5) The alignment and respective heights of the letters in the "e s" combination (p.1304) was one of the most significant differences. In the questioned writings the "e" is always slightly or significantly higher than the "s" (p.1322). Nowhere was the "S" higher. However, in the known writings of Shaughnessy even those which pre-date the litigation, the tendency is reversed. The "S" is always higher (p.1305).

6) Mr. Sulner also noted highly significant distinctions in the formation of the capital "N". In the known writings the first leg is slanted left; the legs are rarely if ever parallel, oneleg extends higher than the other and the bottom of the horizontal downstroke is pointed. In the questioned writings, the portion is rounded and the upswing terminal stroke never extends beyond the height that the originating stroke begins upon. Furthermore, the legs are generally parallel and where not parallel [*44] they go inward (Respondent's go outward) (p.1308). Mr. Sulner felt these distinctions were of the highest significance (p.1322).

Mr. Morrison's formulation of his captial "N" revealed all of the above characteristics (p.1531 and 1368). P.O. Duncan also formulated his "N" in this manner (p.1382).

7) Also of extreme significance (p.1322) and having tremendous identification value (p.1311) was a characteristic found in the "zip codes". In the known writings of Shaughnessy, the zero following the numerical "1" are always equal in height. However, in the questioned documents the zero is always smaller than the "1" which precedes it. This is a highly individualistic writing characteristic, which, by the way, was also found in the writings of P.O. Morrison (p.1531), and P.O. Duncan (p.1382).

8) The formation of the numerical "5" in the questioned writings reveals a number with a fairly prominent neck (vertical center portion), and with a rounded bowl that commences above the baseline. In the known writings of Shaughnessy, the "5's" do not have a neck, the bowl is pointed and the bottom is formed at the baseline (p. 1313-15). These distinctions in formation are highly significant [*45] and indicative of two (2) authors(p. 1322).

Also, the formation of Mr. Morrison's number "5" is strikingly similar to the questioned writings (p.1531), as is P.O. Duncan's (p.1382).

9) Similarly, in the formation of the number "3", the bowl is generally formed above the baseline in the questioned writings. In the writings of Shaughnessy, the bowl not only touches the baseline, but is more pointed. This trait was also consistent throughout the pre-litigation writings of the Respondent and present in none of the "3's"in the questioned writings (p.1316).

10) Mr. Sulner testified that the "i-dot" pattern is an unconscious writing characteristic and highly significant (p.1322) Whereas, Shaughnessy dots his "i" very close to the stem and rarely departs from this (p.1317). The author of the questioned writings dotted his "i" generally high above and significantly to the right of the stem.

Again, the writings of Mr. Morrison showed a similar "i" dot pattern to those in the questioned writings (p.1531).

11) The "T" crossings and center portion of the "H" in the questioned writings reveal an upswing in this marking. Shaughnessy's markings on these letters is downward or horizontal, but [*46] never upward. While this is also an unconscious characteristic, it is not as significant as the others mentioned (p.1318-21).

12) The captial "J" in the questioned documents always had a top to it (p.1329). Further, in Shaughnessy's writings, the "J" was always formulated without a top.

This "J" formation was very similar to that of P.O. Robert Duncan (p.1382) and P.O. Morrison.

13) In the questioned writing, in the formation of the "M", the area in the center where the downstroke and upstroke meet never touches the baseline. Shaughnessy's "M's" in this area always touch the baseline. This, according to Sulner, is indicative of two different writers.

P.O. Morrison writes his "M" very similar to the questioned writings (p.1531).

14) The letter "G" was formulated by Shaughnessy in the traditional manner (p.1230). In the questioned writings, it was formed like a "figure six" (with the backstroke hitting the downstroke). Mr. Sulner testified that this was not only a very significant distinction, but was the most significant similarity between the questioned writings and those of P.O. Morrison.

Mr. Sulner testified further that all of the writing characteristics attributed to [*47] Shaughnessy were present, not only in the exemplars given for this case, but present in the writings which pre-dated the case. Also, more than one person wrote the mail subscription forms attributed to Shaughnessy (p.1541). The subscriptions were uncharged originally, and admitted on the theory of conspiracy,where written by Duncan. In addition, County exhibits 27-28 appear to have been written by Duncan.

Furthermore, with respect to the other items (Co. Ex. 19-26), if there is any indication of a possible author, it was Morrison. Shaughnessy is beyond all doubt not the author of any of these items (p.1372).

However, due to a limited amount of handwriting exemplars by Morrison, Sulner would not make a positive identification. He did acknowledge a distinct similarity between Shaughnessy writings and the questioned documents in the formation of the "L" in the word "Police". However, in view of the overwhelming number of dissimilarities, he felt this factor was insignificant.

I have carefully reviewed both each exhibit and the entire transcript. I must concur with the findings of Mr. Sulner. Aside from the similarities in the "L" in "Police", there do not appear to be any significant [*48] similarites in the questioned writings and those of P.O. Shaughnessy. The numerous distinctions between the two (2) sets of writings compel the conclusion that Shaughnessy was not the author of the questioned writings.

The results of the polygraph examination conducted by Dr. Kaufman support and confirm this conclusion.

FINDINGS

The evidence in insufficient to sustain specifications 1-4. Accordingly, each specification is DISMISSED.

November 18, 1983.

HON. DONALD E. SHELTON

Donald E. Shelton is Associate Professor and Director of the Criminal Justice Studies Program at the University of Michigan-Dearborn. He was a Circuit Judge of the 22d Circuit Court in Ann Arbor, Michigan for over 24 years. He received his PhD in Judicial Studies from the University of Nevada, his M.A. in Criminology and Criminal Justice from Eastern Michigan University, his J.D. from the University of Michigan Law School, and his B.A from Western Michigan University. Dr. Shelton has authored numerous books and articles. His primary research interests include the impact of science and technology on our legal systems, especially the jury system, as well as criminal procedure.

Closing the Gate on Biased Expert Testimony: The Judicial Perspective

Hon. Donald E. Shelton JD, PhD

"I will venture to say that my investigations and decisions are not usually influenced by my hopes and fears." - Mr. Darcy in "Pride and Prejudice"

"I have yet to see a piece of writing, political or non-political, that does not have a slant. All writing slants the way a writer leans, and no man is born perpendicular." — E. B. White

This presentation addresses the issue of bias in the preparation and presentation of forensic science evidence from the judicial perspective. Primarily the issue in the preparation of forensic science evidence is one of "cognitive bias" which is the common tendency to acquire and process information by filtering it through one's own likes, dislikes and experiences". One type of cognitive bias is a "confirmation bias" - the tendency to seek only information that matches what one already believes. Another, is "outcome bias" - the tendency to judge a decision by its eventual outcome instead of based on the quality of the decision at the time it was made. In the context of forensic science expert testimony, some authors have more broadly called this "adversarial bias".

Judges have long recognized the problem of bias in expert testimony. 140 years ago, Sir George Jessel in *Abinger v. Ashton* said "*Undoubtedly there is a natural bias to do something serviceable for those who employ you and adequately remunerate you."* And Learned Hand wrote at the turn of the 20th Century:

> "*The serious objections are, first, that the expert becomes a hired champion of one side . . . . Enough has been said elsewhere as to the natural bias of one called in such matters to represent a single side and liberally paid to defend it. Human nature is too weak for that . . . "*

The Courts started to develop special rules to cope with this "adversarial bias" by experts. Rather than simply leaving the reliability of expert witnesses to the jury, The *Frye* test was developed to make judges the "gatekeepers" to decide which expert evidence was reliable enough for the jury to hear. But nothing in the short *Frye* opinion, or even its progeny, directly addresses the issue of expert witness bias. The *Daubert* trilogy of cases seemed to change the rules dramatically. Judges were indeed instructed to be the "gatekeepers' for forensic science evidence to assure the scientific reliability of that evidence. It replaced the old *Frye* test of "general acceptance", at least in federal courts and set forth a "non-exclusive" list of factors

judges are to analyze at a hearing. The Supreme Court remanded the case for a hearing to reconsider the testimony applying the new factors they announced in the case. But a strange thing happened on *Daubert*'s way back to the trial court. It never got there. It was intercepted by Judge Kozinski on the 9th Circuit who threw out the plaintiff's scientific evidence on a factor he added:

> *"One very significant fact to be considered is whether the experts are proposing to testify about matters growing naturally and directly out of research they have conducted independent of the litigation, or whether they have developed their opinions expressly for purposes of testifying.*

In other words the bias of the expert, based on who was paying the expert, is also a significant factor.

If Judge Kozinski's rationale in the *Daubert* remand were taken seriously in the criminal case context, prosecutors would face a difficult task in getting much of their forensic scientific evidence before a jury. In the criminal forensic science field, most of the testimony has no origin or basis outside of the context of criminal investigation and litigation. It was developed strictly for use by the government to aid in the prosecution of alleged criminal activity in court. But Kozinski dropped in a footnote:

> *"There are, of course, exceptions. Fingerprint analysis, voice recognition, DNA fingerprinting and a variety of other scientific endeavors closely tied to law enforcement may indeed have the courtroom as a principal theatre of operations. . . . As to such disciplines, the fact that the expert has developed an expertise principally for purposes of litigation will obviously not be a substantial consideration."*

Most disturbing is Kozinski's sweeping suggestion that expert testimony in criminal cases is somehow "obviously" so different that the potential bias of the expert is not the "substantial consideration" he thought appropriate in civil cases. If there was any difference to be acknowledged, expert testimony that could deprive a person of life or liberty should be more, not less, rigorous than testimony used to protect the interests of civil defendants.

In criminal cases, it is the nature and structure of the system that creates the adversarial bias that can infect opinions, and outcomes. Roger Koppl (Koppl, *How to Improve Forensic Science*, 20 Eur. J.L. & Econ. 255, 258 (2005) identified some of the structural causes:

- Each jurisdiction typically has just one forensic laboratory; the absence of competition reduces the incentive to perform well.
- Forensic labs are usually attached to police departments and therefore depend on the police department for their budgets, which naturally leads to a desire to please the police, even at the cost of honesty and thoroughness.
- Quality control is weak at most forensic labs.
- Forensic scientists often know what result they are "supposed" to reach, which can lead to an unconscious bias in interpretations of test results, or even conscious fraud. The scientist who performs a particular test typically also interprets the results of the test, reducing the odds that anomalies will be discovered

How can attorneys - and judges - try to deal with serious questions of cognitive bias on the part of forensic experts? *Daubert*, on remand to the 9th Circuit, implicitly recognized that cognitive bias on the part of experts in civil cases is a factor affecting admissibility. Should the relationship of a forensic expert witness to the government or defense in a criminal case be a basis for the gatekeeper to exclude proffered testimony? The Rules of Evidence seem to indicate so:

- Rule 403. Excluding Relevant Evidence for Prejudice, Confusion, Waste of Time, or Other Reasons

The court may exclude relevant evidence if its probative value is substantially outweighed by a danger of one or more of the following: ***unfair prejudice****, confusing the issues, misleading the jury, undue delay, wasting time, or needlessly presenting cumulative evidence.*

- Rule 403 - Notes of Advisory Committee on Proposed Rules

. . . "Unfair prejudice" within its context means an undue tendency to suggest decision on an improper basis, commonly, though not necessarily, an emotional one.

In reaching a decision whether to exclude on grounds of unfair prejudice, consideration should be given to the probable effectiveness or lack of effectiveness of a limiting instruction Bias is the relationship between a party and a witness which might lead the witness to slant, unconsciously or otherwise, his testimony in favor of or against a party. Bias may be induced by a witness' like, dislike, or fear of a party, or by the witness' self-interest. United States v. Abel, 469 U.S. 45 (1984) (emphasis added).The admissibility of evidence regarding a witness's bias . .

. is not specifically addressed by the Rules, and thus admissibility is limited only by the relevance standard of Rule 402. U.S. v. Lindemann, 85 F3d 1232 (1996)

Proving such bias to the satisfaction of a biased judge, however, will be difficult and take preparation. Discovery, although rarely used extensively in criminal cases, may be an approach to that preparation. Shortly after *Daubert*, Federal Rule of Civil Procedure 26(a)(2) was amended to provide that parties planning to use experts must prepare a report containing:

> *a complete statement of all opinions to be expressed and the basis and reasons therefor; the data or other information considered by the witness in forming the opinions; any exhibits to be used as a summary of or support for the opinions; the qualifications of the witness, including a list of all publications authored by the witness within the preceding ten years; the compensation to be paid for the study and testimony; and a listing of any other cases in which the witness has testified as an expert at trial or by deposition within the preceding four years.*

Perhaps criminal defense attorneys can borrow a page from the civil defense attorney playbook. During pretrial discovery regarding a plaintiff's expert, in a medical malpractice case for example, the defense will submit interrogatories and requests for documents seeking to identify a pro-plaintiff bias. Adapting that approach to a criminal case, it could include:

- a complete copy of all documents and materials furnished to the expert by the police or prosecutor
- a complete copy and description of all reports, notes or comments made by the expert about the case
- the time, place and description of all conversations with police or prosecutors related to the case
- a description of any review of the expert's work in the case, including the identity, position and qualifications of any reviewer
- identifying information of all cases in which the witness has been consulted by the particular police agency or prosecutor office
- a list of all cases in which the expert has previously testified
- whether personnel decisions, such as performance evaluations and compensation, are made by persons associated with a police or other government agency.

Whether such discovery will lead to evidence that will convince the judge to exclude prosecution testimony or not, it may still provide information that can be used for cross examination if the witness does testify.

Another alternative for the defense may be to call its own expert witness as to the cognitive bias of the prosecution expert. Courts now generally allow psychologists to testify as experts on the recognized factors affecting the credibility of eye witnesses. Is testimony from a qualified psychologist about the effect of cognitive bias on forensic experts similarly admissible? Similar to expert testimony about the unreliability of eyewitness testimony, should psychologist testimony about the various forms of unconscious bias be offered at trial to discredit a forensic expert? " . . . because bias is not a collateral issue, it was permissible for evidence on this issue to be extrinsic in form". U.S. v. Lindemann, 85 F3d 1232 (1996)

The defense can use all of this evidence to request appropriate jury instructions. Standard jury instructions include criteria for determining the credibility of witnesses in general and include cautionary instructions about the testimony of expert witnesses. Some sample jury instructions that could be tailored include:

- 8th Circuit 3.04 Credibility of Witnesses - *In deciding what testimony to believe, consider . . . any motives that witness may have for testifying a certain way"*
- 1st Circuit 3.06 Credibility of Witnesses - . . . *take into consideration . . . any bias they may have displayed; any interest you may discern that they may have in the outcome of the case*
- 5th Circuit 1.09 Credibility of Witnesses - *"Did the witness have a personal interest in the outcome of the case? Did the witness have any relationship with either the government or the defense?"*

Is a *Daubert* challenge based on bias viable? It certainly has not been an effective means of preventing biased experts from testifying in criminal cases. First, the reality is that criminal defense attorneys rarely even ask for a *Daubert* hearing. Second, judges have not regularly used *Daubert* to examine the admissibility of expert testimony for the prosecution in criminal cases and routinely allow it to come into evidence Risinger (2000) found that Daubert challenges to government evidence were successful less than 10% of the time in federal trial courts and less than 25% percent of the time in state trial courts. Groscup, et al, (2002) studied trial and appellate decisions regarding the admissibility of expert testimony. At the trial court level,

prosecution experts were admitted 95.8% . . . of the time, and defendant experts were admitted only 7.8% . . . of the total number of times they were offered. Another researcher concluded:

> *"Most judges, especially those with prosecutorial experience, presume that most defendants are, in fact, guilty, even though some are, in fact, innocent. This presumption of guilt, pro-prosecution perspective not only affects the manner in which many judges rule on motions, evaluate witnesses, and exercise their discretion, but it also adversely affects the willingness of many judges to police law enforcement agents and prosecutors."*

The reality is that judges often do not weigh the scientific validity of the proffered evidence in any meaningful way. Most of the decisions simply rationalize admissibility based on the prior admission of such evidence by other judges. In other words, the typical analysis becomes one of stare decisis, rather than the scientific inquiry required by *Daubert*. Some judges even take "judicial notice" of the reliability of certain kinds of prosecution evidence, like fingerprints, to avoid even having to hold a *Daubert* hearing. This is in spite of the admonition in FRE 201(b) that *"Judicial notice is appropriate only for matters that are capable of 'accurate and ready determination by resort to sources whose accuracy cannot be ... questioned' "*
There have been a few lower courts that have recognized that precedent in courts does not equate to scientific reliability or even to general acceptance in the scientific community. One court called it "grandfathering in irrationality" And the Court in US v. Saelee said:
"the fact that this type of evidence has been generally accepted in the past by courts does not mean that it should be generally accepted now, after Daubert and Kumho."

But why do so many judges refuse to look at scientific evidence and simply rely on precedent? In essence it is the same type of bias in judges that exists in the "experts" themselves. As to reliance on precedent, some social psychologists refer to a concept of "social proof" - the phenomenon of looking at what other people think is correct to determine what *is* correct". Cognition, in this context, is knowledge – what we believe to be true –what we think we "know" – what we value. Dissonance theory states that when relevant cognitions are inconsistent with one another, they can create dissonance, an uncomfortable state, which then brings about psychological processes to reduce the dissonance, or discomfort.

This process is what leads to the suggestion that there is a systemic pro-prosecution bias on the part of judges and that such a bias is reflected in admissibility decisions, regardless of the

standard of admissibility. As one scholar puts it, "as a general proposition, judges disfavor civil plaintiffs and criminal defendants and, are more likely to rule against them than against their opposites even when presenting equivalent evidence or arguments." Systemic pro-prosecution bias is a function of the same fairly obvious psychological concepts of cognitive bias. including "confirmation bias" and "outcome bias". These are the ways in which judges try to reduce our cognitive dissonance – to do almost anything to reconcile prior beliefs with the new "truths" they are being urged to adopt.

Does this really apply to judges? Dean Chris Guthrie described "confirmaton bias " he found in studying judicial decisions: *"judges come to the bench with political views . . . [that] can predispose them to rule in ways that are consistent with those opinions or attitudes. . . . The evidence [from empirical studies] suggests that attitudinal blinders are an issue not only at the highest court in the land but also in these lower courts"* Dean Guthrie, relying on significant empirical studies of judicial attitudes and actions, described judicial bias as a reflection of an "attitudinal blinder." These "attitudinal blinders" are especially prevalent in criminal cases and especially in the state courts where most criminal cases are tried. Most state court judges, as Professor Rodney Uphoff put it, ". . . are answerable to a tough-on crime electorate and are often reluctant, therefore, to make risky political decisions upholding the constitutional rights of criminal defendants."

It is again related to our reliance on precedent to validate our current choice – consistency! As Linda Morkan said*:*

> *"The human desire for consistency is a powerful tool of influence. Once we have committed to a position, we have an almost overwhelming urge to portray that action as the 'right' choice. People will go to great lengths to keep their thoughts consistent with what they have already decided."*

DAN SIMON

Dan Simon specializes in the field of Law & Psychology. He is the Richard L. and Maria B. Crutcher Professor of Law and Psychology at the USC Gould School of Law and holds a joint appointment at USC's Department of Psychology. He teaches criminal law and courses on the intersection of law and psychology and has been a visiting professor at Yale Law School and Harvard Law School. He earned an S.J.D. degree from Harvard Law School, an MBA from INSEAD in France, and an LL.B. from Tel Aviv University. He worked as an attorney for the Association for Civil Rights in Israel as human rights lawyer on the West Bank, and before joining the USC Gould School of Law in 1999, he was a member of the faculty of the University of Haifa Law School.

Professor Simon was recently selected by the National Institute of Science and Technology (NIST) to be a member of its Human Factors Committee, which will provide guidance throughout the Organization of Scientific Advisory Committees (OSAC) on the influence of systems design on human performance and on ways to minimize cognitive bias and mitigate errors in complex tasks. He also serves as an ad hoc referee for academic presses, peer-reviewed journals in experimental psychology, and the National Science Foundation.

Professor Simon is the author of *In Doubt: The Psychology of the Criminal Justice Process* (Harvard University Press, 2012). His publications in law reviews include "The Limited Diagnosticity of Criminal Trials" (*Vanderbilt Law Review*, 2011); "A Third View of the Black Box: Cognitive Coherence in Legal Decision Making" (*The University of Chicago Law Review*, 2004), and "A Psychological Model of Judicial Decision Making" (*Rutgers Law Journal*, 1988). He has also published a number of articles in experimental psychological journals, including "The Construction of Preferences by Constraint Satisfaction" (*Psychological Science*, 2004; with co-authors), "The Redux of cognitive consistency theories: Evidence judgments by constraint satisfaction" (*Journal of Personality and Social Psychology*, 2004; with co-authors), and Bidirectional Reasoning in Decision Making by Constraint Satisfaction (*Journal of Experimental Psychology—General*, 1999, with Keith J. Holyoak).

IN DOUBT

The Psychology of the Criminal Justice Process

DAN SIMON

Harvard University Press
Cambridge, Massachusetts
London, England
2012

Library of Congress Cataloging-in-Publication Data

Simon, Dan, 1955–
In doubt : the psychology of the criminal justice process / Dan Simon.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-674-04615-3 (alk. paper)
1. Criminal justice, Administration of—Psychological aspects.
2. Criminal investigations—Psychological aspects.
3. Judicial process—Psychological aspects. I. Title.
HV7419.S57 2012
364.01'9—dc23 2011038133

CONTENTS

1

INTRODUCTION

Criminal punishment is the most palpable and ubiquitous means by which the state maintains social order. However, before it unleashes its punitive powers, the state must determine with high certitude which human behaviors amounted to criminal events, and who perpetrated them. This feat requires compliance with an intricate legal regime that constitutes the criminal justice process. The workings of this process and the accuracy of the verdicts it produces are the subject of this book.

The following three cases offer a glimpse into the operation of the criminal justice process. Peter Rose, a California man, was charged with the rape of a thirteen-year-old girl. On the stand, the victim stated that she was 100 percent certain that Rose was her assailant, and a bystander witness stated that the perpetrator was either Rose "or his twin brother."[1] Bruce Godschalk of Pennsylvania was charged with two counts of burglary and forcible rape. The case against Godschalk was replete with incriminating evidence: one of the victims identified him; a jailhouse informant testified that he made inculpatory statements; and a forensic expert provided a blood-typing match. Critically, the prosecution presented a thirty-three-minute tape recording in which Godschalk confessed to the crimes, providing specific details that could not have been known to the public.[2] In his confession, Godschalk blamed his crime on his drinking problem, and added, "I'm very sorry for what I've done to these two nice women."[3] Kirk Bloodsworth was charged with the capital offense of raping and murdering a nine-year-old Maryland girl. At trial, Bloodsworth was identified by five eyewitnesses. The prosecution also provided testimony of statements he made about the rock that was used as the murder weapon, and a forensic investigator testified that the murderer's shoe print matched Bloodsworth's shoes.[4]

In all three cases, the evidence of guilt was indeed compelling, and the men were found guilty beyond a reasonable doubt. Rose was sentenced to twenty-seven years in prison, Godschalk was sentenced to 10–20 years, and Bloodsworth was given a death sentence. For years, nothing seemed out of the ordinary with these convictions, until DNA testing showed that none of these men had actually perpetrated the crime for which he was being punished. The witnesses who testified in these cases were mostly wrong, especially on the crucial aspect concerning the identity of the men who committed the crimes. By the time Rose was released, he had served eight years in prison, Godschalk had served fourteen and a half, and Bloodsworth had served eight years, two of which were on death row.

These cases raise a series of difficult questions pertaining to the functioning of both the investigative and adjudicative phases of the criminal justice processes: What caused the witnesses to provide mistaken testimony? Why did the police investigators, prosecutors, and jurors believe the witnesses? Could the mistakes have been caught? Most importantly, what can be done to prevent such occurrences in the future?

The View from Experimental Psychology

One of the obvious features of the criminal justice process is that it is operationalized mostly through people: witnesses, detectives, suspects, lawyers, judges, and jurors. The wheels of the system are turned by the mental operations of these actors: memories, recognitions, assessments, inferences, social influence, and decisions, all tied in with moral judgments, emotions, and motivations. Criminal verdicts can be no better than the combined result of the mental operations of the people involved in the process. It thus seems sensible to examine the workings of the criminal justice process from a psychological perspective. Fortunately, a large body of experimental psychological research is at our disposal. For some decades now, legal psychologists have been earnestly studying the conditions under which people tend to succeed or fail in fulfilling their designated roles in the operation of the criminal justice process. Likewise, research in a range of related fields—notably cognitive psychology, social psychology, and decision making—has accumulated a wealth of knowledge about the mental processing that is inevitably implicated in the workings of the process.

The principal endeavor undertaken in this book is to apply a part of this vast and dispersed body of experimental psychology toward a better

understanding of the operation of the criminal justice process. The overall observation that emanates from this research is that human performance on the tasks involved in the process can be exceedingly complicated and nuanced. Tasks that are generally taken for granted—such as identifying a stranger, remembering a specific detail from an event, and ascertaining the accuracy of such testimonies—are not as straightforward as they seem. The accuracy of these tasks is contingent on multitudes of factors, many of which are unknown, underappreciated, and easily overwhelmed by the harsh reality of crime investigations and the contentious legal process that ensues.

This observation leads to the twofold claim that lies at the heart of this book: first, in nontrivial criminal cases, the evidence produced at the investigative phase—in particular, human testimony—comprises an unknown mix of accurate and erroneous testimony, and is thus not always indicative of the defendant's guilt. The following four chapters are devoted to providing insight into the prospect of error in criminal investigations. Chapter 2 explores the work of police investigators, focusing on the conditions that can facilitate and even stoke mistaken investigative conclusions. Chapter 3 deals with the topic of identification of perpetrators by eyewitnesses. Chapter 4 examines witnesses' memory of the criminal event. Chapter 5 deals with the interrogation of suspects.

The second key claim is that the ensuing adjudicatory phase is not well suited to ascertain the accuracy of the evidence, and thus cannot distinguish reliably between guilty and innocent defendants. The limited diagnostic capabilities of the adjudicatory process are the subject of the two subsequent chapters. Chapter 6 explores problems that fact finders encounter in determining the truth from the evidence presented at trial. Chapter 7 examines the efficacy of the legal mechanisms that are designed to support the fact finders in performing that task.

In sum, the research will indicate that criminal investigations are prone to produce evidence that contains substantial errors, which the adjudicatory process is generally incapable of correcting. The compounded problems with the accuracy of the investigative phase and the diagnosticity of the adjudicatory phase lead to the conclusion that the criminal justice process falls short of meeting the level of certitude that befits its solemn nature.[5] This shortfall is generally overlooked or denied by the people entrusted with designing and governing the system—notably, police personnel, prosecutors, judges, and law makers—and it is not adequately recognized in the scholarly and public debates. Chapter 8 examines the

implications of this state of affairs and explores some systemic ways to promote the accuracy of the process.

Process Breakdowns

Criminal cases can break down in two ways. A person who perpetrated a crime might escape punishment, or an innocent person might be convicted and punished for a crime he did not commit.[6] The failure to convict guilty people—which can be loosely labeled *false acquittals* (even though most such cases do not make it to a formal acquittal at trial)—is a grave problem for an ordered society. Fewer than one-half of felony crimes are ever reported to the police,[7] and only one of every five reported felonies is cleared by an arrest.[8] Crimes are unlikely to be cleared, for example, when they are not witnessed, when the witnesses refuse to cooperate with the police, or when the witnesses cannot provide the necessary information to solve the case.[9] In these instances, the criminal justice process fails because it lacks the requisite evidence to attain a conviction. The psychological research is best suited to provide insight into cases in which evidence *is* present, particularly by identifying the conditions that make that evidence more or less likely to sustain an accurate conviction. Thus, this book will focus mostly, though not exclusively, on false convictions. It is important to note that some key recommendations proposed in this book are designed to enhance the accuracy of the evidence overall and thus stand also to reduce the incidence of false acquittals.

The steady flow of exonerations in recent years has turned a spotlight onto the accuracy of the criminal justice process.[10] Many of these exonerations have resulted from the work of the Innocence Project, cofounded by Barry Scheck and Peter Neufeld. Some critics of the system describe the recent revelations of false convictions as a momentous, even revolutionary, event.[11] In contrast, proponents of the system steadfastly trivialize their import and dismiss them as "an insignificant minimum."[12] According to a data set maintained by the Innocence Project, 281 convicted inmates have been exonerated on the basis of DNA testing as of the beginning of December 2011,[13] and many more have been exonerated by other types of evidence.[14] The true number of false convictions is unknown and frustratingly unknowable. Based on exoneration data in two categories of capital homicide, the rate of error is estimated at about 3–4 percent, with a possible upper boundary of 5 percent.[15] The rate of false convictions is most likely considerably higher. Given the difficult, even tortuous, legal hurdles that stand in the way of exposing false con-

victions, there is no doubt that a large number of falsely convicted persons have not been, and will never be, exonerated. While a detailed argument on the incidence of false convictions is beyond the scope of this book, it is worth noting that an innocent defendant stands a chance of exoneration if he was convicted for murder or rape;[16] did not accept a plea bargain;[17] was sentenced to a lengthy prison term;[18] and was able to secure good legal representation and investigation in the post-conviction phases. It is essential also that the case centered upon the identity of the perpetrator;[19] physical or otherwise strongly exculpating evidence was present;[20] and that the exculpating evidence was collected,[21] properly preserved,[22] and made available to the defendant.[23] A healthy dose of luck can be very helpful,[24] and in the absence of DNA evidence, it is all but essential.[25] Innocent people for whom any of these conditions do not obtain are unlikely to be exonerated, because the errors underlying their convictions will rarely be detected.

False convictions are the joint product of breakdowns in both the investigative and adjudicative phases of the criminal process. These breakdowns call for a closer look at the system's methods of sorting out criminal responsibility in each of the respective phases.

Investigation breakdowns. An investigative process that results in a false conviction involves a combination of failures. First, the investigation failed to discover the truth, as manifested by the simple fact that the true perpetrator got away. Second, the investigation failed to discern the faulty nature of the evidence it collected, as manifested by the fact that the investigators cleared the case and recommended it for prosecution. The least familiar, though most dire, failure is that in many instances the investigation itself contributed to the mistaken conclusion.

To better understand how mistaken testimony comes about, it would be useful to propose a distinction between two types of error. First, some errors are caused by random cognitive failures that are inherent to human cognition. This category of *spontaneous error* pertains to occasional failures in human performance that cannot be attributed to any obvious external cause. Errors are taken to be spontaneous, for example, when an honest eyewitness mistakenly confuses an innocent person with the perpetrator or when he misremembers a particular detail from the crime scene. Spontaneous errors do not have a directional tendency; they are as likely to inculpate an innocent defendant as they are to exculpate a guilty one. A number of innocent people were spontaneously misidentified by witnesses while walking down the street, shopping in a store, or riding in an elevator.[26]

However serious, these cases do not begin to capture the intricate relationship between the investigative process and the occurrence of error.

Errors can also be caused or exacerbated by situational factors. In the context of the criminal justice process, such situational factors follow from the investigative procedures or from interactions with criminal justice officials and lawyers. Such is the case when a witness picks an innocent person from a skewed lineup or reports an erroneous memory as a result of a suggestive question posed by a detective. These instances represent a second type of error, which can be labeled *induced error.*[27] Induced errors have a directional tendency to coincide with their inducing influences. As discussed in the following chapters, these influences tend more often to pull the case toward conclusions of guilt.

Although law enforcement officials tend to view false convictions as caused by spontaneous errors,[28] induced errors figure more prominently in the studied DNA exoneration cases. In the three abovementioned cases, for example, we see a transformation of the witnesses' statements toward conformity with the police's case against the suspect. The evidence presented in court was considerably different from—and indeed, more incriminating than—the witnesses' initial statements given to the police. Notwithstanding her certain identification of Peter Rose in the courtroom, the victim was initially adamant that she did not see the face of the man who dragged her into an alley, raped her from behind, and fled. When presented with a photograph lineup that contained Rose's photo, she could not pick anyone out. At the lineup, the bystander witness, who later testified that Rose was either the perpetrator or his twin brother, had actually selected a photo of an innocent filler.[29] At first, Bruce Godschalk denied his involvement in the crime, and he could not provide any details about it. By the end of the interrogation, however, he confessed to horrible deeds that he had not perpetrated, and provided intricate corroborating details that he could not possibly have known.[30] Four of the five witnesses who testified against Kirk Bloodsworth had provided the police with inconsistent and unreliable statements. One witness had previously tipped the police that the suspect matched a different person whom she knew, and a second witness had initially told investigators that she did not see the face of the perpetrator. At the lineup, one of the two child witnesses picked an innocent filler and the other failed to choose anyone.[31] Similar transformations of evidence were observed in the cases of Walter Snyder,[32] Edward Honaker,[33] Darryl Hunt,[34] William O'Dell Harris,[35] Ronald Cotton (discussed in Chapters 2 and 3), and numerous others.[36]

These cases illustrate that criminal investigations can overwhelm the often weak and vague remnants of the truth, and thus shape the testimony to substantiate a prosecution.

Another observation that has come to light from the studied exoneration cases is that the inculpating evidence presented at trial often hinged not just on a single mistaken evidence item. Rather, as seen in the cases of Rose, Godschalk, and Bloodsworth, prosecutions are typically based on an array of seemingly independent pieces of evidence, all of which connect the defendant to the crime.[37] An analysis of DNA exoneration cases shows that 71 percent of the cases involved mistaken identification, 63 percent involved forensic science errors, 27 percent involved false or misleading testimony by forensic scientists, 19 percent involved dishonest informants, 17 percent involved false testimony by lay witnesses, and 17 percent involved false confessions.[38] These evidentiary causes sum to 214 percent of the cases, which means that on average, each case was afflicted by more than two types of bad evidence.[39] In reality, the number of mistaken evidence items is much greater. For example, many misidentification cases contain erroneous testimony from multiple witnesses, and each mistaken identification typically includes numerous additional incorrect corroborating statements.

Given that these convicted persons were ultimately found to have not perpetrated the crimes, it follows that the bulk of the evidence used to convict them, if not all of it, was wrong. While it is theoretically possible that all the errors just happened to coincide, there is strong reason to suspect that they were induced by the investigative process. As discussed in Chapter 2, due to the dynamic nature of police investigations, errors can beget more errors. By way of illustration, a mistaken fact suggested by one witness can lead the detective toward a mistaken conclusion, which can then induce the forensic examiner to confirm the hypothesis, and so on. This *escalation of error* can transform even a flimsy mistake into a full-blown case replete with overwhelming evidence strong enough to support the conviction of an actually innocent person.

Adjudication breakdowns. Mistaken verdicts also entail a breakdown of the adjudicative phase. The failure to distinguish between guilty and innocent defendants typically follows from the failure to tell apart accurate and erroneous testimony. Virtually every exoneration follows a conviction by a jury or judge who believed that the faulty evidence was true beyond a reasonable doubt. Some prosecutions of innocent defendants

failed to raise even the slightest suspicion from the fact finder. One jury, for example, took no more than seven minutes to convict an innocent man for a crime that resulted in a life sentence.[40] By the same token, the limited diagnosticity of the adjudicative process can also lead to false acquittals. Indeed, some juries have refused to convict defendants in the face of compelling evidence of guilt.[41] Hence jury verdicts are often perceived to be unpredictable, even by professionals who sit through the whole trial and see all the evidence.[42]

Case Typologies

Easy and difficult cases. Not all criminal events are born equal, nor are the ensuing investigations. Police studies show that the majority of serious criminal events that get cleared are solved quite easily. In fact, most of them are solved at the first encounter with the responding patrol officer, that is, without any investigatory effort by detective units.[43] For example, crimes are solved easily when a witness identifies the perpetrator by his name, address, vehicle, or place of employment. Solving cases is also relatively straightforward when the perpetrator is caught in the act, in possession of the contraband, or singled out by means of forensic tests, surveillance cameras, or telecommunication records. This category of *easy cases* accounts for a large majority of the people sitting in prisons, but it tends to account for only a small fraction of the investigative and adjudicative resources expended. Habitually, these cases are disposed through plea bargaining, and when they do go to trial, they hardly challenge the adjudicative process. At the other extreme, there is a large category of crimes that are exceedingly difficult to solve because of a dearth of evidence, lack of resources, or noncooperation by victims and witnesses. Although some of these cases consume heavy investigative resources, most are readily abandoned. Either way, these cases tend not to be cleared, and thus do not evolve into prosecutions, not to mention convictions.

The criminal justice process is brought to bear mostly in the middle category of *difficult cases,* where solving the case is neither easy nor impossible. In these instances, the initial information made available to the responding officer falls short of enabling her to clear the crime or to single out the perpetrator. The investigative effort required to overcome this evidentiary shortfall is what makes these cases difficult. This relatively narrow category of cases consumes the bulk of the investigatory and adjudicatory resources, and it also puts the criminal justice process to the test. This category of cases is the focal point of this book.

Identity cases and culpability cases. At the most general level, criminal cases center upon two types of questions. Some cases are concerned primarily with figuring out who committed the crime, the *whodunit* question. These can be labeled *identity cases*. Accurate verdicts in this category mean that the true culprit was convicted, whereas false convictions typically mean that an innocent person was found guilty. *Culpability cases* center upon determining the criminality of a suspect whose identity is not in question. Accurate verdicts in these cases mean that the perpetrator was appropriately convicted for his criminal actions. False convictions in this category mean that the defendant's innocent behavior was mistakenly taken to be guilty, or that he was convicted on a charge that was more severe than warranted by his conduct.

This book is concerned primarily with the factual accuracy of verdicts, and thus focuses on case outcomes that can, in principle, be determined as being correct or incorrect. As such, the book pertains straightforwardly to almost all identity cases, which can be resolved by showing that the defendant was or was not the person who committed the crime. The vast majority of exonerations stem from identity cases, where subsequent evidence demonstrated that the inmate did not perpetrate the crime for which he was convicted. The book does not apply directly to questions of culpability that hinge on value judgments, such as the morality of a behavior, the reasonableness of an act, or the fairness of the law. It does, however, apply to culpability cases that revolve around determinations of factual questions such as the defendant's actions and mental states. It should be noted that culpability cases rarely result in exonerations. Culpability questions tend to hinge on subtle and elusive aspects of the criminal event, and thus are not readily subject to objective confirmation or refutation. It follows that mistaken determinations of the defendant's culpability are rarely traceable. The dearth of culpability cases among the exoneration cases should not be taken to suggest that these mistakes do not occur.

Some Caveats and Qualifications

It is imperative to keep this book's claims and objectives in perspective. While the book attempts to provide a relatively broad application of legal psychology to the criminal justice process, it necessarily leaves out some important aspects of the research. For one, it does not examine differences in performance among people, which are bound to influence verdicts under some conditions.[44] Rather, it seeks to capture broader

phenomena entailed in legal procedures and practices, and thus focuses on the overall performance of legal actors. Nor does the book deal with the performance of special populations, such as children, the elderly, and people affected by mental disease, retardation, drug dependence, and the like. By concentrating on healthy adults, the book examines the performance of the criminal justice process as it is operationalized by well-functioning actors.

The book does not offer an examination of the ubiquitous practice of plea bargaining, the process by which some 95 percent of felony convictions are obtained.[45] Plea bargaining is one of the most obscure and troubling aspects of the criminal justice system,[46] but it does not readily lend itself to psychological experimentation. Still, it warrants noting that the problems with the integrity of the evidence discussed in the following chapters are bound to affect plea negotiations no less—and, probably, even more—than they do criminal trials. Effectively, defendants' decisions to plead guilty are based on sparse, uncertain, and questionable evidence that will rarely be subjected to any meaningful scrutiny.

A substantial number of known mistaken verdicts have been caused at least in part by conscious and deliberate efforts to distort the truth. The culprits in these transgressions have been people with a stake in the outcome, such as codefendants, and overreaching or corrupt detectives, prosecutors, and forensic examiners.[47] Numerous convictions that resulted in DNA exonerations were driven by police misconduct,[48] prosecutorial misconduct,[49] and misleading or fraudulent forensic testimony.[50] Deliberate distortions are the most egregious type of miscarriage of justice, especially when perpetrated by state officials. This book, however, focuses primarily on the working of the process when all the actors seek to fulfill their roles honestly and dutifully.

The book should not be taken to stand for the proposition that the legal system is entirely insensitive to any psychological aspects involved in the production of criminal verdicts. Indeed, the criminal justice system embodies a considerable amount of psychological insight. For example, the law recognizes the possible effects of leading questions, coercion in the interrogation room, and prejudicial evidence.[51] Still, law's psychological sensibilities are mostly frozen at the state of the pre-experimental psychological knowledge that prevailed at the time these common-law rules were forged. Law's intuitions tend to overestimate the strengths of human cognition and to underappreciate its limitations. There is good reason to update the system with more reliable and nuanced knowledge of this complex matter.

The book's focus on psychological causes of mistaken verdicts should not obscure the fact that the criminal justice process is plagued by a host of other factors, which have not been the subject of substantial psychological experimentation and are thus not discussed here in any detail. The late William Stuntz repudiated the system for the excessive discretion awarded to prosecutors, inconsistent policing, the infrequency of jury trials, and the inordinate reliance on plea bargaining.[52] Other factors include insufficient access to appropriate legal representation and investigation,[53] inadequate training and lack of discipline of law enforcement personnel, improper forensic procedures, and the frequent reliance on unreliable evidence such as informants.[54]

Methodological Concerns

It must be acknowledged that the research that underlies this project is naturally susceptible to methodological concerns. No single study, body of research, or experimental method is devoid of methodological limitations. Most notably, applying psychological research to the legal world raises concerns over its *external validity,* that is, the degree to which the findings can be generalized beyond the experimental setting to the naturalistic environment.[55] Psychologists, who are habitually attuned to situational influences on human behavior,[56] are the first to acknowledge that experimental findings are sensitive to the specifics of the experimental design.[57] Critics of legal-psychological research observe that the controlled environment of the laboratory differs from the real world in important ways. The research has been criticized for overstating the import of experimental results, and more specifically, for the nonrepresentativeness of the participants, the disconnectedness from institutional contexts, and the inconsequentiality of the tasks.[58] This critique places a serious burden on researchers' shoulders. It does not, however, warrant a wholesale dismissal of the research.[59]

The concerns over the external validity of this body of research are largely allayed by its *convergent validity*.[60] The convergent validity of this research refers to the combined empirical support derived from replications of the results from studies that test different stimuli, on different populations, in different laboratories, and focusing on different facets of the issues. The convergent validity is enhanced also by triangulating a variety of methodologies, namely, basic- and legal-psychological experimentation, survey data, field studies, and archival research.[61] To be sure, not every finding mentioned in this book has been subjected to the

complete panoply of external-validity verification, though the available data invariably indicate consistency and convergence in the findings.

One valid criticism of the experimental method is that it cannot fully capture the richness of human performance, which is invariably multi-determined. By design, psychological experiments focus on only one or two aspects of the task, while keeping all the other dimensions under tight control. The research, then, is not capable of explaining how any of its observations would fare if the focal aspect were allowed to interact with each of the numerous other aspects that were controlled in the particular study. This limitation must be acknowledged, but it does not lead to the conclusion that this body of experimentation necessarily exaggerates the problems with the criminal justice process. In fact, it seems more likely to underrepresent them because many of the hidden interactions actually detract from the accuracy of the process.[62] The experimental environment tends to block out biasing factors such as the actors' motivations, incentives, subcultures, and personalities, as well as adverse social dynamics, emotional arousal, prejudice, and the like.[63] Moreover, many of the human processes involved in operating the criminal justice process are basic-psychological phenomena. People's performance on these tasks is barely amenable to improvement, but is very susceptible to contamination from poor procedures.

Still, there are reasons to guard against overstating the conclusions that can be drawn from the research results. First, the prevalent criterion for the validity of experimental findings is the statistical probability that they are attributable to the experimental treatment, as opposed to mere chance. This criterion does not speak to the strength of the treatment or to its absolute values.[64] Second, difficult cases contain an unspecifiable fraction of the array of factors that have the potential to skew the process. These factors vary in strength, and they do not all sway the process in the same direction. With the exception of extreme cases, the net effect of the biasing factors is unknowable. Thus, the experimental findings are best understood as heightened propensities, or tendencies. It would be imprudent to attempt to determine unequivocally the exact effect of any factor or whether a particular outcome was accurate or mistaken. The research can, however, enrich our understanding about which factors present a risk of error and how best to avoid them.

Toward Reform: Accurate and Transparent Evidence

The primary objectives of this book are to energize the debate about the accuracy of the criminal process and to suggest reforms that would enable it to better meet its exacting goals. Given the depth of the foregoing critique, one might well be tempted to advocate a thoroughgoing restructuring of the criminal justice system. A fundamental institutional redesign, however, is not a proximate objective of this book. Deep institutional reforms would relinquish much of the reformative potential of the psychological research. Unlike most other disciplines that are employed in the analysis of the legal system, experimental psychology operates at a granular level that enables offering direct and immediate solutions to specific problems. It would be a mistake to forego the benefits that these solutions can yield. Over the years, a number of scholars have proposed profound institutional changes to the criminal justice process. Most of these proposals have entailed adopting elements from the inquisitorial system practiced in continental European countries.[65] These proposals warrant serious consideration, but they run against the grain of the current Anglo-American legal culture,[66] and would likely require deep legislative changes and perhaps also constitutional amendments.[67] Hence, these proposals seem unlikely to be implemented in the foreseeable future. In the vein of pragmatism, the recommendations offered in this book will be limited to reforms that are practical, feasible, and readily implementable in the short or medium term. Most of these reforms are targeted directly at law enforcement officials, lawyers, and judges, and they could be adopted at the departmental level and even by the individuals themselves.[68]

The reasons for reducing the incidence of false acquittals hardly need mentioning: escaping a deserved criminal sanction negates the very purpose of the criminal justice system, and thus can undermine the foundation of an ordered society. There are also strong reasons for reducing the incidence of false convictions to the lowest feasible level. Most obviously, inflicting punishment on innocent people constitutes a grave moral transgression, and it can also devastate that person's family and dependents. Preventing false convictions also serves a public safety interest, in that every conviction of an innocent person effectively averts the pursuit and incapacitation of the true perpetrator. By the same token, uncovering false convictions can lead to the apprehension of the actual perpetrators. In almost one-half of the DNA exonerations, the evidence that cleared the innocent suspect also inculpated the true perpetrator.[69] In the

long run, minimizing false verdicts is bound to enhance the legitimacy of the criminal justice system.

Reforming the criminal justice system is a delicate and complex endeavor, particularly given the pointed adversarialism that tends to pervade all things related to criminal justice. Reforms must be designed not to reduce the overall rate of convictions or acquittals, but should be targeted as narrowly as possible at *false* convictions and acquittals.[70] Accurate evidence and correct verdicts, rather than partisan advantage, should be the goal. The two central recommendations made throughout the book are designed to address the most serious problem that affects the accuracy of criminal verdicts, namely, the problematic quality of the evidence presented in many criminal trials.

First, criminal investigations ought to be conducted meticulously according to best-practice procedures. Best-practice investigative procedures will ensure that criminal verdicts and plea bargains will be based on the most accurate account of the criminal event that can be obtained. Specific recommendations for some best-practice procedures will be offered at the end of each chapter.

In determining which procedures ought to be considered "best practice," one ought to think through the implications of the proposed reform for both false convictions and false acquittals. Contrary to widely held beliefs, criminal justice reform is not always a zero-sum game in which reducing one type of error necessarily increases the opposite one. Indeed, some of the key recommendations proposed in this book are designed to improve the quality of the evidence across the board and thus reduce both types of error at once. These win-win reforms include the use of computerized systems in eyewitness identification procedures, resorting to sophisticated interviewing protocols such as the cognitive interview, and, as discussed below, the creation of a complete record of the investigative process. A proposed reform should be noncontroversial also when it reduces one type of error substantially while causing a marginal increase, or none at all, in the opposite error. Reforms should be deemed justified also when they entail a moderate increase in the opposite error, when the marginal cases are based on evidence that is nominally correct, but unreliable.[71] It is, however, inescapable that some policy decisions entail tradeoffs between the two types of error, where the respective evidence is of similar reliability. The calculation of such tradeoffs is a complex undertaking because of the unknown distribution of truly guilty and innocent defendants in the mix of the cases and the

perplexing weighting of the social costs of the respective errors. A full discussion of all possible costs and benefits of each of the recommended reforms lies beyond the scope of this book. While even the more controversial of these recommendations seem to strike a correct balance, they could benefit from further analysis and debate.[72]

Second, all encounters with witnesses should be recorded in their entirety and the recordings should be made openly available to all parties. In other words, the goal is to make the evidence as transparent as possible. It is important to appreciate that courtroom testimony is usually proffered months, sometimes years, following the criminal event.[73] During this time, witnesses typically have numerous encounters with the legal process. They interact with investigators, cowitnesses, lawyers, and other people who have a stake in the outcome of the case, and they are subjected to procedures that have the potential to induce error. Over the natural course of the process, testimony often changes, as previously unreported details come to be included in witnesses' statements, narratives are crystallized, gaps get filled, ambiguity fades away, and tentativeness is replaced by certitude. In other words, the *synthesized* testimony that is presented at trial often differs from—and is invariably stronger than—the witnesses' *raw* statements they initially gave the police. Although raw testimony is usually the best approximation of the truth, verdicts are invariably based on the inferior synthesized version.[74]

Enhancing the transparency of the evidence should have a very favorable impact on the process. Creating a reliable record of the criminal investigation stands to improve the investigation itself. The record will provide law enforcement agencies with a tool for training, oversight, and quality assurance. This should promote adherence to best practices and deter misconduct. The record could also serve as an informational tool by capturing forensic details that would otherwise be missed. Importantly, the record will provide access to the witnesses' raw statements and thereby offer a way around the effects of memory decay, contamination, and any biases or distortions arising from the investigative and pretrial processes. The availability of a record should also have a direct effect on the witnesses themselves because their testimony could be checked against their statements to the police. Effectively, courtroom testimony will be given under the shadow of the witnesses' own raw statements. The availability of the record should also reduce any pressures applied on the witnesses to alter their testimony, and when necessary it could be used to supplement or replace the testimony given in court. Transparent procedures will

enable fact finders to focus on drawing correct inferences from the evidence, rather than conjecturing about its reliability. Greater transparency should also help jurors determine whether the testimony might have been induced or otherwise biased by the investigation itself.

The combined effect of heightened accuracy and transparency has tremendous potential to improve the performance and enhance the integrity of the process. More accurate and transparent evidence is bound to improve the ability of all decision makers—investigators, prosecutors, defense attorneys, judges, defendants, and jurors—to make more informed and well-reasoned decisions. Most notably, criminal verdicts are bound to be more accurate, and plea bargains are expected to be fairer and better calibrated with the defendant's actual guilt. Greater accuracy and transparency are bound to increase the legal actors' trust in the evidence and limit their ability to distort and hide it, which should lead to a reduction in the distrust between the adversarial parties and a softening of the contentiousness of the process. The range of plausible claims will be curbed, narrowing the opportunities for both unjust prosecutions and frivolous defenses. Greater accuracy and transparency should reduce the need to sort out murky facts through the costly, cumbersome, and imprecise process of litigation. One can also expect that the heightened level of factual clarity will result in fewer appeals, *habeas* proceedings, civil suits, and damage payouts.

However promising, the proposed recommendations should be constantly subjected to reassessment. Future research may yield somewhat different findings and contribute new insights to the policy debate. While the available psychological literature is neither perfect nor fixed in stone, it offers a wealth of sorely needed insight into the workings of the criminal justice system and it can show the way toward important reforms.

2

"WE'RE CLOSING IN ON HIM"

Investigation Dynamics

The criminal process is as good as the evidence on which it feeds. In all but the simplest of cases, the fact finder at trial is bound to be presented with a mixed fare, containing unknown shares of accurate and inaccurate testimonies. A central claim of the next four chapters is that the single most important determinant of evidence accuracy is the police investigation. This chapter examines the dynamic process by which evidence is sought and evaluated. It highlights the risk that investigations will arrive at faulty conclusions, even absent any malicious intent. The following three chapters examine the accuracy of the types of evidence that are commonly used in criminal prosecutions. These examinations emphasize both the risk of spontaneous error by the witness and the proneness of the investigation to induce and shape their testimony. Understanding the workings of police investigations is thus key to an appreciation of the verdicts they propagate.

The case of Ronald Cotton provides a rare opportunity to peer into the investigative process and to appreciate how closely the psychological research maps onto real-life investigations. Early in the morning of July 29, 1984, Jennifer Thompson, a white twenty-two-year-old student in Burlington, North Carolina, awoke to find a stranger hovering beside her bed. The man put a knife to her throat, forced himself on her, and sexually assaulted her. Throughout the ordeal, Thompson made an effort to memorize any feature that could help her identify her assailant. At some point, Thompson managed to convince him to allow her to go to the kitchen to fix them drinks. She seized the opportunity to escape through the back door, and ran for shelter in a nearby house.

A tip communicated to the police implicated Ronald Cotton, who was out on parole for a conviction for breaking and entering. Cotton, who

was African-American, had been convicted as a juvenile for attempting to rape a fourteen-year-old white girl. Cotton was tried twice and convicted. At his second trial he was convicted for sexual crimes against both Thompson and a second woman who was assaulted in a nearby apartment that same night. He was sentenced to life plus fifty-four years in prison, and his conviction and sentence were ultimately held up on appeal.[1]

The evidence produced at trial amounted to a compelling incrimination of Cotton. Thompson provided forceful testimony, which included a confident identification of Cotton as her assailant. Police investigators and the prosecutor stated that she was the best witness they had ever put on the stand. At the second trial, Cotton was also identified by the second victim. Both victims provided similar descriptions of the man, and both reported that he was wearing a distinctive navy blue sports shirt with white stripes circling the arms. A bystander witness testified that around the time of the crime she saw Cotton riding a bicycle near Thompson's apartment and wearing the blue shirt described by the victims. The restaurant owner for whom Cotton had worked testified that Cotton had worn a similar blue shirt to work, and that he had also been seen wearing white gloves similar to the distinctive gloves described by Thompson. The employer testified also that Cotton had a habit of fondling white female waitresses and talking to them about sex. The prosecution's case was bolstered by physical evidence collected from Cotton's residence: a flashlight that was said to be similar to a flashlight removed by the assailant from the second victim's apartment, and sneakers that appeared to have been the source of a piece of foam found in Thompson's apartment. Cotton's defense was based on testimony by his family members stating that he was at home that night, watching TV and sleeping on the living-room couch. That alibi was undermined by the fact that Cotton had previously given the police an alibi that turned out to be untrue.

Some ten years after his arrest, a DNA test proved that Cotton was not the man who assaulted Jennifer Thompson. He was exonerated and released from prison after serving more than ten years of his life sentence. The biological evidence was traced to a convicted rapist, Bobby Poole, whose name as a suspect arose after Cotton was already in custody. It turns out, then, that the bulk of the evidence implicating Cotton—if not every material bit of it—was flawed. Thompson's identification of Cotton was wrong, as was the identification by the second victim. It is most likely that Cotton's employer never saw him wearing that distinctive shirt or uncommon gloves. The flashlight picked up from Cotton's residence was

not the one taken from the second victim's house, and his sneakers were not the source of the foam found in Thompson's apartment.

How did all of this inaccurate evidence come into play? A static snapshot of the prosecution's case cannot provide an answer to this question, as evidence does not normally provide an account of its own production. What is needed is a dynamic account of how the case evolved from the initial report to the police through the adjudicative process. This case is exceptionally informative thanks to the uncommonly frank and detailed reports provided by Thompson and police detective Mike Gauldin. There is every reason to believe that Thompson's testimony was sincere and that Gauldin's investigation was performed conscientiously.[2]

The key to this prosecution lies in the fact that from the inception of the investigation, Thompson had apparently only a faint memory of the face of her assailant, Bobby Poole. As discussed in Chapter 3, the frailty of her memory was likely due to the dim lighting, the stressful assault, and the fact that her assailant was a member of a different race. That memory was soon cluttered and possibly also morphed by the arduous task of constructing the facial composite sketch. The investigation gained considerable momentum once Cotton's photograph was placed in the photo array, and it escalated considerably once Thompson picked it out. The identification bore the markings of a weak recognition: it was tentative, doubtful, and protracted. Any qualms that she might have harbored were probably allayed by Gauldin's assurance that she chose the man whom they had suspected. By the same token, Thompson's identification emboldened Gauldin, placing Cotton in the crosshairs of the investigation. Gauldin's search of Cotton's residence yielded the physical evidence that further implicated him in the act. After arresting and questioning Cotton, he also discovered that Cotton's alibi was untrue. Thompson, in return, was emboldened by Gauldin's findings.

The case against Cotton was boosted further when he was picked out by Thompson at a live lineup (he was the only person included in both procedures). Again, Thompson's identification was hesitant, slow, and insecure. Once again she was reassured and relieved by Gauldin's confirmation that she had chosen the same man. With the investigation closing in on Cotton, information that indicated the possible involvement of a convicted rapist, Bobby Poole, was disregarded. The case became stronger after the bystander stated that she saw Cotton near the crime scene wearing the particular blue shirt, and Cotton's employer linked him to the shirt and to the gloves described by Thompson. At some

point, the second victim claimed that she recognized Cotton as her assailant despite having picked someone else at the lineup.

This dynamic account indicates that the mass of evidence against Cotton was triggered by the tip that connected Cotton to the composite sketch and was solidified and energized by Thompson's initial recognition of Cotton in the photographic array. That flimsy and erroneous identification propelled a process that ultimately produced a confident identification by Thompson herself, a reversal of a key witness's testimony (the second victim), another misidentification (the bystander witness), statements about Cotton's clothing that were probably false (from Cotton's employer), and two items of misleading physical evidence. Importantly, the investigation had a parallel effect on Thompson herself, transforming her initial hesitance into a formidably confident and persuasive testimony. By the end of the process, that initial error had escalated into a powerful prosecution that easily convinced two juries and passed the muster of an appellate court.

The case of Ronald Cotton epitomizes the phenomenon of the *escalation of error.* At bottom, even the most compelling prosecutions can be the product of a flimsy or erroneous piece of information that became amplified and reinforced as a result of the dynamics of the investigation. Similar escalations are observed in the investigations that resulted in a large number of DNA exonerations.[3] To understand better how investigations can go awry, we turn to criminological and psychological research that illuminates the investigative process.

The Investigative Task

Any discussion of criminal investigations in the United States must be qualified by the fact that investigations vary widely among the some 20,000 law enforcement agencies at the federal, state, and local levels. Most criminal investigations are performed by the 13,500 local police departments, many of which comprise just a handful of officers, with few if any trained and specializing in investigative work.[4] The following discussion will refer generally to investigators, a category that encompasses mostly police detectives, but also forensic examiners and even patrol officers, who perform a great deal of evidence gathering. Much of the discussion pertains also to prosecutors, who are often involved in one way or another in major investigations, and who are subjected to similar incentives and pressures in the performance of their role. In many respects, the investigative and prosecutorial processes share similar dynamic properties.

It must be appreciated that investigating crimes is a genuinely difficult task. Crimes that receive investigative attention lie mostly in the gray zone between easy cases and unsolvable ones. In many instances, investigators have too little information to generate leads, while in others they are inundated with information that is contradictory or dubious.[5] Investigators are entrusted with a great deal of discretion,[6] much of which is not readily teachable.[7] For example, investigators have discretion in deciding whether a crime occurred, which leads to pursue, what physical evidence to collect, which witnesses to question, which testimonies to trust, when to make an arrest, when to declare the case solved, and when to give up on it. Investigations follow an array of formal and informal policies, practices, and idiosyncratic habits.[8] The investigator's work is encumbered and complicated by departmental directives,[9] public expectations,[10] media exposure,[11] and the passage of time,[12] as well as by limited resources and departmental politics. The prevailing legal rules are onerous,[13] and often confusing, at least until interpreted by courts many months, even years, down the road. Most importantly, as described below, police investigators are encumbered by strong conflicts that pervade their roles. Overall, investigations are conducted in an environment that is hardly suited for the delicate and solemn task. The blue-ribbon committee commissioned by the National Research Council was rather pessimistic about the prospects of reforming the investigatory environment or improving the capabilities of the police to solve crimes.[14]

The accuracy of criminal investigations is bound to be determined by the interrelated cognitive and motivational dimensions of the task. The former pertains to the inferential reasoning involved in any investigative endeavor, while the latter pertains to the particular context of police investigative work. Each of these aspects can contribute to investigative breakdowns.

Cognitive Factors

Abductive Reasoning

In any investigative task, the process of winnowing the field of possible hypotheses to the single substantiated conclusion entails a conceptual problem. To determine the validity of a hypothesis, one needs to obtain evidence that supports or refutes it. Conversely, because it is impossible to seek and test the infinite amount of evidence that might have any bearing on the case, one needs a hypothesis in order to decide which evidence

to test. Hence the circular nature of investigative reasoning: evidence is necessary to test hypotheses, while hypotheses are necessary to decide which evidence to pursue. This dialectical tension makes the investigator's task a most delicate cognitive endeavor.

A form of bootstrapping, known as *abductive* reasoning, is probably the only feasible method suited for conducting criminal investigations.[15] Abductive reasoning is a recursive process of generating and testing hypotheses, geared toward eliminating invalid hypotheses and substantiating the correct one. The testing of hypotheses has two components: a *search* for information, followed by its *evaluation,* that is, the drawing of correct inferences from that information. While the evaluation of the information entails logical inference, the generation of hypotheses and decisions about which information to pursue require intuitive and conjectural thinking. Hence, police investigative work is described not only as a science, but also as a craft, even an art.[16] Following in the mold of Sherlock Holmes, investigators are valued, even valorized, for the creativity of their intuitions.[17]

Performing this bootstrapping task correctly requires fine balancing. A lack of imagination will generate too few hypotheses and thus stands to miss useful information. Excessive creativity, on the other hand, is bound to drain resources on improbable hypotheses and, more importantly, it can lead the process astray. The primary concern is that the evaluative task may be swayed by both cognitive limitations and the motivational aspects of the investigative task.

The Confirmation Bias

A serious concern with the integrity of the investigative process stems from the potential stickiness of the focal hypotheses. While to some degree all reasoning processes rely on underlying knowledge and beliefs,[18] unwarranted conformity of incoming information to extant beliefs is a cause for concern. Investigative hypotheses are, by definition, merely hypothetical scenarios, generated for the sake of exploring particular lines of inquiry. The focal hypotheses must be readily abandoned when they are not adequately supported by the evidence, correctly construed. The threat of bias borne by inertia in investigative reasoning is highlighted by the experimental research on the relationship between extant beliefs and the evaluation of new evidence.

The research indicates that even flimsy thoughts can easily gain traction in people's minds. A number of studies show that merely providing hypo-

thetical explanations or reasons for an imagined scenario strengthens one's belief in the likelihood of its occurrence. For example, asking people to explain why a particular sports team will win a future game increases their belief in the likelihood of that team's victory.[19] Similarly, asking people to imagine a certain outcome of a political election increases their belief in the occurrence of that outcome;[20] and asking people to explain why a particular mental patient might end up joining the Peace Corps (or, conversely, committing suicide) increases their belief in the corresponding future scenario.[21] The research demonstrates also that people tend to anchor their judgments on salient values even when those values are patently arbitrary[22] and adhere to newly formed beliefs even after the evidence that purported to support them has been debunked.[23]

Research on the *confirmation bias* verifies what Francis Bacon described as the "pernicious predetermination" that ensures that one's "former conclusion may remain inviolate,"[24] and what Arthur Conan Doyle's fictional character depicted as the "twist[ing] of facts to suit theories."[25] The key observation of this body of research is that incoming evidence is evaluated in a manner that conforms to the person's extant beliefs.[26] The bias is defined as the "inclination to retain, or a disinclination to abandon, a currently favored hypothesis,"[27] and has also been dubbed the *belief bias,*[28] and the *prior belief effect*.[29] Researchers have also identified the reciprocal *disconfirmation bias,* by which evidence that is incompatible with one's prior beliefs is judged to be weak and thus unlikely to disrupt them.[30]

Confirmatory reasoning has been demonstrated in a number of classic studies. Scientists refereeing an article for publication were more accepting of it when the results comported with their own beliefs than when they contradicted them.[31] The academic performance of a child was judged more favorably when participants were led to believe that she was a high-performing student than when they expected low performance.[32] The bias is strongest in the absence of alternative plausible theories,[33] thus confirming the adage that "nothing is more dangerous than an idea when it is the only one you have."[34] The research on the confirmation bias spans a breadth of domains, including judgments of people,[35] public policy,[36] scientific research,[37] consumer products,[38] and real estate.[39] The bias is observed among novices and experts alike. Doctors and medical students have been found to generate hypotheses at very early stages of the examination, and to adhere to them even in the face of counterevidence.[40] Indeed, premature diagnoses are a leading cause of faulty medical decisions.[41] Psychotherapists have been found to detect psychopathology when

observing normal people whom they mistakenly believed to be psychiatric patients,[42] and to interpret ambiguous psychiatric test results as consistent with disorders that were tentatively suggested to them.[43]

The confirmation bias has also been observed in studies conducted with police personnel, some of whom were experienced investigators. A series of studies with Swedish police officers found that incoming evidence was judged to be stronger when it confirmed the officers' preliminary hypotheses than when it disconfirmed them. For example, eyewitness identifications were deemed more accurate and photographic evidence was deemed more reliable when they supported the investigators' notions than when they contradicted them.[44] A study conducted with Dutch police crime analysts found that even when special analysts are assigned to challenge the investigative team's prevailing theories, they tend to endorse those theories, at the expense of more plausible alternatives.[45] A small study found that a majority of international fingerprint experts judged print matches in a manner that conformed to (misleading) information about the case. In doing so, most of the experts negated their own previous judgments of the same prints.[46]

In the context of criminal investigations, confirmation biases have been labeled *tunnel vision*.[47] The incidence of tunnel vision in criminal investigations is boosted by the fact that the majority of arrests are made in the early stages of the investigation, often by the responding patrol officer.[48] That means that the bulk of the investigatory work is performed well after the suspect has been named and placed under arrest. In other words, investigations are often conducted under strong prior hypotheses regarding the identity of the perpetrator. The guilt-proneness of the bias is probably strengthened also by law enforcement personnel's prevailing attitudes toward questions of law and order. The research suggests that the bias is strongest when the prior beliefs are positively related to the person's stable attitudes toward the topic at hand.[49] Law enforcement personnel tend to subscribe to tough-on-crime worldviews, and are thus more inclined to prioritize the value of crime control over the countervailing values attached to the protection of innocence.[50] It follows that they are also more likely to infer guilt.

In sum, the confirmation bias can wreak havoc in the delicate dialectical task of abductive reasoning. When the evidence conforms to the hypothesis rather than serving to check it, the reasoning process can lose its internal backbone and become even more susceptible to other biasing factors, in particular to the motivational forces discussed below. As with

the majority of psychological phenomena discussed in this book, the confirmation bias need not be driven by conscious or explicit errors. Rather, it occurs almost automatically, under the level of conscious awareness,[51] and is likely to be sincerely denied.[52] Like most other biases, the confirmation bias is most likely to occur when the evidence itself is ambiguous.[53] When the evidence is clear-cut, people are less susceptible to be swayed by biases.

Motivational Factors

Conflicting Roles

Even greater threats to the integrity of investigations stem from motivational factors pertaining to police investigative work. The criminal investigation is a delicate endeavor also in that investigators are entrusted with two distinct tasks. For one, investigators must solve the crime. In whodunit cases, that typically amounts to identifying and locating the perpetrator. In this vein, investigators are expected to search for the best explanation for the event. In addition, investigators are entrusted with constructing the case in preparation for the state's prosecution of the suspect. This task of case construction typically starts at the point at which the suspect is named or taken into custody, and intensifies as the investigation progresses toward its conclusion. The investigative task, then, contains an inherent tension between an objective inquiry and an adversarial-like endeavor of building a case against the suspect.[54] This duality can cause a palpable role conflict. Similar role conflicts are apparent in the work of forensic examiners, whose primary task is to apply scientific methods to discover the truth, but do so almost exclusively on behalf of law enforcement agencies, by whom they are typically employed.[55] Prosecutors, too, are burdened with a dual role, bearing both the responsibility to act as an adversarial advocate and as a "minister of justice."[56] Operating under these discordant goals might prove to be a tall order. The concern is that under some circumstances, the truth-seeking goal will be eclipsed by the adversarial one.

Research on *motivated reasoning* shows that people's reasoning processes are readily biased when they are motivated by goals other than accuracy. These *directional goals* pertain to any "wish, desire, or preference that concerns the outcome of a given reasoning task."[57] Distortions borne by motivated reasoning have been observed in the way people interpret

information suggesting a threat to their health,[58] handle challenges to their competence,[59] perceive the performance of their preferred political candidate,[60] judge the sportsmanship of their sports team,[61] predict their future performance,[62] and assess the odds of winning a bet on a horse race.[63] Motivated reasoning has been observed also outside the laboratory.[64]

It does not take much motivation to skew a reasoning process. A recent study simulating an investigation suggests that the mere assignment to an adversarial role can trump the objectivity of the process. The study found that participants assigned to investigate a case either for the prosecution or for the defense were motivated to see their respective side win the case, and endorsed a biased view of the evidence that was consistent with their role. Those assigned to the prosecution side judged the suspect to be more guilty, whereas the opposite assignment led to more judgments of innocence. A third group of participants, assigned to investigate jointly for both parties, judged the case around midway between the two polarized versions, a result that suggests that they were more neutral in the evaluation of the facts and their judgment of the suspect's guilt.[65] This adversarial mindset was accompanied by a distrust of the (fictional) investigator assigned to work on behalf of the opposite side.[66] These adversarial tendencies were observed in a relaxed experimental setting, in the absence of incentives or tangible goals, and despite an instruction to be fair and objective.

The motivations in real life are considerably stronger. Not unlike any other professional group, criminal investigators take pride in their vocation and derive satisfaction from the execution of their professional duty, namely, solving crimes. For example, fingerprint analysts in the United Kingdom reported feelings of satisfaction, pride, and "a buzz" when finding a match.[67] Yet the motivation to clear the case runs deeper. Most law enforcement personnel identify themselves as fighters in the War on Crime.[68] Bringing criminals to trial is the noble cause to which they devote their careers and for which some risk their lives.[69]

Perhaps the strongest goal motivating investigators stems from the pressure to clear cases—that is, to arrest and charge the suspect.[70] In a public announcement of the arrest in the case of Darryl Hunt, the chief of police of Winston-Salem, North Carolina, stated: "We spent hundreds of man hours on this case but of course, our objective, from the very beginning was to make a charge, and we have accomplished that."[71] Clearing cases is the most common measure of departmental effectiveness.[72] At the level of the individual investigator, clearing cases is a measure of per-

sonal success. It reflects on her professional reputation, standing among her peers, and prospects for promotion.[73] By the same token, the failure to close cases can be costly at the departmental and personal levels. Low clearance rates are used as a key tool in disciplining management within police departments.[74] Low rates can also result in demotion of investigators to positions of lesser status.[75] Indeed, the pressures to clear cases has led to occasional distortions and misrepresentation of crime data by police departments in the United States and the United Kingdom.[76] The costs of failure are particularly steep in cases considered to be high-profile. Such cases are not limited to sensational or heinous crimes. Most violent crimes—especially rape-murders, sex offenses against children, and serious felonies committed in small towns or neighborhoods—have the potential to destabilize the community and generate heightened pressure for its resolution.

The pressure to clear cases is exacerbated by the generally low rate of solving cases through detective work. As mentioned in Chapter 1, only half of the serious crimes committed are reported to the police, and only one in five of these are cleared by arrest. A landmark study by the RAND Corporation found that a majority of the serious crimes that get cleared are solved during the first encounter with the responding patrolman. Even serious crimes are often resolved with the provision of the name of the suspect by victims or witnesses, thus obviating the need for detective work.[77] Many crimes that cannot be solved promptly will never be cleared.[78] Thus, detectives walk away from a great many crime scenes knowing that they are unable to do anything to solve them.[79] This reality flies in the face of the pragmatic, action-oriented, and getting-the-job-done culture that pervades investigative units.[80] The frustration of the investigators' goals is likely to increase their motivation to clear the relatively few crimes that are actually investigated.

Effects of Emotion

It is not hard to see how investigators can get emotionally involved in their cases. For one, investigators can develop personal relationships with the victims or their families, thus strengthening their resolve to apprehend the perpetrator.[81] Investigators are often exposed to the human tragedy inflicted by crime and confronted with gruesome crime scenes. This exposure has the potential to arouse intense negative emotions, particularly anger and also disgust.[82]

The research indicates that high levels of anger arousal tend to result in shallow processing of evidence and hostile judgments of other people. Specifically, anger has been found to result in stronger attributions of personal blame for negative outcomes, higher propensities to perceive other people's conduct as intentional, lower thresholds of evidence, and a stronger tendency to discount alternative explanations and mitigating circumstances.[83] Anger has also been found to increase reliance on stereotypes,[84] desire for retaliation,[85] and motivation to take action to remedy the transgression.[86] In a study simulating fingerprint analysis, participants displayed a heightened tendency to find positive matches after being shown gruesome photographs of the putative murder victims.[87] A study of experienced Swedish police officers showed that the arousal of anger resulted in superficial processing of information and a lack of sensitivity toward exculpating evidence.[88] In all of these studies, the person being judged was unrelated to the source of the anger. In other words, when in a state of anger, people are more harsh in their judgment of *any* other person. It is not hard to see how people would react angrily toward a person who is believed to have committed a heinous crime.

Group Membership

Another notable feature of criminal investigations is that they are performed within the social setting of group membership. Investigators generally view themselves as belonging to the group of people working for law enforcement agencies, and who share the common goal of fighting crime. This in-group includes detectives, patrol officers, forensic examiners, prosecutors, and sometimes also victims and witnesses for the prosecution. Importantly, the in-group contrasts itself starkly with the out-group, consisting primarily of criminal offenders—who are generally deemed to be bad people, often referred to as "scumbags"—and sometimes also their defense attorneys. Suspects can readily be lumped into this out-group, whether because of their status as the presumed perpetrator or because of their criminal history.

The research indicates that group membership constitutes an important component of people's identity and is integral to their self-concept.[89] People tend to consider their groups to be trustworthy, competent, moral, and peaceful, while out-groups are generally regarded as untrustworthy, competitive, and aggressive. This *in-group favoritism* and *out-group derogation* have been observed in numerous laboratory studies as well as in

anthropological work.[90] This polarization is fueled by the adversarial nature of the legal process.

The group setting has the potential to sway criminal investigations toward conclusions of guilt. Group members tend to share similar worldviews, beliefs, and stereotypes about out-group members.[91] Groups tend to exert cohesive forces on their members when working toward the group's shared goal,[92] which in the case of criminal investigations is fighting crime. The joint endeavor makes the members more prone to reach a consensus and to conform to group norms.[93] Reflecting on the successful prosecution of a man who was subsequently exonerated by DNA evidence, the prosecutor stated: "Maybe I was too willing to believe what the law-enforcement officers told me. Maybe I got caught up in the sense that the prosecutor and the investigators are all on the same team."[94]

Groups, particularly homogeneous groups, have been found to search for information in a selective manner,[95] display the confirmation bias,[96] and respond to threats with arousal of anger followed by superficial processing of information.[97] Excessive cohesion can reach the pathological state of groupthink.[98] Importantly, the group setting has disinhibiting effects on its members, enabling them to overcome inhibitions that would normally prevent them from acting in their individual capacities.[99] For example, people are more likely to engage in binge drinking when that conduct is an acceptable norm of their group.[100] Groups have been found to be more aggressive than individuals in electrocuting another person,[101] and in forcing adversaries to eat hot sauce.[102] This heightened aggression is accompanied by a reduced sense of moral responsibility.[103] Group members are particularly prone to shed moral responsibility when they can attribute primary responsibility for aggressive behavior to other members of the group.[104] Group membership also makes it easier for individuals to discount, overlook, or turn a blind eye to the misdeeds of other members of the group.[105]

Commitment

Another potential problem stemming from the dynamics of police investigations is that as the process unfolds, investigators become increasingly invested in the focal hypothesis. They devote significant time and resources to pursuing their theory of the crime, and sometimes they invest personal capital in proving it to be correct. This sense of personal

investment is likely to be heightened when a suspect has been placed under arrest, something that happens routinely once he has been named.

The admission of an error poses a threat to the ubiquitous need to maintain a positive self-conception,[106] particularly in regard to one's competence,[107] morality,[108] and consistency.[109] This motivation is understood to serve both private and social needs, that is, to maintain a positive conception in one's own eyes as well as in the eyes of others.[110]

A substantial body of experimental research demonstrates that people tend to adhere to their prior courses of action, even in the face of indications that they were wrong in the first place.[111] One explanation for this *escalation of commitment* stems from a favorable distortion of one's original course of action, which serves to negate any prior fault.[112] Studies show also that committed people tend to search selectively for information to justify their prior decisions rather than prepare themselves for future ones.[113] Committed people tend also to interpret incoming information in a distorted manner that serves to justify those decisions.[114] The escalation of commitment has been observed also in naturalistic settings. A study of NBA teams reveals that costly players receive preferential treatment that is not warranted by their performance on the court,[115] bank managers tend to be committed to bad loans that they had personally approved,[116] holders of theater season passes are more likely to attend the shows if they paid full price for them,[117] and managers provide inflated ratings of employees whom they hired.[118] Commitment effects were observed in a study that simulated a criminal investigation. The mere naming of the suspect at an initial phase of the study led participants to a stronger belief in that suspect's guilt, which resulted in a failure to explore alternative theories adequately. These participants sought additional information to confirm those initial hypotheses and evaluated it in a way that corresponded to those beliefs.[119]

The research has identified a number of task features that exacerbate the escalation of commitment, many of which are likely to be present in criminal investigations that go astray. Commitment has been found to increase along with increases in the actor's responsibility for the original error,[120] the room for concealing the failure,[121] the adversity of the outcome of the original decision,[122] the perceived threat entailed by the exposure of the error,[123] and the publicity of the original error.[124] Paradoxically, the more egregious the error and the longer it has persisted, the less likely it is that it will be corrected.[125]

Commitment to a faulty course of action is bolstered also by the group setting. Groups are prone to escalate their commitment to failing courses

of action,[126] at times more so than individuals.[127] More importantly, groups exert strong disciplining powers on their individual members. Group deviance draws criticism, hostility, and ostracism, as evidenced by the typically harsh reaction to whistle-blowers.[128] Groups also retaliate more forcefully than individuals.[129] The more cohesive the group, the more strongly it condemns its deviants.[130] Thus, any doubt expressed by a law enforcement agent might be regarded as a challenge to the group's consensus and thus also as a breach of loyalty, a value that is valorized in police culture.[131]

Admitting error is complicated further by investigators' sense of commitment to the soundness of their investigative methods. A botched investigation punctuates that eyewitnesses cannot always be trusted, memories can be mistaken, confessions can be elicited from innocent people, and forensic tests can be off the mark. These prospects could be disconcerting, especially given that investigators live by these methods, defend them in court, and are bound to use them in investigations to come.

In sum, the commitment effect can deflect incoming information that challenges the focal hypothesis, that is, evidence that either indicates the innocence of the suspect or implicates a different person in the crime. Commitment can thus make it difficult to upend an advanced investigation or indictment, not to mention a conviction.[132] While it does not take much to become the target of a criminal procedure, it can be very difficult to reverse that status.[133] Strong commitment is observed also in the behavior of prosecutors, most notably when they proceed to prosecute defendants even after the latter have been exculpated by compelling evidence such as DNA testing.[134] For example, prosecutors in Orange County, California, went forward with the prosecution of an innocent man on charges of carjacking and armed robbery even though he had been excluded by both DNA and fingerprint tests. In defending the decision to prosecute, Assistant District Attorney Marc Rozenberg explained: "If nobody had identified him, we wouldn't have prosecuted this case."[135] Prosecutors persisted with the prosecution of some fifteen other defendants in the face of exculpatory DNA tests, all of whom were subsequently exonerated.[136]

Motivations Combined: The Adversarial Pull

The discussion thus far has indicated that in contested cases, investigators experience a cumulative set of motivations that drive them toward conclusions of guilt. In actuality, the picture is more nuanced. Investigators

are undoubtedly motivated also by the goal of finding the true perpetrator and avoiding the incrimination of innocent people. Balancing these opposing pulls places investigators in the thick of a difficult and potentially stressful[137] role conflict.[138] Just how investigators resolve the conflict will depend on a host of circumstantial and personality factors.

The concern is that in the demanding circumstances of contested investigations, the truth-seeking goal will hold less sway. First, from the investigator's perspective, the need to avoid false charges is an abstract principle, not a concrete incentive. It is a constraint that is often experienced as a hindrance to the goal of clearing crimes, hardly a desideratum in its own right. Investigators get rewarded and recognized predominantly for making arrests, not for refraining from charging innocent people (imagine a chief of police convening a press conference to announce that although a dangerous perpetrator remains on the loose, the department succeeded in not arresting any innocent persons). Second, the opposing goals have very different feedback mechanisms. As mentioned, a failure to clear the case can bring immediate and tangible negative repercussions upon both the officer and the department, especially under the spotlight of the media. In contrast, a mistaken suspicion of an innocent person might well go undetected. Ironically, mistaken suspicions can be buried under compelling (yet mistaken) evidence that was induced by them. The prospect that mistakes may come to light months, years, or decades down the line tends not to weigh heavily in the rough-and-tumble exigencies of the moment.[139] Third, not unlike most other people, investigators are likely to hold unrealistically positive views of their own performance,[140] and thus tend to believe that they have not erred. Indeed, law enforcement agents and judges tend to believe that virtually no innocent people are convicted, at least not in their jurisdiction.[141] Finally, investigations often gravitate toward the usual suspects, that is, people with criminal records. These suspects are often engaged in some form or another of criminal activity, and may be deemed to have escaped punishment in the past. The investigator might feel less troubled by a false charge made against such suspects, and perhaps even welcome the opportunity to rectify their impunity for past misdeeds.

In sum, it is not hard to imagine that under certain circumstances, the truth-seeking objective will be overridden by goals and motivations that lead investigators toward a more adversarial, conviction-prone stance.[142] This *adversarial pull* was captured by Justice Robert Jackson's characterization of "the often competitive enterprise of ferreting out crime."[143]

To be sure, there are differences in the degree of adversarial pressures generated across jurisdictions, departmental procedures, and local professional cultures. There is variance also among the people who conduct investigations. Investigators differ in their professional temperament, which is probably affected by their introspection, integrity, conformity, and susceptibility to incentives. There is reason to believe that most law enforcement personnel in most police departments withstand the adversarial pressures, and conduct thorough and fair investigations. The adversarial pull, however, is likely to wreak havoc in investigations conducted under intense pressures and performed by those who lack a disciplined professional temperament. The adversarial pull is evident in the inculpatory bent of the scientific methods developed by forensic scientists. Many of these methods lack adequate scientific grounding, and some are plainly junk science.[144] The truth-seeking objective is most likely to be overridden in high-profile cases, where the pressures to solve the crimes are the strongest.[145] In some instances, the adversarial pull results in deliberate police malfeasance,[146] and even entails lying outright in court, a practice known as *testilying*.[147] Recall, however, that we are interested primarily in conduct that does not involve deliberate dishonesty.

Over time, the adversarial pressures that develop in the course of criminal investigations and prosecutions are bound to become internalized in the prevailing culture and practices of law enforcement agencies.[148] This internalization helps explain the gradual strengthening of tough-on-crime attitudes among police officers.[149] These attitudes, in turn, contribute to the inclination to infer guilt.

The Coherence Effect

One of the distinctive features of difficult cases is that they entail drawing inferences from multiple evidence items, all of which need to be integrated into a singular factual assessment and expressed in the form of a binary conclusion. This task is no light matter given the uncertainty, incommensurability, and conflict within the information that is often encountered in the course of investigations. For example, an analysis of the evidence presented in the trial of Sacco and Vanzetti identified more than 300 facts and propositions.[150] A cognitive process is needed also to integrate the available information with other aspects of the task, including the person's motivational and emotional responses to it. The

cognitive process that performs this integrative task poses another threat to the accuracy of the investigative task.

The integration of evidence in complex decision tasks lies at the core of the body of research on the *coherence effect*. This psychological phenomenon can be encapsulated by the Gestaltian notion that *what goes together, must fit together.* Complex tasks can be solved effectively and comfortably when they are derived from coherent mental models of the case at hand,[151] that is, when the conclusion is strongly supported by the bulk of the evidence. This coherence effect is driven by a bidirectional process of reasoning: just as the facts guide the choice of the preferred conclusion, the emergence of that conclusion radiates backward and reshapes the facts to become more coherent with it.[152] This process occurs primarily beneath the level of conscious awareness.[153] The coherence effect has been observed in decision-making tasks as well as in tasks that involve general cognitive processing such as memorization of information or recounting it to another person.[154] In itself, the coherence effect is probably adaptive, in that it enables people to reach conclusions and make decisions even when the task is most complicated and difficult. Still, this phenomenon has serious implications for both the investigative and adjudicative processes.

First, coherence is achieved by spreading the evidence apart into two (or more) clusters, each corresponding to a different conclusion. The evidence supporting the emerging conclusion becomes stronger, while the evidence supporting the rejected conclusion wanes. Thus, the cognitive process transforms the evidence from an initial state of conflict into a lopsided evidence set that clearly supports the decision. In other words, the evidence comes to cohere with the emerging decision. This *spreading apart* results in the dominance of one conclusion over the other, thus enabling confident action. For example, one study presented participants with a theft case that contained a range of unrelated evidence items, including an eyewitness identification, a possible motive, an unexplained possession of money, and an alibi claim. The study found that people tended strongly to evaluate the evidence in a coherent block, all pointing toward either inculpation or exculpation.[155] The spreading apart enables people to reach concrete conclusions even when they originally perceived the evidence as ambiguous and conflicting. It must be appreciated that to some degree, the apparent strength of the evidence that enables confident action is an artifact of the cognitive process rather than an objective assessment of the case at hand. Thus, investigators will tend

to perceive the evidence that supports their conclusion as stronger and more corroborative than it really is.

In itself, the coherence effect is nondirectional, in that it can inflate judgments of guilt and innocence alike. However, in combination with other biasing factors—notably, motivations and confirmatory biases—it can sway the judgments of the entire case in the direction of those factors.[156] The coherence effect can be seen operating in the abovementioned study of the confirmation bias in judgments of Swedish police investigators. When the witness's account was consistent with the investigators' theory of the crime, the investigators also judged her to be more reliable, the witnessing conditions to be better, and the memory loss over the seven-day interval to be less harmful.[157] Similarly, the abovementioned study that simulated an investigation found that the ambiguous fact pattern was evaluated in a manner that cohered with the participants' assigned roles.[158]

A second feature of the coherence effect is that information items are not evaluated independently, but rather according to how they fit into the mental model of the task. As a result of the interconnectivity of the Gestaltian process, any evidence item can impact all other items, and ultimately the entire case. One important facet of this *nonindependence* feature is that including an evidence item that is strongly inculpating can make the entire evidence set appear inculpating, just as including an exculpating item can result in a conclusion of innocence. This nonindependence naturally adds a directional dimension to coherence shifts, driving the entire set of evidence toward the corresponding conclusion. This phenomenon of *circuitous influences* was observed, for example, in the abovementioned study of a theft case. Adding information that placed the suspect near the scene of the crime resulted in a higher rate of convictions, as one would expect. Interestingly, it also resulted in more inculpating evaluations of all the other evidence items, such as greater trust in the eyewitness's identification, and a weaker belief in the defendant's explanation for his possession of money days after the theft.[159] Similarly, describing the defendant in a libel suit as motivated by good intentions led to the strengthening of various legal and factual reasons supporting his defense, whereas portraying him as motivated by greed resulted in opposite inferences.[160] In the absence of any direct relationship between these extraneous manipulations and the rest of the case, one must infer that these influences occurred through the circuitous connections of the cognitive system.

Circuitous influences were observed incidentally in a number of studies that showed how extraneous evidence can contaminate witnesses' statements. Eyewitnesses were more likely to pick the suspect at a lineup when they were told that he had confessed to the crime, and were less likely to identify him when told that another suspect had made a confession.[161] In a series of studies discussed in Chapter 3, eyewitnesses who identified the wrong person at the lineup were provided with (fictitious) affirmation of their identifications ("Good, you identified the suspect"). In addition to inflating the witnesses' confidence in their identifications, this feedback also distorted a range of judgments surrounding the witnessing of the criminal event, including the witnesses' assessment of how good a view they got of the gunman, how well they were able to make out the specific features of the gunman's face, how much attention they were paying to the gunman's face, how easy it was to identify him, and how quickly they picked him out at the lineup.[162] These judgments can be of considerable significance to the outcome of the case, in that they are typically viewed by third parties—investigators, prosecutors, and jurors—as indicators of testimony reliability. Given that all these identifications were wrong (actual targets were not placed in the lineups), the apparent corroboration is misleading.[163]

Circuitous influences were observed incidentally also in studies conducted with experienced criminal investigators. Polygraph examiners were more likely to interpret ambiguous physiological data as indicative of deception when told (fictitiously) that the suspect had confessed to the crime.[164] Fingerprint examiners from a number of countries were likewise influenced by (fictitious) knowledge that the suspect had confessed to the crime or that he was under arrest at the time.[165] The contaminating potential of circuitous influences poses a serious concern in police investigations, as it does for jury decision making as discussed in Chapter 6. Investigators are often exposed to a variety of information of varying reliability from informants, fellow detectives, witnesses, media reports, and physical evidence. Exposure to any such erroneous information can sway the evaluation of the incoming evidence and thus influence the direction of the investigation.

Five Mechanisms of Biased Reasoning

It would be helpful to briefly describe five common mechanisms by which biasing processing operates. Recognizing these mechanisms could

assist in identifying biased reasoning processes. The mechanisms can work independently or in combination.

Selective framing strategy. One way of enhancing the compatibility of evidence with a preferred conclusion is to frame the inquiry in a manner that affirms the salient hypothesis. This mechanism, observed early on by Jerome Bruner and colleagues,[166] has been replicated in numerous studies and been labeled the *positive test strategy*[167] and the *verification bias*.[168] The strategy has been described as looking for "features that are expected to be present if the hypothesis is true."[169] For example, when people are instructed to determine whether a conversation partner is an introverted person, they tend to ask questions that confirm introversion (for example, "In what situations do you wish you could be more outgoing?"), and to phrase the questions in the opposite form when being asked to determine whether the person is an extrovert (for example, "What would you do if you wanted to liven things up at a party?").[170] It is not hard to see that different ways of framing the investigative task will result in different courses of action. As discussed in Chapter 4, subtle differences in the phrasing of questions can readily affect the responses given by witnesses. A positive test strategy will naturally yield leading questions that drive the witness toward affirming the interviewer's implicit assumptions. As discussed in Chapter 5, a study simulating an interrogation found that interrogators who were led to believe in a higher likelihood of guilt asked more guilt-presumptive questions, which in turn elicited responses that made the suspect look more culpable.[171]

Selective exposure. Another way to reach particular conclusions is to choose which evidence to examine for the purpose of testing the chosen hypotheses. The research indicates that people tend to selectively expose themselves to information that confirms their focal hypothesis and shield themselves from discordant information.[172] This pattern is readily apparent in people's choices of news media (compare the political attitudes of the viewers of Fox News and MSNBC).[173] Likewise, recent car purchasers tend to read advertisements of the car they bought more often than of other cars they considered but did not buy.[174] Experimental evidence includes findings that people are more likely to seek favorable rather than unfavorable information about themselves.[175] The selectivity of exposure becomes more acute when the information is scarce,[176] which is often the case in criminal investigations. Selectivity can also

take the form of actively seeking evidence that is expected to inhibit a countervailing hypothesis.[177]

Selective scrutiny. Desired conclusions can also be derived by altering the standard for validating the incoming information. The research demonstrates that people tend to scrutinize information that is incompatible with their conclusion, but apply lax standards when assessing the validity of compatible information.[178] When people contest adverse positions, they spend more effort, generate more refutational thoughts, and muster more redundant counterarguments.[179] People who receive unfavorable results from a putative intelligence test tend to challenge its validity, but accept it at face value when they receive a favorable score.[180] Likewise, people react skeptically to a medical diagnostic test when it indicates that they are susceptible to a disease, but tend to accept it readily when it finds no such indication.[181] A study of peer reviewers of a scientific publication found that the referees were more likely to notice a typo in the submitted article when the result of the research contradicted their beliefs.[182]

Biased evaluation. The objectivity of the evaluation is key to the integrity of any investigation. Yet the most ubiquitous form of biased reasoning occurs through a distorted evaluation of evidence. Biased evaluation[183] features in the bulk of the abovementioned research, including evaluating a shove as either jovial or aggressive depending on the race of the actor,[184] maintaining that your preferred political candidate did better at a debate than his rival,[185] believing that a physical encounter in a football game was a foul if it was committed by the rival team but a legitimate hit if it was committed by a player on your favorite team,[186] and inflating the odds that your chosen horse will win a race.[187] Biased evaluation figures also in the studies that find distorted judgments in the testing of forensic evidence.[188]

Selective stopping. Finally, a limited body of research suggests that people tend to shut down inquiries after having found a sufficient amount of evidence to support their leading hypothesis.[189] This means that police investigations might be aborted prematurely. In particular, the inquiry might be stopped before information that tends to refute the police's hypotheses has been adequately considered.[190] As a veteran Dallas detective explained after learning that his investigation had incriminated an innocent man, "You think you have a slam-dunk case, and so you don't go in

there and dot your I's and cross your T's." The detective added that it is only after the conviction has proved to have been mistaken that it "comes back to bite you."[191]

The Opacity of Investigations

There is reason for hope that investigative errors would be corrected by the mechanisms of oversight that are embedded in the criminal justice process. After all, the investigator's work product is subjected to the judgments of superiors, prosecutors, judges, defense attorneys, and jurors, all of whom are designated to assess the evidence with a critical eye. This institutional oversight parallels the psychological construct of *accountability,* wherein people anticipate being called on to justify their performance to others. The construct implies that one expects to gain praise or suffer negative consequences depending on how she is deemed in the eyes of an intended audience. As developed in the research of Philip Tetlock and his colleagues, accountability can improve otherwise inferior performance through a process of preemptive self-criticism, by which people anticipate and preempt the expected objections of their would-be critics. Accountability has been found to lead to closer attention to evidence, higher calibration between confidence and accuracy, increased sophistication of thought processes, and lower effects of emotions on unrelated judgments.[192]

Still, despite the multilayered oversight built into the criminal justice process, accountability might not yield its intended effects. Naturally, the ameliorative effects of accountability are limited to situations in which the relevant audience is well informed about the issue at hand.[193] Accountability is not a viable construct for people whose conduct remains out of sight. In other words, accountability depends on transparency, but criminal investigations are invariably opaque. Recording practices vary among investigative agencies, but are rarely complete or objective. By their own admission, 33 percent of lineup administrators fail to keep any written reports of the lineups, and 27 percent do not keep a photographic record of the procedure.[194] In many jurisdictions, only 7 percent of administrators videotape the lineup procedures.[195] This opacity deprives outside reviewers of information such as witnesses' choices, confidence, other statements about the suspect, the speed of the choice, and any statements made by the administrator. As discussed in Chapter 6, this information could be vital to the assessment of the identification. In about one-half of

the eyewitness identification cases that have been decided by the U.S. Supreme Court, the Court noted the incompleteness of the record of the procedure (yet proceeded to discuss the reliability of the identifications—invariably, favorably—with no apparent concern over the missing information).[196]

In the bulk of interviews with cooperative witnesses, the record consists mostly of retrospective paraphrases jotted down by the investigator. These practices result in a loss of a considerable amount of information provided by the witness in almost all of the questions asked. For example, an experiment with forensic and child service interviewers found that between 20 percent and 40 percent of details provided by children and more than 80 percent of the interviewers' questions were omitted from the interview reports.[197] Similarly, a field study that tested real-life interviews of child abuse cases found that even when taking contemporaneous verbatim notes, interviewers missed about one-quarter of the details reported by witnesses and omitted more than one-half of the substantive questions that they had asked.[198] A study with experienced police investigators in Florida found that their reports missed some two-thirds of the information stated by the witnesses and did not include any of the questions that they had asked.[199] There is little reason to expect that witnesses will remember much more. This opacity is problematic, since errors can be induced in barely noticeable ways, such as by subtle phrasing of questions or mere hints of suggested information. The absence of a reliable record is particularly acute in the context of interrogations of suspects, where disputes often arise over both the content of the statements attributed to the suspect and the investigative means used to elicit them. As investigators are aware, in the bulk of these *swearing contests,* their word is trusted over the suspect's. In sum, the opacity of investigations obscures much evidence from the legal actors and gives investigators little reason to be worried about accountability.[200]

The Investigation of Brandon Mayfield

The investigation of Brandon Mayfield, an Oregon lawyer suspected of involvement in an Al Qaeda terrorist attack, provides a good illustration of the potential of police investigations to go astray. This case is instructive in that it has been the subject of especially thorough inquiries, one conducted by the FBI and one by the Office of the Inspector General of the Department of Justice (DOJ).[201] The Mayfield affair jolted the finger-

print identification community, which has long insisted on an error rate of zero, a claim that has been endorsed repeatedly by the courts.[202] The affair also demonstrates that even highly regarded professionals in a flagship law enforcement laboratory can get swept up in the dynamics of an investigation gone awry and ultimately insist on improbable conclusions.

In March of 2004, the FBI was called on to assist the Spanish National Police with the investigation of a massive terrorist attack by Al Qaeda on commuter trains in Madrid. A computerized fingerprint identification suggested Mayfield as a possible match with the latent prints of a person implicated in the attack. It is not hard to see why Mayfield seemed like a fitting suspect. Mayfield, an army veteran, had converted to Islam and was married to an Egyptian woman. He had previously represented a Muslim man convicted of terrorist conspiracy in a child custody dispute.[203]

A high-ranking FBI fingerprint specialist examined the prints and concluded a positive match. The match was subsequently confirmed by a retired FBI examiner with over thirty years of experience. The process was overseen by a specialist who headed the FBI's Latent Print Unit. Two weeks later, federal prosecutors applied to federal court for a warrant to search and detain Mayfield as a "material witness." The application was based primarily on FBI affidavits stating that the match provided a "100% identification of Mayfield."[204] Mayfield was arrested, and reportedly was told that he was being investigated in connection with crimes punishable by death.[205] The FBI's conclusion was subsequently confirmed by another fingerprint analyst who was appointed by the court. Soon thereafter, however, the Spanish Police found that the latent prints actually matched an Algerian national by the name of Daoud Ouhnane. After reviewing the indisputable match with Ouhnane's prints, the FBI withdrew its identification of Mayfield and released him.

From the moment that Mayfield's prints were declared a match with the Madrid train bomber, the investigation rolled ahead like an unstoppable freight train. As the author of the FBI report stated, "Once the mind-set occurred with the initial examiner, the subsequent examinations were tainted."[206] The FBI's motivation to name Mayfield as the suspect is also apparent. It is not a fancy speculation to state that the FBI was keenly interested in cracking the identity of the Madrid train bombings. Identifying the terrorists would have been beneficial in securing the cooperation of the Spanish government, then a reluctant ally of the United States in the Iraq war. The prospect of solving an Al Qaeda attack perpetrated on the soil of a friendly European country would also have been a

boon for the United States in its Global War on Terror. Finally, linking an American Muslim to an Al Qaeda ring would have provided a justification for the government's domestic antiterrorism efforts and garnered support for its controversial legislative agenda, notably the Patriot Act. Indeed, the high profile of the Mayfield case was cited in the FBI report as a central reason for the faulty fingerprint match.[207]

The Mayfield investigation manifests the FBI examiners' selective exposure to the available evidence. As the DOJ report points out, some of the print similarities considered important by the FBI were visible in only one of the several sets of Mayfield's prints that were available.[208] In other words, the examiners focused on the source of information that provided evidence that confirmed the hypothesis, while ignoring equally reliable information that contradicted it. The examiners also treated the evidence with selective scrutiny. The comparison of the prints relied on a number of similarities within extremely tiny details of the prints ("Level 3" details), whose validity is considered controversial.[209] Unhindered by these concerns, the Mayfield examiners cited similarities in numerous tiny details to justify their match.[210] At the same time, they completely dismissed the fact that the entire upper left portion of the latent print did not correspond with Mayfield's print.[211]

Biased evaluation permeated this investigation. The FBI claimed to have found fifteen points of similarity between the sets of prints. According to the DOJ report, the examiners interpreted some murky and ambiguous details in the latent print as similar to Mayfield's.[212] It also turned out that the latent print originated from Daoud's right middle finger, whereas the FBI matched it to Mayfield's left index finger.[213] The examiners also engaged in "backward" reasoning that led them to "find" additional similarities that did not exist.[214] The case also manifests the effects of group membership. As noted in the FBI report, the analyses of the second and third FBI examiners were likely constrained by the pressures of cohesion: "To disagree was not an expected response."[215] The effect of group membership was evident also in the FBI examiners' apparent overconfidence and sense of superiority to their Spanish counterparts.[216]

Perhaps most notable in this case was the FBI team's commitment to the initial identification of Mayfield. Days after learning of the FBI's conclusion, the Spanish police notified the FBI that the match was "negativo." This red flag failed to spur a reexamination of the FBI's findings. Rather, the agency elected to arrange a meeting with its Spanish counterparts in Madrid to persuade them of the validity of the match. The meeting, held

eight days later, did not go well. As reported by a Spanish official, the FBI insisted that the prints shared fifteen similar "points," whereas the Spaniards found only seven similarities.[217] The Spanish representatives kept pointing out discrepancies between their analysis and that of the FBI, but these "did not seem to sink in with the Americans." "They had a justification for everything," explained the head of the Spanish fingerprint unit, "But I just couldn't see it."[218] At the conclusion of the meeting, the FBI extracted a promise from the Spaniards to reexamine the prints. The pressure from the FBI persisted. As the Spanish official explained, for three weeks following the meeting, the FBI "called us constantly," "they kept pressing us."[219] The Mayfield case also punctuates the problems that stem from the opacity of the investigation. Even though this investigation was conducted in the comfort of an FBI facility, the examiners did not record the reasons that led them to their conclusions.[220] The precise reasons for this investigative debacle thus remain unknown.

This case also demonstrates that a commitment to an erroneous investigation can cause a secondary contamination of the adjudicative procedure. In attempting to justify the faulty investigation to the court, FBI personnel and their lawyers made unfounded and distortive statements in support of Mayfield's arrest warrant. The government misrepresented the uncomfortable discrepancy with the Spanish police analysis, stating that the Spaniards had "felt satisfied" with the FBI's conclusions and promised to reexamine their findings.[221] The FBI was also hard pressed to show that Mayfield had actually traveled to Spain: there was no record of his travel, and he did not own a passport. With no evidence to support the claim, the federal government's affidavit stated: "It is believed that Mayfield may have traveled under a false or fictitious name."[222] In an apparent attempt to link Mayfield with the Madrid bombings, federal agents claimed to have confiscated "miscellaneous Spanish documents" from Mayfield's office and home. According to a source close to Mayfield, these documents were his young children's Spanish homework.[223]

It is worth noting that the evidence produced by the FBI would probably have sufficed to have Mayfield convicted and sentenced to death. He was confronted with unwavering inculpating evidence from the most prestigious crime laboratory in the land, grounded in scientific testimony and backed up by a court-appointed expert. Under normal circumstances, it would have been close to impossible to uncover the FBI's error. Absent the verifiable and indisputable match to the true suspect, Mayfield's fate could have been quite different.

Recommendations for Reform

This chapter has examined the manner in which police investigations are conducted and has emphasized their dynamic properties. The discussion indicates that properly conducted investigations require delicate cognitive processing that might not be afforded in the harsh realities of contested criminal investigations. In themselves, the cognitive biases are mostly nondirectional, in that they merely bolster investigative conclusions regardless of whichever conclusion they support. However, the process can be swayed strongly by a variety of motivational forces, which tend to pull investigations toward adversarial, guilt-confirming conclusions. In reality, the cognitive and motivational phenomena often operate contemporaneously, producing a potent recipe for biased processing.

The foregoing analysis has focused on the investigators themselves, particularly how they seek, test, and evaluate information. It must be acknowledged that this discussion speaks to only one dimension of the dynamic process. It omits the crucial dimension pertaining to the impact of the investigation on the evidence that it produces, primarily, on human testimony. The next three chapters examine how investigative beliefs can seep into witnesses' testimony and induce it to conform to them. Hence the escalating dynamic of investigations: investigators' hypotheses tend to generate confirmatory testimony that bolsters those hypotheses and turns them into firm conclusions. As discussed in Chapter 6, this *pseudo-corroboration* can have a strong impact on criminal outcomes. Investigations that begin with an initial mistaken conception of the case are prone to perpetuate that error. This is mostly likely to occur in investigations in which the adversarial pull is particularly strong.

While adversarialism is one of the hallmarks of the Anglo-American adjudicative process, it seems hardly controversial that it is fundamentally unsuitable in the investigative phase of the process. To have any chance to succeed as a fact-finding device, adversarialism requires a contest between opposing accounts of the facts. In reality, investigations are virtually monopolized by the police and other state investigative agencies. The state has virtually exclusive access to the crime scene, the physical evidence, the databases, the victim, and most witnesses. The state also has exclusive power to search, seize evidence, place people under arrest, and wield its prosecutorial power as a threatening device. In contrast, the defendant is afforded a very limited scope for conducting investigations, especially when she is incarcerated. Yet even if suspects enjoyed equal investigative

powers, the vast majority of non-white-collar suspects could not afford to avail themselves of them. Effectively, investigations are driven by a one-sided quasi-adversarial process, on which the accounts of the state go largely unchecked and unopposed. Like one hand clapping, this one-sidedness guts any virtue that the adversarial procedure might have harbored, yielding an unfitting method of truth discovery. The following discussion examines possible avenues to enhance the accuracy of police investigations.

Debiasing: Considering Alternatives

A natural approach to tackle the confirmation bias is to debias it. One possible way to do so is by promoting a healthy skepticism and lateral thinking, that is, by introducing mechanisms for the generation of alternative hypotheses.[224] Suggestions along these lines have been introduced in police training in the United Kingdom,[225] and have been mandated by Canadian courts.[226] One concrete experimental intervention that has been used with some success is instructing people to "consider the opposite" hypothesis.[227] Instituting this practice in criminal investigations could be done via forcing investigators to consider alternative hypotheses and elaborate on the reasons for rejecting them.

To be sure, this intervention should be welcomed, but it should be acknowledged that its effectiveness will likely be limited. Debiasing instructions have been found to be successful in correcting relatively weak cognitive failures, such as where lines of thought were neglected because of lack of sufficient attention.[228] They have proved less successful in correcting reasoning processes where cognitive biases were compounded by motivational factors.[229] There is reason to suspect that debiasing instructions will fall short of overcoming the strong motivational biases that are often present in contested criminal investigations. Moreover, this intervention can backfire, resulting in the bolstering of the focal hypothesis.[230]

Functional Separation

Another possible way of promoting the integrity of investigations is to introduce procedures designed to provide a critical appraisal of the focal hypothesis. The objective would be to scrutinize investigations and correct them when they go askew. Dialectical reasoning is an intervention

that designates some of the team members to offer a countertheory to the prevailing focal hypothesis in order to instigate a structured debate about the merits and weaknesses of the vying hypotheses.[231] This technique has been found to reduce commitment to prior choices.[232]

Although functional separation is an elegant solution, in practice it is a complicated proposition and there is reason to doubt whether it will reap the intended benefits. The interventions can be effective when the designated intervener is driven by authentic dissent to the emerging hypothesis, but are ineffective when the dissent is contrived through techniques such as role-playing.[233] Genuine separation is difficult to generate, especially when the personnel designated to propound opposing views actually share the same viewpoints. In the context of criminal investigations, the designated dissenters will typically come from the ranks of the same agency as their counterparts, and they are likely to have undergone similar training and to hold similar attitudes toward issues of law and order. A failure of separation appears to have been the case in the study with the Dutch police crime analysts, whose function is to serve the role of devil's advocates. Recall that the critical contribution of those analysts was undermined by their tendency to confirm the investigative team's prevailing theory, while ignoring plausible alternative hypotheses.[234] Similar problems seem to be limiting the effectiveness of magistrates in the French legal system.[235]

Even if successful, functional separation might not endure. Dissenters are generally disliked,[236] and that could render them less influential in the long run. Moreover, genuine psychological separation is bound to be accompanied by the pathological features of intergroup conflict, which could encumber the investigative process and obstruct meritorious investigations. Furthermore, failed interventions might even backfire. Acknowledging countertheories and summarily refuting them might provide decision makers with a hollow sense of having addressed all objections, and thus result in heightened confidence in the correctness of their conclusions.[237] That appears to have happened in the Brandon Mayfield investigation. Recall that the expert appointed by the court arrived at the same erroneous conclusion as did his FBI colleagues, thus bolstering the FBI's blunder.

Still, with careful design and in the correct cultural climate, functional separation can operate successfully. The office of the District Attorney of Dallas County, under the stewardship of Craig Watkins, is a promising example. Watkins has brought the question of accuracy to the fore, and

in 2007 he established an internal Conviction Integrity Unit. This unit is designed to review and re-investigate legitimate claims of innocence by convicted inmates.[238] Within four years of its operation, fourteen convicted inmates have been exonerated.[239] Other jurisdictions have established innocence commissions, which are quasi-judicial entities designed to reexamine convictions of inmates who can show a plausible case of innocence.[240] Functional separation would be most effective and beneficial if it prevented innocent people from being charged and convicted in the first place.

Restructuring the Investigative Authorities

The most ambitious way to defuse the quasi-adversarial nature of investigations is to revamp the institutional incentives and motivations under which investigators operate. Most importantly, eliminating the goal of clearing crimes should go a long way to reduce the institutional pressures on them to reach conclusions of guilt. This reform would replace the current incentive structure with one driven by the goal of truth-seeking. To achieve this, one could imagine transferring the investigative responsibilities from the police to an authority that is not directly responsible for fighting crime. A good candidate would be a judicial branch body established for this purpose.[241] Investigations would be overseen by specially trained judges and conducted by professional investigators. Criminal investigations would be conducted much like investigations of aviation accidents conducted under the authority of the Federal Aviation Authority. Investigative reports would contain a full exposition of both inculpating and exculpating evidence, and would be shared with the prosecution and the defense.

This proposed reform would require a sweeping overhaul of large bureaucratic entities. As discussed in Chapter 1, deep reforms of this nature ought to be considered seriously. The recommendations offered in this book, however, focus on changes that are feasibly implementable in the short and medium term, and that can be adopted even at the departmental or personal level.

Transparency: Electronic Recording of Investigations

The most promising and feasible avenue for enhancing the objectivity of criminal investigations is to make them transparent. This is one of the

two single most important recommendations proposed in this book. All interactions with would-be witnesses—including all lineups, interviews, and interrogations—should be recorded and preserved in their entirety. Meticulous records should be made also of other investigative procedures, especially forensic testing. The recording should be made in the best available medium, which would normally be audiovisual. The recordings should also include unfruitful investigative efforts, even if they are not used in court, such as interviews with witnesses whose statements do not support the prosecution. Importantly, the record should be made available to all parties involved in the case.

As discussed in the following chapters, the creation of a complete and reliable record of investigations is bound to soften their quasi-adversarial bite and enhance the twin objectives of accuracy and transparency. In addition to improving the quality of evidence consumed by the entire criminal justice process, transparent investigations are expected also to improve the investigative process itself. The availability of a record would increase investigators' sense of accountability for their work. The awareness that their performance will be exposed to the critical eye of other actors should make investigators think harder when deciding which hypotheses to generate, which information to test, how to collect that information, and how to evaluate it. Transparency would help ensure that investigators adhere to best practices by providing law enforcement agencies with a tool for training, oversight, and quality assurance. Transparency should also help deter police misconduct.[242] Furthermore, recording investigations is bound to serve as an information-gathering tool. A complete and accurate record is bound to capture forensic details that would otherwise go unnoticed or be forgotten.[243]

The following recommendations seek to further promote more accurate and transparent evidence, and to diminish the adversarial pull.

1. Investigative departments should professionalize and systematize their investigative procedures. The procedures should be based on best-practice protocols.
2. Investigative departments should create full recordings of their investigative processes.
3. Investigative departments should encourage and reward open-minded thinking, and appoint personnel who demonstrate a temperament that is suited for the complexities of the task.

4. Investigative departments should promote greater sensitivity toward the possibility of error.
5. Investigative mistakes should be debriefed candidly.

Chapters 3–5 will offer specific recommendations for conducting best practices with respect to the most common types of evidence used in criminal trials.

NOTES

1. Introduction

1. For a detailed account of the Rose case, see Rutenberg, S. (2006). Anatomy of a miscarriage of justice: The wrongful conviction of Peter J. Rose. *Golden Gate University Law Review, 37,* 7–37; Innocence Project, profile, Peter Rose. http://www.innocenceproject.org/Content/Peter_Rose.php.

2. On the Godschalk case, see Kreimer, S. F., & Rudovsky, D. (2002). Double helix, double bind: Factual innocence and postconviction DNA testing. *University of Pennsylvania Law Review, 151,* 547–617; Innocence Project, profile, Bruce Godschalk. http://www.innocenceproject.org/Content/Bruce_Godschalk.php.

3. *Commonwealth of Pennsylvania v. Bruce Godschalk,* 00934-87, Montgomery County, Jury Trial, May 27, 1987, pp. 138–139.

4. Junkin, T. (2004). *Bloodsworth: The true story of the first death row inmate exonerated by DNA*. Chapel Hill, NC: Algonquin Books; Dwyer, J., Neufeld, P., & Scheck, B. (2000). *Actual innocence: Five days to execution and other dispatches from the wrongfully convicted,* pp. 213–222. New York: Doubleday. See also Department of Justice (1996). Convicted by juries, exonerated by science: Case studies in the use of DNA evidence to establish innocence after trial. http://www.ncjrs.gov/pdffiles/dnaevid.pdf.

5. For early critiques of the process, see Borchard, E. M. (1932). *Convicting the innocent.* Garden City, NY: Garden City Publishing; Frank, J., & Frank, B. (1957). *Not guilty.* Garden City, NY: Doubleday.

6. As the large majority of the crimes discussed in this book are perpetrated by males, the text will normally adopt that gender.

7. It is estimated that only 48 percent of violent crimes and 38 percent of property crimes committed in the United States are reported to the police. Department of Justice, Bureau of Justice Statistics (data for 2006/2007). http://bjs.ojp.usdoj.gov/content/glance/tables/reportingtypetab.cfm.

8. The clearance rates are 45 percent for violent crime and 17 percent for property crime. The rates are 64 percent for murder, 55 percent for aggravated assault, 40 percent for rape, and 27 percent for robbery. Federal Bureau of Investigation (2008). *Uniform crime reporting handbook, 2008: Crime in the United* States, table 25. http://www2.fbi.gov/ucr/cius2008/data/table_25.html.

The data are similar in the United Kingdom, where about one-half of serious crimes are reported, of which only one-fifth are prosecuted. Crown Prosecution Service (2002). *Narrowing the justice gap.* http://www.cps.gov.uk/publications/prosecution/justicegap.html.

9. Failures to punish guilty individuals can also be caused by legal rules that exclude reliable inculpating evidence or thwart the prosecution in the first place. See, e.g., Pizzi, W. T. (1999). *Trials without truth: Why our system of criminal trials has become an expensive failure and what we need to do to rebuild it.* New York: NYU Press.

10. For recent literature on the topic, see Gross, S. R. (2008). Convicting the innocent. *Annual Review of Law & Social Science, 4*, 173–192; Garrett, B. L. (2011). *Convicting the innocent: Where criminal prosecutions go wrong*. Cambridge, MA: Harvard University Press; Westervelt, S. D., & Humphrey, J. A. (2002). *Wrongfully convicted: Perspectives on failed justice*. Piscataway, NJ: Rutgers University Press; Gould, J. B. (2008). *The Innocence Commission: Preventing wrongful convictions and restoring the criminal justice system*. New York: NYU Press; Marshall, L. C. (2004). The innocence revolution and the death penalty. *Ohio State Journal of Criminal Law, 1,* 573–584 (p. 573).

11. See Gross (2008), *supra* note 10; Garrett (2011), *supra* note 10; Marshall (2004), *supra* note 10.

12. See Justice Scalia's concurring opinion in *Kansas v. Marsh,* 548 U.S. 163, 193, 200 (2006). Justice Scalia relies on an appealingly simple mathematical calculation proposed by Joshua Marquis: dividing the number of known exonerees by the number of people incarcerated across the country yields a quotient 0.027 of 1 percent (or 0.00027), which is indeed a small number. Marquis, J. (2005). Myth of innocence. *Journal of Criminal Law & Criminology, 95,* 501–522. See also Hoffman, M. B. (2007). The myth of factual innocence. *Chicago-Kent Law Review, 82,* 663–690.

This mathematical computation is flawed because the rate of false convictions (as opposed to exonerations) is obscure, and because the denominator should be limited to cases in which false convictions are realistically possible and the detection of error is feasible. While both the numerator and denominator resist accurate quantification, under any realistic assumptions they are bound to lead to an error rate that is dramatically larger than the values proposed by Marquis and Scalia. Sam Gross argues convincingly that the rate that Justice Scalia advocates is "flat wrong and badly misleading." Gross, S. R. (2006).

Souter passant, Scalia rampant: Combat in the marsh. *Michigan Law Review First Impressions, 105,* 67–72 (p. 69).

13. Innocence Project, http://www.innocenceproject.org/.

14. As there is no official record of exonerations, their exact number is unknown. Data compiled by Samuel Gross and his colleagues from the University of Michigan indicate that there were some 340 known cases of individual exonerations between 1989 and 2003. Since 2000, the rate of exonerations has been about 40 cases per year. Fewer than half of the cases were overturned through DNA testing. The remainder of the exonerations were spurred by factual findings made by means of conventional types of evidence. Gross, S. R., Jacoby, K., Matheson, D. J., Montgomery, N., & Patil, S. (2005). Exonerations in the United States 1989 through 2003. *Journal of Criminal Law & Criminology, 95,* 523–560.

15. Sam Gross and colleagues have found an estimated error rate of about 4 percent among death row inmates. This rate was estimated for inmates who were sentenced to death between 1973 and 2004 and who remained under threat of execution for up to twenty-one years. Gross, S. R., O'Brien, B., Hu, C., & Kennedy, E. H. (under review). The rate of false convictions among criminal defendants who are sentenced to death. Michael Risinger examined the rate of DNA exonerations in the category of capital rape-murders and found a minimal error rate of 3.3 percent, with a likely ceiling of 5 percent. Risinger, D. M. (2007). Innocents convicted: An empirically justified factual wrongful conviction rate. *Journal of Criminal Law & Criminology, 97,* 761–806. These errors are mostly taken from within the select one-third of capital verdicts and sentences that survived criminal appeals. Gelman, A., Liebman, J. S., West, V., & Kiss, A. (2004). A broken system: The persistent patterns of reversals of death sentences in the United States. *Journal of Empirical Legal Studies, 1*, 209–261. For a survey of these and other studies, see Zalman, M. (in progress). Qualitatively estimating the incidence of wrongful convictions—a postscript.

16. While the crimes of rape and murder account for fewer than 2 percent of the total felony convictions, they make up for 96 percent of the exonerations. See Gross et al. (2005), *supra* note 14. For all practical purposes, the probability of discovering mistaken convictions in other crimes is very slim.

17. It is clear that many innocent people plead guilty rather than go to trial. For example, 135 people most of whom were innocent pled guilty in the Rampart scandal in Los Angeles and in the infamous case of Tulia, Texas. See Burcham, D. W., & Fisk, C. L. (2001). Symposium: The Rampart scandal: Introduction. *Loyola of Los Angeles Law Review, 34,* 537–543; Open Society Policy Center (2005). Tulia: Tip of the drug war iceberg. http://www.soros.org/resources/articles_publications/publications/tulia_20050101/tulia.pdf.

Overturning a conviction is close to impossible for inmates who were convicted based on their pleas. These inmates are at a substantial disadvantage when it comes to tapping into the limited legal assistance resources, convincing

prosecutors, judges, and even defense attorneys to entertain their claim of innocence. Most important, in many states these inmates are barred from testing the evidence that could exonerate them.

18. The average time it takes from conviction to exoneration is more than ten years. See Gross et al. (2005), *supra* note 14.

19. As discussed in "Identity Cases and Culpability Cases," exonerations are all but impossible in culpability cases. Almost all the false convictions that have come to light were in identity cases, that is, cases in which the wrong person was convicted.

20. It is estimated that with the exception of rape crimes, biological evidence is available in only a limited subset of cases, around 10–15 percent of serious felonies. Liptak, A. (2007). Study of wrongful convictions raises questions beyond DNA. *New York Times,* July 23. Quoting Peter Neufeld. http://select.nytimes.com/2007/07/23/us/23bar.html?_r=1.

21. For illustration, in Dallas County, Texas, the standard procedure has been to preserve evidence post-trial, whereas Harris County, Texas has historically destroyed the evidence. As of November 2011, Dallas County has exonerated twenty-two people on the basis of DNA tests, whereas only eight people have been exonerated in Harris County, which has almost double the population. In Virginia, eight innocent convicts have been exonerated on the basis of biological evidence preserved by the late Mary Jane Burton, who worked in the state's crime lab. Burton's habit of preserving evidence was contrary to the laboratory's policies. See Associated Press (2011). Man exonerated in 1979 Newport News rape. April 13. http://hamptonroads.com/2011/04/man-exonerated-1979-newport-news-rape.

22. Some exonerations were based on tests of physical evidence that was supposed to have been destroyed. Dwayne Dail of North Carolina was convicted for raping a twelve-year-old girl and was sentenced to more than two life sentences. Eighteen years into his sentence, Dail was told that the evidence from his trial had long ago been destroyed. Dail's lawyer, Christine Mumma, discovered that the since-deceased police detective had kept the victim's nightgown in a private storage unit. A DNA test of the gown excluded Dail and inculpated a convicted inmate. Mumma, C. (2009). Wrongfully convicted: One lawyer's perspective. *NIJ Journal* no. 262 (March). http://www.nij.gov/journals/262/one-lawyers-tale.htm.

In the abovementioned case of Pete Rose, the bulk of the evidence used to convict Rose was destroyed, but one sample of the semen had been left by mistake at a laboratory in Berkeley for nearly ten years. Rose was set free on the basis of a DNA test of that sample. Rutenberg (2006), *supra* note 1. In the case of Kevin Byrd, the exculpating evidence was preserved most likely as a result of a clerical error. Byrd served twelve years of his life sentence. Innocence Project, profile, Kevin Byrd. http://www.innocenceproject.org/Content/Kevin_Byrd.php.

In a Nebraska case, six innocent people were exonerated thanks to a police sergeant who preserved the biological evidence from another suspect (the real perpetrator) for twenty-three years. DNA tests clear 6 in 1985 slaying: Group's guilty pleas coerced, attorney general says (2008). KETV7.com, November 7. http://www.ketv.com/news/17936340/detail.html. Likewise, the biological evidence that exonerated Ronald Cotton (discussed in Chapters 2–4) was preserved only thanks to the personal initiative of Detective Mark Gauldin.

23. Alan Newton's exoneration for rape, assault, and robbery came twelve years after his first motion to conduct the testing. Dwyer, J. (2006). 22 years after wrongful conviction—and after 12 years fighting for access to evidence—DNA proves Alan Newton's innocence. *New York Times,* June 6. Likewise, in the case of Anthony Capozzi, Erie County Medical Center recovered the biological evidence fifteen years after his first request, following multiple subpoenas issued by the county's district attorney. Capozzi was exonerated after spending twenty-two years in a New York prison for two rapes. Staba, D. (2007). Located in hospital, DNA clears Buffalo man convicted in '80s rapes. *New York Times,* March 29. http://www.nytimes.com/2007/03/29/nyregion/29bike.html.

24. Good fortune was critical in the exoneration of Clarence Elkins, who was sentenced to life for murdering and raping his mother-in-law and raping his six-year-old niece. Elkins's requests for DNA testing were rebuffed by the Ohio courts. Coincidentally, Elkins was placed in the same prison block with the inmate who he suspected had perpetrated the crime. Surreptitiously, Elkins obtained a cigarette butt from the inmate and arranged for it to be tested. The DNA test inculpated that man and set Elkins free. He was released after serving seven and a half years of his sentence. Innocence Project, profile, Clarence Elkins. http://www.innocenceproject.org/Content/Clarence_Elkins.php.

James Curtis Giles became the suspect in a rape case based on a tip from an informant. He was subsequently identified by a rape victim in a photographic lineup, even though he was a decade older and much heavier than the man she had described to the police. Giles was given a DNA test only after the informant admitted that he had intended to inform on a different man by the name of James Giles, who turned out to be James Earl Giles. Giles was exonerated on the basis of a DNA test, while on parole, some twenty-five years after his conviction for a brutal rape. Bustillo, M. (2007). Texas men's innocence puts a county on trial. *Los Angeles Times,* April 9. http://www.latimes.com/news/nationworld/nation/la-na-exonerate9apr09,1,265991.story.

Roy Brown of New York was convicted for sexual assault and murder, and was sentenced to 25 years to life imprisonment. Due to a fire at the home of his step-father, Brown requested copies of the police reports and was accidentally given previously undisclosed reports that implicated the true perpetrator. Brown was ultimately exonerated based on a DNA test of the deceased suspect's daughter. Santos, F. (2006). With DNA from exhumed body, man finally wins

freedom. *New York Times,* December 24. http://www.nytimes.com/2007/01/24/nyregion/24brown.html; Innocence Project, profile, Roy Brown. http://www.innocenceproject.org/Content/Roy_Brown.php.

25. Some of the exonerations in cases that did not have DNA evidence were precipitated by very uncommon circumstances. For example, the innocence of an Illinois man convicted for murdering his parents came to light only because the actual perpetrators were caught bragging about murdering the couple while under surveillance in an unrelated investigation. On the case of Gary Gauger, see Warden, R. (2005). Illinois death penalty reform: How it happened, what it promises. *Journal of Criminal Law & Criminology, 95,* 381–426. After Anthony Porter's family began making funeral arrangements in anticipation of his execution, journalism students from Northwestern University managed to disprove his guilt and obtain a taped confession from the true culprit. See Warden (2005), p. 423. The innocence of two Illinois death row inmates was revealed only because a state attorney happened to recall that the actual perpetrator (who was also the prosecution star witness) had confessed committing the murder to him when they were co-workers at a summer job some years prior. On the case of Perry Cobb and Darby Tillis, see Warden (2005), pp. 412–413.

26. See discussion of such cases in the section "Lineups in the Wild" in Chapter 3.

27. Since much of human behavior is multidetermined, it is impossible to pinpoint a single, precise cause of any given error. Still, one can make rough distinctions between internal stochastic errors and ones that are triggered or exacerbated by a specific type of situation or an input from another person.

28. For example, one prosecutor reflected on his prosecution of an innocent man exonerated after serving twenty-four years in prison: "His name got up in a lineup, and she picked him out. It just turned out to be the wrong man." Another prosecutor commented on her prosecution of a man exonerated after serving sixteen of a forty year sentence: "the police thought they had the right man. And the victim thought she had the right man, and they were wrong." Interviews with Mike O'Connor and Lana Myers, former prosecutors in Dallas County, Texas, commenting on the false convictions of James Curtis Giles and Willy Fountain. Council, J. (2008). Witnesses to the prosecution. *Texas Lawyer,* June 9.

29. The victim agreed to identify Rose only after a protracted and testy exchange with the detectives, and the bystander witness followed suit. Rutenberg (2006), *supra* note 1; Innocence Project, profile, Peter Rose, *supra* note 1.

30. See Kreimer & Rudovsky (2002), *supra* note 2; Innocence Project, profile, Bruce Godschalk, *supra* note 2.

31. See Junkin (2004), *supra* note 4, chap. 12. Two of the witnesses were intoxicated when they saw the perpetrator, and identified Bloodsworth only after seeing him being paraded by the police on television.

32. Dwyer, Neufeld, & Scheck (2000), *supra* note 4, pp. 45–77; Innocence Project, profile, Walter Snyder. http://www.innocenceproject.org/Content/Walter_Snyder.php.

33. See Gould, J. B. (2008), *supra* note 10.

34. For an excellent work of investigative reporting, see Zerwick, P. (2007). Murder, race, justice: The state vs. Darryl Hunt. *Winston-Salem Journal,* November 16; Vertuno, J. (2009). Judge clears dead Texas man of rape conviction. *Austin American-Statesman,* February 7. The Hunt case was also the subject of a compelling documentary film; see Brown, K., Rexer, W., Stern R., & Sundberg, A. (Producers), & Stern, R., & Sundberg, A. (Directors). (2006). *The trials of Darryl Hunt* [Motion picture]. United States: Break Thru Films. See also Innocence Project, profile, Darryl Hunt. http://www.innocenceproject.org/Content/Darryl_Hunt.php.

35. See Castelle, G., & Loftus, E. F. (2002). Misinformation and wrongful convictions. In S. D. Westervelt & J. A. Humphrey, eds., *Wrongfully convicted: Perspectives on failed justice,* pp. 17–35. New Brunswick, NJ: Rutgers University Press; Innocence Project, profile, William O'Dell Harris. http://www.innocenceproject.org/Content/William_ODell_Harris.php.

36. For more examples, see cases listed in the section "Lineups in the Wild" in Chapter 3.

37. As mentioned above, Rose was identified confidently by two witnesses (Rutenberg 2006, *supra* note 1). Bruce Godschalk was convicted on the basis of an identification by a victim, testimony of a second victim, testimony of a jailhouse informant, forensic evidence of a blood test, plus his own confession (Innocence Project, *supra* note 2). The capital prosecution of Kirk Bloodsworth included identifications by five eyewitnesses, a shoe impression, and a putatively incriminating statement made by the defendant, all leading the prosecutor to describe the evidence as being "extremely strong" (Dwyer, Neufeld, & Scheck 2000, *supra* note 4, p. 222). Ronald Cotton, whose case is discussed in the following chapters, was convicted on the basis of identifications by two victims and a bystander, testimony of his employer, and physical evidence.

38. Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science, *Science, 309,* 892–895. A similar analysis of the first 225 DNA exonerations done by the Innocence Project yielded a total of 376 percent; see http://www.innocenceproject.org/understands/. An analysis of the first 250 DNA exonerations indicates that invalid forensic testimony was present in about 60 percent of the cases that contained forensic evidence (80 out of 137 cases in which the trial transcripts are available), which amounts to about one-third of this sample of exonerations. Garrett, B. L., & Neufeld, P. J. (2009). Invalid forensic science testimony and wrongful convictions. *Virginia Law Review, 95,* 1–97.

39. As observed by Saks and Koehler (2005, *supra* note 38), false convictions are caused also by nonevidential factors: 44 percent of the cases involved police

misconduct, 28 percent involved prosecutorial misconduct, and 19 percent involved incompetent legal representation.

40. Julius Ruffin was convicted by a Virginia jury on rape and burglary charges. He was exonerated on the basis of a DNA test after serving twenty-one years in prison. Gould (2008), *supra* note 10.

41. See, e.g., Uviller, H. R. (1990). Acquitting the guilty: Two case studies of jury misgivings and the misunderstood standard of proof. *Criminal Law Forum, 2,* 1–43; Rosen, J. (1998). After "One Angry Woman." *University of Chicago Legal Forum,* 179–195.

42. In many instances, litigators and other courtroom observers will not risk predicting the jury's decision.

43. National Research Council (2004). *Fairness and effectiveness in policing: The evidence.* Ed. W. Skogan & K. Frydl, pp. 74, 227–228. Washington, DC: National Academies Press.

44. One type of systemic differences among types of people in relation to the criminal justice process is demonstrated in the research by Dan Kahan and his colleagues. See Kahan, D. M. (2010). Culture, cognition, and consent: Who perceives what, and why, in "Acquaintance Rape" cases. *University of Pennsylvania Law Review, 158*, 729–813; Kahan, D. M., & Braman, D. (2008). The self-defensive cognition of self-defense. *American Criminal Law Review, 45*, 1–65.

45. The rate of plea bargaining is high even for serious felonies such as rape (88 percent), robbery (89 percent), and aggravated assault (92 percent). Even for murder, almost two-thirds of convictions (61 percent) are obtained by plea bargain. Rosenmerkel, S., Durose, M., & Farole, D. (2009). Felony sentences in state courts, 2006—statistical tables. *U.S. Department of Justice, Bureau of Justice Statistics,* table 4.1. http://bjs.ojp.usdoj.gov/content/pub/pdf/fssc06st.pdf.

46. For critiques of the practice, see Alschuler, A. W. (1976). The trial judge's role in plea bargaining. *Columbia Law Review, 76,* 1059–1154; Stuntz, W. J. (2004). Plea bargaining and criminal law's disappearing shadow. *Harvard Law Review, 117,* 2548–2569. Cf. Church, T. W. (1979). In defense of bargain justice. *Law & Society Review, 13,* 509–525. On the heightened risk that plea bargaining poses for innocent defendants, see Alschuler, A. W. (2003). Straining at gnats and swallowing camels: The selective morality of professor Bibas, *Cornell Law Review, 88,* 1412–1424.

47. See Gross et al. (2005), *supra* note 14.

48. The Innocence Project identified police misconduct in thirty-seven of the first seventy-four cases. The misconduct included suppression of exculpatory evidence, undue suggestiveness, evidence fabrication, coercion of witnesses, and coercion of confessions; http://innocenceproject.org/understand/Government-Misconduct.php. As observed by Saks and Koehler (2005, *supra* note 38), police misconduct was present in 44 percent of the first eighty-six DNA exoneration cases.

49. Prosecutorial misconduct was identified in thirty-three of the first seventy-four DNA exoneration cases. The misconduct included suppression of exculpatory evidence, knowing use of false testimony, coercion of witnesses, improper closing arguments, false statements to jury, and fabrication of evidence; http://innocenceproject.org/understand/Government-Misconduct.php. Saks and Koehler (2005, *supra* note 38) identified prosecutorial misconduct in 28 percent of the first eighty-six DNA exoneration cases.

50. According to Saks & Koehler (2005, *supra* note 38), 27 percent of the first eighty-six DNA exoneration cases involved false or misleading scientific testimony. For disconcerting abuses of scientific testimony, see Garrett, B. L., & Neufeld, P. J. (2009). Invalid forensic science testimony and wrongful convictions. *Virginia Law Review, 95,* 1–97; Giannelli, P. C. (1997). The abuse of scientific evidence in criminal cases: The need for independent crime laboratories. *Virginia Journal of Social Policy & the Law, 4,* 439–478; Mills, S., McRoberts, F., & Possley, M. (2004). Forensics under the microscope—When labs falter, defendants pay. *Chicago Tribune,* October 20.

51. See Federal Rules of Evidence 403, 404(a), and 404(b); and Strong, J. W., ed. (1999). *McCormick on evidence* (5th ed.). St. Paul, MN: West Group.

52. Stuntz, W. (2011). *The collapse of American criminal justice*. Cambridge, MA: Harvard University Press.

53. The work of public defenders and appointed counsel is hindered by both excessive caseloads and limited compensation. For illustration, the estimated average hourly rate of court-appointed attorneys in noncapital felony cases ranges from $50 to $65, which is about one-fourth of the hourly rate of lawyers in private practice. Public funding for investigation is all but nonexistent. The Constitution Project (2009). Justice denied: America's continuing neglect of our constitutional right to counsel. http://www.constitutionproject.org/pdf/139.pdf.

54. For a good account on how law works in parts of the country, see Bach, A. (2009). *Ordinary injustice: How America holds court*. New York: Henry Holt.

55. For this body of research, there is less reason for concern over internal validity and construct validity, both of which speak to the extent to which the observations support the stated conclusions. See Aronson, E., Wilson, T. D., & Brewer, M. B. (1998). Experimentation in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey, eds., *The handbook of social psychology,* vol. 1 (4th ed.), pp. 99–142. New York: McGraw-Hill.

56. See, e.g., Lewin, K. (1935). *A dynamic theory of personality.* New York: McGraw-Hill; Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology.* New York: McGraw-Hill.

57. See Simon, D. (2010). In praise of pedantic eclecticism: Pitfalls and opportunities in the psychology of judging. In D. E. Klein & G. Mitchell, eds., *The psychology of judicial decision making,* pp. 131–147. New York: Oxford University Press.

58. See, e.g., McCloskey, M., Egeth, H., & McKenna, J. (1986). The experimental psychologist in court: The ethics of expert testimony. *Law and Human Behavior, 10,* 1–13; Konecni, V. J., & Ebbesen, E. B. (1986). Courtroom testimony by psychologists on eyewitness identification issues: Critical notes and reflections. *Law and Human Behavior, 10,* 117–126; Yuille, J. C., & Cutshall, J. L. (1986). A case study of eyewitness memory of a crime. *Journal of Applied Psychology, 71,* 291–301.

For an uncharitable treatment of the research on the exclusion of jurors from death penalty panels ("death qualification"), see Chief Justice Rehnquist's opinion in *Lockhart v. McCree,* 476 U.S. 162 (1986). For a response, see Ellsworth, P. C. (1991). To tell what we know or wait for Godot? *Law and Human Behavior, 15(1),* 77–90.

59. See Diamond, S. S. (1997). Illuminations and shadows from jury simulations. *Law and Human Behavior, 21,* 561–571; Bornstein, B. H. (1999). The ecological validity of jury simulations: Is the jury still out? *Law and Human Behavior, 23,* 75–91; Simon (2010), *supra* note 57.

60. On the construct of convergent validity, see Aronson, Wilson, & Brewer (1998), *supra* note 55.

61. On the construct validity of legal psychological research, see Simon (2010), *supra* note 57.

62. Whether an experiment will overstate or understate the results will depend on the manner in which the controlled factors would have interacted with the focal factor in a natural setting. Controlling factors that would otherwise moderate the focal factor are bound to result in the overstatement of the finding, whereas controlling factors that would otherwise exacerbate it are likely to understate the finding. For illustration, a finding that individual jurors are susceptible to a certain bias might be said to exaggerate the problem for the criminal justice system because, in real life, that bias might be corrected by jury deliberation. On the other hand, the deliberation itself might exacerbate the bias, which would mean that the focal finding actually understates the problem.

63. As discussed in Chapter 7, the institutional context of legal adjudication is hardly a guarantee for the accuracy of the process.

64. According to the conventions of experimental psychology, a finding is deemed statistically significant based on the probability that the effect attributed to the experimental treatment was not caused by chance (typically, using a threshold criterion of 0.05). In itself, statistical significance does not distinguish between weak effects (for example, an increase in the rate of an error from 24 percent to 29 percent) and strong ones (an increase in the rate of the error from 24 percent to 60 percent). Nor does it speak to the absolute levels of the observed phenomena, namely, whether the treatment results in an error rate of 30 percent (up from 20 percent) or of 90 percent (up from 80 percent).

65. Lloyd Weinreb proposed the establishment of an "investigating magistracy." Weinreb, L. L. (1977). *Denial of justice*. New York: Free Press, p. 119. George Thomas proposed that criminal investigations and pretrial procedures be overseen by a "screening magistrate": Thomas, G. C., III (2008). *The Supreme Court on trial: How the American justice system sacrifices innocent defendants*. Ann Arbor: University of Michigan Press, pp. 193–227. Along similar lines, Keith Findley has suggested a system that blends the strengths of the adversarial and inquisitorial systems. Findley, K. A. (in press). Adversarial inquisitions: Rethinking the search for the truth. *New York Law Review*.

On the differences between Anglo-American and continental criminal justice systems, see Hatchard, J., Huber, B., & Vogler, R. (1996). *Comparative criminal procedure*. London: British Institute of International and Comparative Law; van Koppen, P. J., & Penrod, D. S. (2003). *Adversarial versus inquisitorial justice: Psychological perspectives on criminal justice systems*. New York: Kluwer Academic Publishing.

66. On the prevailing aversion toward inquisitorial systems, see Sklansky, D. A. (2009). Anti-inquisitorialism. *Harvard Law Review, 122,* 1634–1704.

67. It should be acknowledged that the inquisitorial system is no panacea, as its lofty goals of seeking the truth are not entirely immune to the harsh reality of criminal investigations. Jacqueline Hodgson has observed that the French system does not routinely meet the ideal of the inquisitorial model. Some 95 percent of crimes are investigated not by the magistrate *(juge d'instruction),* but by the regular police force under the supervision of a prosecutor *(procureur)*. The latter process functions much like the one in the Anglo-American model, with perhaps weaker protection for suspects' rights. Hodgson, J. (2005). *French criminal justice: A comparative account of the investigation and prosecution of crime in France*. Oxford: Hart Publishing.

68. A small number of suggestions might require legislative intervention.

69. Forty-two percent of the first 250 exonerations resulted in positive identifications of the true perpetrators. Innocence Project (2010). 250 exonerated, too many wrongfully convicted. http://www.innocenceproject.org/news/250.php.

70. On the importance of protecting only innocent defendants, see Amar, A. R. (1997). *The constitutional and criminal procedures: First principles,* chap. 4. New Haven, CT: Yale University Press.

71. For an example of this type of tradeoff, see "Recommendations for Reform" in Chapter 3.

72. Designing reform should take into consideration the potential for unintended and unwanted consequences. As noted by Carol and Jordan Steiker, arguments based on actual innocence can be used as a double-edged sword, and they have been deployed successfully by the U.S. Supreme Court and by Congress to justify policies that deprive defendants of a fair review of their cases. Steiker, C. S., & Steiker, J. M. (2005). The seduction of innocence: The attraction and

limitations of the focus on innocence in capital punishment law and advocacy. *Journal of Criminal Law & Criminology, 95,* 587–624.

For a skeptical view of the need for reform, see, e.g., Allen, R. J., & Laudan, L. (2008). Deadly dilemmas. *Texas Tech Law Review, 41,* 65–92.

73. The median time from arrest to adjudication for various felonies including rape, robbery, and assault ranges from four to eight months, and for murder it is about one year. Cohen, T. H., & Kyckelhahn, T. (2010). Felony defendants in large urban counties, 2006. *Department of Justice, Bureau of Justice Statistics,* table 10. http://bjs.ojp.usdoj.gov/content/pub/pdf/fdluc06.pdf. In the cases that actually go to trial, the periods are oftentimes considerably longer.

74. There are exceptions to the superior accuracy of *raw* evidence. On occasion, a witness's original statement could have been mistaken, while the contamination from cowitnesses' statements can actually make his or her testimony more accurate. As a policy matter, planting accurate information in witnesses' testimony must be discouraged. The possible increase in accuracy cannot justify the host of legal, ethical, and practical concerns raised by this prospect.

2. "We're Closing In on Him"

1. The case is the subject of a moving book: Thompson-Cannino, J., Cotton, R., & Torneo, E. (2009). *Picking cotton.* New York: St. Martin's Press. The case was first exposed in a documentary film produced and directed by Ben Loeterman, *What Jennifer Saw*, *Frontline* series, PBS (1997). http://www.pbs.org/wgbh/pages/frontline/shows/dna/. The account of the case in the text is based also on the transcripts of the first trial (*State v. Cotton,* No. 257A85 Alamance Co. Super. Ct., January 7, 1985), and the appellate documents of the second trial (*State v. Cotton,* 318 N.C. 663 [1987], No. 257A85).

2. Although Gauldin's identification procedures do not meet the current best-practice standards, they are not much different from the way procedures are conducted today in many jurisdictions. Unlike many defendants lacking the means to afford effective counsel, Cotton was lucky to have been represented by a court-appointed attorney, Philip Moseley, who appears to have performed proficiently both at trial and on appeal.

3. For example, the interrogation that led to Bruce Godschalk's false confession came on the heels of his identification in a photographic array by one of the victims. Although this woman testified in court that she was absolutely certain in her identification, she had been very hesitant when she picked him out at a photographic array. She picked Godschalk only on the third viewing, each of which lasted twenty to thirty minutes. *Commonwealth of Pennsylvania v. Bruce Godschalk,* 00934-87, Montgomery County, Suppression hearing, May 26, 1987, p. 23.

4. National Research Council (2004). *Fairness and effectiveness in policing: The evidence*. Ed. Wesley Skogan & Kathleen Frydl, pp. 48–49. Washington, DC: National Academies Press. No fewer than 771 departments employ only one police officer. Bear in mind that police departments are also entrusted with noninvestigative tasks, such as enforcing the law (mostly through patrolling), maintaining order, and providing miscellaneous public services.

5. Innes, M. (2003). *Investigating murder: Detective work and the police response to criminal homicide,* p. 127. Oxford: Oxford University Press.

6. National Research Council (2004), *supra* note 4, p. 2; Skolnick, J. (1966). *Justice without trial: Law enforcement in democratic society,* pp. 66–68. New York: Wiley; Waegel, W. B. (1981). Case routinization in investigative police work. *Social Problems, 28,* 263–275.

7. See Brownlie, A. R. (1984). *Crime Investigation, art or science? Patterns in a labyrinth*. Edinburgh: Scottish Academic Press. Much of detectives' training is done informally on the job, effectively by means of apprenticeships with senior agents. Manning, P. K. (2006). Detective work/culture. In J. Greene, ed., *Encyclopedia of police sciences,* p. 394. New York: Routledge.

8. Innes (2003), *supra* note 5.

9. Skolnick (1966), *supra* note 6; Neyroud, P., & Disley, E. (2007). The management, supervision, and oversight of criminal investigations. In T. Newburn, T. Williamson, & A. Wright, eds., *Handbook of criminal investigation,* pp. 549–571. Portland, OR: Willan Publishing.

10. See U.S. Department of Justice, National Institute of Justice (2003). *Factors that influence public opinion of the police.* Washington, DC. http://www.ncjrs.gov/pdffiles1/nij/197925.pdf.

11. Handling the media in high-profile cases has become one of the more onerous tasks facing police investigators. Mawby, R. C. (2007). Criminal investigation and the media. In T. Newburn, T. Williamson, & A. Wright, eds., *Handbook of criminal investigation,* pp. 146–169. Portland, OR: Willan Publishing.

12. Innes (2003), *supra* note 5, p. 127. Investigations that fail to make progress in the "golden hour" stand the risk of contagion of crime scene evidence, deterioration and contamination of witness memory, and greater opportunities for the perpetrator to cover up his leads. Innes, M. (2007). Investigation order and major crime inquiries. In T. Newburn, T. Williamson, & A. Wright, eds., *Handbook of criminal investigation*, pp. 255–276. Portland, OR: Willan Publishing.

13. As observed by Skolnick, abiding by the rule of law can interfere with getting the job done. Skolnick (1966), *supra* note 6, chaps. 9–11. See also National Research Council (2004), *supra* note 4, p. 159.

14. National Research Council (2004), *supra* note 4, p. 3. Since the 1970s, the field of police investigation has been researched only sparsely in the United

States; ibid., p. 23. Much of the recent research has been performed in the United Kingdom.

15. A study of murder investigations conducted in the United Kingdom found that abductive reasoning was by far the most commonly used form of investigative logic. Innes (2003), *supra* note 5, p. 184. On abductive reasoning, see Anderson, T., Schum, D., & Twining, W. (2005). *Analysis of evidence*. 2nd ed. Cambridge: Cambridge University Press.

16. Carson, D. (2009). Detecting, developing and disseminating detectives' "creative" skills. *Policing & Society, 19,* 216–225.

17. Risinger, M. D. (2006). Boxes in boxes: Julian Barnes, Conan Doyle, Sherlock Holmes and the Edalji case. *International Commentary on Evidence, 4(2),* article 3.

18. Spalding, T. L., & Murphy, G. L. (1996). Effects of background knowledge on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 525–538.

19. This effect has been labeled the *explanation bias*. Markman, K. D., & Hirt, E. R. (2002). Social prediction and the "allegiance bias." *Social Cognition, 20,* 58–86. This effect, however, can easily be swamped by motivation. For example, it does not hold up for participants who are fans of either one of the teams.

20. Carroll, J. S. (1978). The effect of imagining an event on expectations for the event: An interpretation in terms of the availability heuristic. *Journal of Experimental Social Psychology, 14,* 88–96.

21. Ross, L. D., Lepper, M. R., Strack, F., & Steinmetz, J. (1977). Social explanation and social expectation: Effects of real and hypothetical explanations on subjective likelihood. *Journal of Personality and Social Psychology, 35,* 817–829. For a review, see Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin, 110,* 499–519. See also Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology, 59,* 601–613.

22. Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Biases and heuristics. *Science, 185,* 1124–1130. For findings of inflated belief in conditional probabilities, see Koriat, A., Fiedler, K., & Bjork, R. A. (2006). Inflation of conditional predictions. *Journal of Experimental Psychology: General, 135,* 429–447.

23. This effect has been labeled *belief perseverance*. Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology, 39,* 1037–1049.

24. Bacon, F. (1620/1960). *The new organon and related writings,* p. 50. New York: Liberal Arts Press.

25. Doyle, A. C. (1891). The adventures of Sherlock Holmes: A scandal in Bohemia. *Strand Magazine,* July 1981, p. 2.

26. For reviews, see Klayman, J. (1995). Varieties of confirmation bias. In J. R. Busemeyer, R. Hastie, & D. L. Medin, eds., *Decision making from the perspective of cognitive psychology,* vol. 32: *The psychology of learning and motivation,* pp. 385–418. New York: Academic Press; Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2,* 175–220.

27. Klayman (1995), *supra* note 26, p. 386.

28. See Revlin, R., Leirer, V., Yopp, H., & Yopp, R. (1980). The belief-bias effect in formal reasoning: The influence of knowledge on logic. *Memory & Cognition, 8,* 584–592.

29. See Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology, 71,* 5–24.

30. Ibid.

31. Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes, 56,* 28–55; Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1,* 161–175.

32. Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology, 44,* 20–33.

33. Hirt, E. R., & Markman, K. D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology, 69,* 1069–1086 (study 3).

34. Attributed to French philosopher Emile Chartier (1868–1961).

35. Cohen, C. E. (1981). Person categories and social perception: Testing some boundaries of the processing effect of prior knowledge. *Journal of Personality and Social Psychology, 40,* 441–452.

36. Edwards & Smith (1996), *supra* note 29.

37. Greenhoot, A. F., Semb, G., Colombo, J., & Schreiber, T. (2004). Prior beliefs and methodological concepts in scientific reasoning. *Applied Cognitive Psychology, 18,* 203–221.

38. Fraser-Mackenzie, P. A. F., & Dror, I. E. (2009). Selective information sampling: Cognitive coherence in evaluation of a novel item. *Judgment and Decision Making, 4,* 307–316.

39. Kempton, J., Alani, A., & Chapman, K. (2002). Potential effects of the confirmation bias in house condition surveys. *Structural Survey, 20,* 6–12.

40. Wallsten, T. S. (1981). Physician and medical student bias in evaluating diagnostic information. *Medical Decision Making, 1,* 145–164. Another study found that 71 percent of medical students and residents converged on diagnoses that were tentatively presented to them, while fewer than 10 percent endorsed

nonsuggested yet plausible alternative ones. Leblanc, V. R., Brooks, L. R., & Norman, G. R. (2002). Believing is seeing: The influence of a diagnostic hypothesis on the interpretation of clinical features. *Academic Medicine, 77(10),* supp. See also Pines, J. M. (2005). Profiles in patient safety: Confirmation bias in emergency medicine. *Academic Emergency Medicine, 13,* 90–94.

41. Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. *Archives of Internal Medicine, 165,* 1493–1499.

42. Psychotherapists described an interviewee presented as a job applicant in neutral terms, such as "attractive and conventional looking" and "candid and innovative." When the same interviewee was presented as a patient, he was described as "dependent, passive-aggressive" and a "tight, defensive person, conflicted over homosexuality." The reported differences were stark for the two groups of analytic therapists, but less so among behavior therapists. Langer, E. J., & Abelson, R. P. (1974). A patient by any other name . . . : Clinician group difference in labeling bias. *Journal of Consulting and Clinical Psychology, 42,* 4–9.

43. Ben-Shakhar, G., Bar-Hilel, M., Bilu, Y., & Shefler, G. (1998). Seek and ye shall find: Test results are what you hypothesize they are. *Journal of Behavioral Decision Making, 11,* 235–249.

44. Ask, K., & Granhag, P. A. (2007a). Motivational bias in criminal investigators' judgments of witness reliability. *Journal of Applied Social Psychology, 37,* 561–591; Ask, K., Rebelius, A., & Granhag, P. A. (2008). The "elasticity" of criminal evidence: A moderator of investigator bias. *Applied Cognitive Psychology, 22,* 1245–1259.

45. No differences were found between seasoned and relatively inexperienced analysts. Kerstholt, J. H., & Eikelbloom, A. R. (2007). Effects of prior interpretation on situation assessment in crime analysis. *Journal of Behavioral Decision Making, 20,* 455–465.

46. Unbeknownst to these experts, they had previously analyzed the same pairs of prints in a real-life case and determined them to be a positive match. The researchers informed them (incorrectly) that the prints were a nonmatch. This misinformation appears to have had a strong impact on all but one of the experts tested. Of the five experts, three changed their judgments to nonmatches, and one changed to "cannot decide." Dror, I. E., Charlton, D., & Péron, A. E. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International, 156,* 74–78.

47. For informative discussions of tunnel vision in police investigations, see Martine, D. L. (2002). The police role in wrongful convictions: An international comparative study. In Saundra D. Westervelt & John A. Humphrey, eds., *Wrongfully convicted: Perspectives on failed justice,* pp. 77–95. New Brunswick, NJ: Rutgers University Press; Findley, K. A., & Scott, M. S. (2006). The multiple dimensions of tunnel vision in criminal cases. *Wisconsin Law Review,* 291–397;

and Risinger, M. D., Saks, M. J., Thompson, W. C., & Rosenthal, R. (2002). The *Daubert/Kumho* implications of observer effects in forensic science: Hidden problems of expectation and suggestion. *California Law Review, 90,* 1–56.

48. This is true for both the United States and the United Kingdom. National Research Council (2004), *supra* note 4, p. 74; Bayley, D. H. (2005). What do the police do? In T. Newburn, ed., *Policing: Key readings,* p. 145. Portland, OR: Willan Publishing; Bayley, D. H. (1994). *Police for the future,* p. 27. New York: Oxford University Press.

49. See, e.g., Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37,* 2098–2109; Edwards & Smith (1996), *supra* note 29.

50. Findings to this effect were made with police personnel in Canada, Australia, France, the United Kingdom, and the United States. See Perrott, S. B., & Taylor, D. M. (1995). Attitudinal differences between police constables and their supervisors: Potential influences of personality, work environment, and occupational role. *Criminal Justice and Behavior, 22,* 326–339; Wortley, R. K., & Homel, R. J. (1995). Police prejudice as a function of training and outgroup contact: A longitudinal investigation. *Law and Human Behavior, 19,* 305–317; Furnham, A., & Alison, L. (1994). Theories of crime, attitudes to punishment and juror bias amongst police, offenders and the general public. *Personality and Individual Differences, 17,* 35–48; Sidanius, J., Liu, J. H., Shaw, J. S., & Pratto, F. (1994). Social dominance orientation, hierarchy attenuators and hierarchy enhancers: Social dominance theory and the criminal justice system. *Journal of Applied Social Psychology, 24,* 338–366. An experiment comparing Swedish experienced criminal investigators to students found that the former were more prone to interpret mixed evidence as inculpating of the focal suspect. Ask, K., & Granhag, P. A. (2005). Motivational sources of confirmation bias in criminal investigations: The need for cognitive closure. *Journal of Investigative Psychology and Offender Profiling, 2,* 43–63.

51. The confirmation bias thus fits the category of *mental contamination;* see Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin, 116,* 117–142.

52. Jay Koehler found that although scientists' evaluations of research were biased by its compatibility with their own beliefs, they denied and deplored any such influence. See Koehler (1993), *supra* note 31.

53. See, e.g., Ask, Rebelius, & Granhag (2008), *supra* note 44; Dror, I. E., & Charlton, D. (2006). Why experts make errors. *Journal of Forensic Identification, 56,* 600–616.

54. See Weinreb, L. L. (1977). *Denial of justice,* chap. 2. New York: Free Press.

55. Cole, S. A. (2005). More than zero: Accounting for error in latent fingerprint identification. *Journal of Criminal Law & Criminology, 95,* 985–1078; Giannelli, P. C. (1997). The abuse of scientific evidence in criminal cases: The need for independent crime laboratories. *Virginia Journal of Social Policy & the Law, 4,* 439–478.

56. Comment on Rule 3.8 of Model Rules of Professional Conduct. American Bar Association. See also *Berger v. United States,* 295 U.S. 78, 88 (1935); Weinreb (1977), *supra* note 54, chap. 3.

57. The theoretical foundation of the research on motivated reasoning was formulated by Ziva Kunda. See Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108,* 480–498, p. 480.

58. Ditto, P. H., Munro, G. D., Apanovitch, A. M., Scepansky, J. A., & Lockhart, L. K. (2003). Spontaneous skepticism: The interplay of motivation and expectation in responses to favorable and unfavorable medical diagnoses. *Personality and Social Psychology Bulletin, 29,* 1120–1132.

59. Wyer, R. S., & Frey, D. (1983). The effects of feedback about self and others on the recall and judgments of feedback-relevant information. *Journal of Experimental Social Psychology, 19,* 540–559.

60. Munro, G. D., Ditto, P. H., Lockhart, L. K., Fagerlin, A., Gready, M., & Peterson, E. (2002). Biased assimilation of sociopolitical arguments: Evaluating the 1996 U.S. presidential debate. *Basic and Applied Social Psychology, 24,* 15–26.

61. Hastorf, A. H., & Cantril, H. (1954). They saw a game: A case study. *Journal of Abnormal and Social Psychology, 49,* 129–134.

62. Boiney, L. G., Kennedy, J., & Nye, P. (1997). Instrumental bias in motivated reasoning: More when more is needed. *Organizational Behavior and Human Decision Processes, 72,* 1–24.

63. Brownstein, A. L., Read, S. J., & Simon, D. (2004). Effects of individual expertise and task importance on pre-decision reevaluation of alternatives. *Personality and Social Psychology Bulletin, 30,* 819–904.

64. See, e.g., Larwood, L., & Whittaker, W. (1977). Managerial myopia: Self-serving biases in organizational planning. *Journal of Applied Psychology, 62,* 194–198; Risucci, D. A., Tortolani, A. J., & Ward, R. J. (1989). Ratings of surgical residents by self, supervisors and peers. *Surgery, Gynecology and Obstetrics, 169(6),* 519–526; Bass, B. M., & Yammarino, F. J. (1991). Congruence of self and others' leadership ratings of naval officers for understanding successful performance. *Applied Psychology, 40,* 437–454.

65. This study used a quasi-criminal setting, in which a university student was being investigated for academic misconduct. Simon, D., Stenstrom, D., & Read, S. J. (2008). *On the Objectivity of Investigations: An Experiment.* Paper presented at Conference for Empirical Legal Studies, Cornell Law School, September 9–10.

66. The adversarial distrust was manifested by the perception of the other investigator as less objective and more distrusting than they viewed themselves. Ibid.

67. Charlton, D., Fraser-Mackenzie, P., & Dror, I. E. (2010). Emotional experiences and motivating factors associated with fingerprint analysis. *Journal of Forensics Sciences, 55,* 385–393.

68. The war metaphor generates an aura of legitimacy and a network of entailments that provide a license for special lines of action. Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

69. Klockers, C. B. (1985). *The idea of police.* Beverly Hills, CA: Sage Publications; Stenross, B., & Kleinman, S. (2003). The highs and lows of emotional labor: Detectives' encounters with criminals and victims. In M. R. Pogrebin, ed., *Qualitative approaches to criminal justice: Perspectives from the field,* pp. 107–115. Thousand Oaks, CA: Sage Publications; Charlton, Fraser-Mackenzie, & Dror (2010), *supra* note 67.

An illustration of the intensity of conviction in the noble cause was provided by Dean Bowman, the prosecutor who obtained the conviction of Darryl Hunt in the second trial: "The most rewarding thing about being a prosecutor is that you . . . know that you're doing the right thing or you're doing your best to do the right thing, and that you have a moral conviction that no matter what the odds are against you, that the truth will come out. You have to just trudge on, and you somehow know it will, and believe it will, and you move toward that. And I think, in doing that, you become very passionate about that, no matter what the obstacles are, if you believe that to be the truth, then it will prevail. And I found that it does." Interview with Dean Bowman, in *The trials of Darryl Hunt* (2005). Think Film, produced and directed by Ricki Stern and Anne Sundberg. Darryl Hunt was exonerated by DNA and released after spending eighteen and a half years in prison. http://www.innocenceproject.org/Content/Darryl_Hunt.php.

70. The official standards of reporting crime data to the FBI require that the suspect be arrested, charged with the crime, and turned over to the courts for prosecution. Federal Bureau of Investigation (2010). *Uniform crime reporting handbook*. http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/methodology. In investigators' day-to-day practical terms, arresting the suspect can suffice. See Skolnick (1966), *supra* note 6, pp. 167–173; Waegel (1981), *supra* note 6.

71. See statement of chief of police Joseph Masten, in Stern & Sundberg (2005), *supra* note 69. Recall that Hunt was exonerated by DNA and released after spending eighteen and a half years in prison.

72. See Wilson, O. W. (1962). *Police planning,* p. 3. Springfield, IL: Charles C. Thomas; Skolnick (1966), *supra* note 6, chap. 8.

73. See Skolnick (1966), *supra* note 6; Waegel (1981), *supra* note 6.

74. In the New York Police Department, low clearance rates have been used to scold and embarrass precinct commanders in front of their peers and subordinates. Rashbaum, William K. (2010). Retired officers raise questions on crime data. *New York Times,* February 6. http://www.nytimes.com/2010/02/07/nyregion/07crime.html?scp=1&sq=Retired%20officers%20raise%20questions%20on%20crime%20data&st=cse.

75. Waegel (1981), *supra* note 6.

76. Rashbaum (2010), *supra* note 74; Baker, A. (2010). Former commander recalls pressures to alter reports. *New York Times,* February 7. http://www.nytimes.com/2010/02/08/nyregion/08captain.html; Rayman, G. (2010). The NYPD tapes: Inside Bed-Stuy's 81st precinct. *The Village Voice*, May 4, 2010; Davies, N. (2003). Fiddling the figures: Police cheats who distort force records. *The Guardian,* July 11. http://www.guardian.co.uk/uk/2003/jul/11/ukcrime.prisonsandprobation1.

77. The study estimated that some 30 percent of cleared crimes were solved on-scene, and 50 percent were solved on the basis of an initial identification made by the patrol officers. Detectives were involved in the remaining 20 percent, though a majority of these cases were solved by information volunteered by witnesses or by mundane tracking of information that could be done by clerical personnel. Petersilia, J. (1977). The investigative function. In P. W. Greenwood, J. M. Chaiken, & J. Petersilia, eds., *The criminal investigation process*. Lexington, MA: D. C. Heath. See also Waegel (1981), *supra* note 6; Stenross & Kleinman (2003), *supra* note 69. See also National Research Council (2004), *supra* note 4, pp. 74, 227–228.

In the United Kingdom, it has been estimated that some 70 percent of homicide cases can be considered "self-solvers." Innes, M. (2002). The "process structures" of police homicide investigations. *British Journal of Criminology, 42,* 669–688. In some jurisdictions in the United States, investigators call easy cases "dunkers." Manning (2006), *supra* note 7.

78. Bayley (2005), *supra* note 48, p. 145; Eck, J. (1992). *Solving crimes: The investigation of burglary and robbery.* Washington, DC: Police Executive Research Foundation.

79. National Research Council (2004), *supra* note 4, p. 74; Tilley, N., Robinson, A., & Burrows, J. (2007). The investigation of high volume crime. In T. Newburn, T. Williamson, & A. Wright, eds., *Handbook of criminal investigation,* pp. 226–254. Portland, OR: Willan Publishing.

80. Innes (2003), *supra* note 5, p. 15.

81. Joseph McCarthy, who prosecuted the Walter Snyder case (discussed in Chapter 4), explained: "And in a rape case, there is often a bonding between the victim and the prosecutor, and the investigator. They are going through a bad time. The psychology is that you're the last line of defense between them and the guy's getting out on the street." Dwyer, J., Neufeld, P., & Scheck, B. (2000). *Actual innocence: Five days to execution and other dispatches from the wrong-*

fully convicted, p. 238. New York: Doubleday. After spending thirteen years on a homicide case, a California homicide detective described a close relationship with the victim's mother: "She's part of my family and I am part of hers." Therolf, G. (2007). A "bitter joy" at murder arrests; After 13 years, Placentia police say new DNA evidence ties two cousins to the stabbing death of a Cal State Fullerton student. *Los Angeles Times,* July 7. http://www.latimes.com/news/local/la-me-torrez7jul07,1,2429489.story?coll=la-headlines-california.

82. On the relationship among these emotional reactions, see Kahneman, D., & Sunstein, C. R. (2005). Cognitive psychology of moral intuitions. In J. P. Changeux, A. R. Damasio, W. Singer, & Y. Christen, eds., *Neurobiology of human values,* pp. 91–105. Berlin: Springer.

83. Lerner, J. S., Goldberg, J. H., & Tetlock, P. E. (1998). Sober second thought: The effects of accountability, anger, and authoritarianism on attributions of responsibility. *Personality and Social Psychology Bulletin, 24,* 563–574; Goldberg, J. H., Lerner, J. S., & Tetlock, P. E. (1999). Rage and reason: The psychology of the intuitive prosecutor. *European Journal of Social Psychology, 29,* 781–795; Quigley, B. M., & Tedeschi, J. T. (1996). Mediating effects of blame attributions on feelings of anger. *Personality and Social Psychology Bulletin, 22,* 1280–1288. Anger was also found to mediate judgments of blame in apportioning responsibility for accidents. Feigenson, N., Park, J., & Salovey, P. (2001). The role of emotions in comparative negligence judgments. *Journal of Applied Social Psychology, 31,* 576–603. Heightened states of anger increase attributions of fault to human conduct rather than to situational conditions. Keltner, D., Ellsworth, P. C., & Edwards, K. (1993). Beyond simple pessimism: Effects of sadness and anger on social perception. *Journal of Personality and Social Psychology, 64,* 740–752.

84. For example, arousal of anger increased participants' tendency to believe an allegation that a Hispanic person behaved violently and that a student athlete cheated on an exam. Bodenhausen, G. V., Sheppard, L. A., & Kramer, G. P. (1994). Negative affect and social judgment: The differential impact of anger and sadness. *European Journal of Social Psychology, 24,* 45–62.

85. Ferguson, T. J., & Rule, B. G. (1983). An attributional perspective on anger and aggression. In R. G. Geen & E. I. Donnerstein, eds., *Aggression: Theoretical and empirical reviews,* vol. 1, pp. 41–74. New York: Academic Press.

86. Mackie, D. M., Devos, Thierry, & Smith E. R. (2000). Intergroup emotions: Explaining offensive action tendencies in an intergroup context. *Journal of Personality and Social Psychology, 79,* 602–616.

87. Dror, I. E., Péron, A. E., Hind, S. L., & Charlton, D. (2005). When emotions get the better of us: The effect of contextual top-down processing on matching fingerprints. *Applied Cognitive Psychology, 19,* 799–809.

88. The arousal of sadness had no such effect. Ask, K., & Granhag, P. A. (2007b). Hot cognition in investigative judgments: The differential influence of anger and sadness. *Law and Human Behavior, 31,* 537–551.

89. Tajfel, H., & Turner, J. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel, eds., *The Social psychology of intergroup relations,* pp. 33–47. Belmont, CA: Wadsworth; Abrams, D., & Hogg, M. A. (1990). *Social identity theory: Constructive and critical advances.* New York: Springer-Verlag; Abrams, D. (1999). Social identity, social cognition, and the self: The flexibility and stability of self-categorization. In D. Abrams, & M. A. Hogg, eds., *Social identity and social cognition,* pp. 197–229. Malden, MA: Blackwell.

90. For a review of experimental work, see Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin, 86,* 307–324. Recent research emphasizes the importance of morality in in-group evaluations. Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology, 93,* 234–249. For anthropological illustrations, see Brewer, M. B., & Campbell, D. T. (1976). *Ethnocentrism and intergroup attitudes: East African evidence.* Beverly Hills, CA: Sage Publications; Phalet, K., & Poppe, E. (1997). Competence and morality dimensions of national and ethnic stereotypes: A study in six eastern-European countries. *European Journal of Social Psychology, 27,* 703–723.

91. As mentioned above, law enforcement personnel tend to hold attitudes associated with the position of *law and order.* See, e.g., Perrott & Taylor (1995), *supra* note 50; Wortley & Homel (1995), *supra* note 50; Furnham & Alison (1994), *supra* note 50; Sidanius et al. (1994), *supra* note 50.

92. White, K. M., Hogg, M. A., & Terry, D. J. (2002). Improving attitude-behavior correspondence through exposure to normative support from a salient ingroup. *Basic and Applied Social Psychology, 24,* 91–103.

93. Back, K. W. (1951). Influence through social communication. *Journal of Abnormal and Social Psychology, 46,* 9–23; Swann, W. B., Jr., Gómez, Á., Seyle, D. C., Morales, J. F., & Huici, C. (2009). Identity fusion: The interplay of personal and social identities in extreme group behavior. *Journal of Personality and Social Psychology, 96,* 995–1011. For a review of members' deference to the group norm, see Roccas, S., Sagiv, L., Schwartz, S., Halevy, N., & Eidelson, R. (2008). Toward a unifying model of identification with groups: Integrating theoretical perspectives. *Personality and Social Psychology Review, 12,* 280–306.

94. Dwyer, Neufeld, & Scheck (2000), *supra* note 81, p. 238. On the case of Walter Snyder, see Chapter 4.

95. Kerschreiter, R., Schulz-Hardt, S., Mojzisch, A., & Frey, D. (2008). Biased information search in homogeneous groups: Confidence as a moderator for the effect of anticipated task requirements. *Personality and Social Psychology Bulletin, 34,* 679–691.

96. Schulz-Hardt, S., Frey, D., Lüthgens, C., & Moscovici, S. (2000). Biased information search in group decision making. *Journal of Personality and Social Psychology, 78,* 655–669.

97. Rydell, R. J., Mackie, D. M., Maitner, A. T., Claypool, H. M., Ryan, M. J., & Smith, E. R. (2008). Arousal, processing, and risk taking: Consequences of intergroup anger. *Personality and Social Psychology Bulletin, 34,* 1141–1152.

98. As described by Irving Janis, groupthink encompasses illusions of invulnerability, collective rationalization, belief in the inherent morality of the group, stereotypes of out-groups, pressure on dissenters, self-censorship, illusions of unanimity, and self-appointed mind-guards. On the basis of historical case studies, Janis demonstrated that groupthink mindsets result in poor decisions. Janis, I. L. (1972). *Victims of groupthink*. Boston: Houghton Mifflin; Janis, I. L. (1982). *Groupthink: Psychological studies of policy decisions and fiascoes,* 2nd ed.. Boston: Houghton Mifflin.

99. The discrepancy between individual and group behavior has been labeled the *discontinuity effect*. Insko, C. A., & Schopler, J. (1998). Differential distrust of groups and individuals. In C. Sedikides, J. Schopler, & C. A. Insko, eds., *Intergroup cognition and intergroup behavior,* pp. 75–107. Mahwah, NJ: Lawrence Erlbaum.

100. See Johnston, K. L., & White, K. M. (2003). Binge-drinking: A test of the role of group norms in the theory of planned behaviour. *Psychology & Health, 18,* 63–77.

101. Jaffe, Y., Shapir, N., & Yinon, Y. (1981). Aggression and its escalation. *Journal of Cross-Cultural Psychology, 12,* 21–36; Jaffe, Y., & Yinon, Y. (1979). Retaliatory aggression in individuals and groups. *European Journal of Social Psychology, 9,* 177–186.

102. Meier, B. P., & Hinsz, V. B. (2004). A comparison of human aggression committed by groups and individuals: An interindividual-intergroup discontinuity. *Journal of Experimental Social Psychology, 40,* 551–559.

103. Jaffe, Shapir, & Yinon (1981), *supra* note 101.

104. Milgram, S. (1974). *Obedience to authority: An experimental view,* experiment 18, pp. 121–122. New York: Harper & Row.

105. Valdesolo, P., & DeSteno, D. (2007). Moral hypocrisy: Social groups and the flexibility of virtue. *Psychological Science, 18,* 689–690. For a real-life example, see Moser, K. (2010). San Francisco DA says office didn't know about problems at scandal-ridden crime lab, *The Recorder,* April 27. http://www.law.com/jsp/article.jsp?id=1202453216864&San_Francisco_DA_Says_Office_Didnt_Know_About_Problems_at_ScandalRidden_Crime_Lab.

106. Baumeister, R. F., & Leary, M. F. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motive. *Psychological Bulletin, 117,* 497–529.

107. For example, people hold in high regard traits such as "creative" and "bright," and disapprove of traits such as "lazy" and "incompetent." Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology, 49,* 1621–1630.

108. For the need to be seen as fair, see Loewenstein, G., Issacharoff, S., Camerer, C., & Babcock, L. (1993). Self-serving assessments of fairness and pretrial bargaining. *Journal of Legal Studies, 22,* 135–159.

109. Aronson, E. (1969). The theory of cognitive dissonance: A current perspective. In L. Berkowitz, ed., *Advances in experimental social psychology,* vol. 4, pp. 1–34. San Diego: Academic Press; Aronson, E. (1992). The return of the repressed: Dissonance theory makes a comeback. *Psychological Inquiry, 3,* 303–311.

110. As suggested by Alicke (1985), *supra* note 107, some of the traits that are associated with good investigative work are also those that people assign themselves most strongly, such as "perceptive," "level headed," and "reliable." See also Aronson (1969), *supra* note 109.

111. See, e.g., Staw, B. M., & Fox, F. V. (1977). Escalation: The determinants of commitment to a chosen course of action. *Human Relations, 30,* 431–450; Garland, H., & Conlon, D. E. (1998). Too close to quit: The role of project completion in maintaining commitment. *Journal of Applied Social Psychology, 28,* 2025–2048.

112. The distortion of one's prior choices bears a resemblance to cognitive dissonance theory. See Festinger, L. (1957). *A theory of cognitive dissonance.* Evanston, IL: Row, Peterson; Harmon-Jones, E., & Mills, J., eds. (1999). *Cognitive dissonance: Progress on a pivotal theory in social psychology.* Washington, DC: American Psychological Association. The reference to cognitive dissonance theory was proposed in Bazerman, M. H., Giuliano, T., & Appelman, A. (1984). Escalation of commitment in individual and group decision making. *Organizational Behavior & Human Performance, 33,* 141–152.

113. This discrepancy is indicated by a preference for retroactive information that speaks to the decision already made, rather than prospective information about the decision to be made. See Beeler, J. D., & Hunton, J. E. (1997). The influence of compensation method and disclosure level on information search strategy and escalation of commitment. *Journal of Behavioral Decision Making, 10,* 77–91; Conlon, E. J., & Parks, J. M. (1987). Information requests in the context of escalation. *Journal of Applied Psychology, 72,* 344–350.

114. For example, when rating employees whom they originally hired, managers tend to inflate the ratings of their effectiveness, likelihood of improvement, and potential for promotion. See Schoorman, F. D. (1988). Escalation bias in performance appraisals: An unintended consequence of supervisor participation in hiring decisions. *Journal of Applied Psychology, 73,* 58–62; Bazerman, M. H., Beekun, R. I., & Schoorman, F. D. (1982). Performance evaluation in a dynamic context: A laboratory study of the impact of a prior commitment to the ratee. *Journal of Applied Psychology, 67,* 873–876; Slaughter, J. E., & Greguras, G. J. (2008). Bias in performance ratings: Clarifying the role of positive versus negative escalation. *Human Performance, 21,* 414–426.

115. Specifically, players picked high in the draft were given more game time and retained longer than players picked lower in the draft. Staw, B. M., & Hoang, H. (1995). Sunk costs in the NBA: Why draft order affects playing time and survival in professional basketball. *Administrative Science Quarterly, 40,* 474–494.

116. Staw, B. M., Barsade, S. G., & Koput, K. W. (1997). Escalation at the credit window: A longitudinal study of bank executives' recognition and write-off of problem loans. *Journal of Applied Psychology, 82,* 130–142.

117. One study found that season ticket holders who paid full price for the subscription attended more performances than did subscribers who bought it at a discount. Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes, 35,* 124–140 (study 2).

118. Schoorman (1988), *supra* note 114.

119. O'Brien, B. (2009). Prime suspect: An examination of factors that aggravate and counteract confirmation bias in criminal investigations. *Psychology, Public Policy, and Law, 15,* 315–334.

120. Staw & Fox (1977), *supra* note 111; Bobocel, D. R., & Meyer, J. P. (1994). Escalating commitment to a failing course of action: Separating the roles of choice and justification. *Journal of Applied Psychology, 79,* 360–363; Whyte, G. (1993). Escalating commitment in individual and group decision making: A prospect theory approach. *Organizational Behavior and Human Decision Processes, 54,* 430–455.

121. Harrison, P. D., & Harrell, A. (1993). Impact of "adverse selection" on managers' project evaluation decisions. *Academy of Management Journal, 36,* 635–643. See also Simonson, I., & Nye, P. (1992). The effect of accountability on susceptibility to decision errors. *Organizational Behavior and Human Decision Processes, 51,* 416–446 (study 6).

122. Staw & Fox (1977), *supra* note 111.

123. Zhang, L., & Baumeister, R. F. (2006). Your money or your self-esteem: Threatened egotism promotes costly entrapment in losing endeavors. *Personality and Social Psychology Bulletin, 32,* 881–893; Harrison & Harrell (1993), *supra* note 121.

124. Beeler & Hunton (1997), *supra* note 113; Bobocel & Meyer (1994), *supra* note 120.

125. Garland & Conlon (1998), *supra* note 111; Moon, H. (2001). Looking forward and looking back: Integrating completion and sunk-cost effects within an escalation-of-commitment progress decision. *Journal of Applied Psychology, 86,* 104–113; Boehne, D. M., & Paese, P. W. (2000). Deciding whether to complete or terminate an unfinished project: A strong test of the project completion hypothesis. *Organizational Behavior and Human Decision Processes, 81,* 178–194.

126. Greitemeyer, T., Schulz-Hardt, S., & Frey, D. (2009). The effects of authentic and contrived dissent on escalation of commitment in group decision making. *European Journal of Social Psychology, 39,* 639–647; Bazerman, Giuliano, & Appelman (1984), *supra* note 112.

127. See, e.g., Whyte (1993), *supra* note 120.

128. Marques, J., Abrams, D., & Serôdio, R. G. (2001). Being better by being right: Subjective group dynamics and derogation of in-group deviants when generic norms are undermined. *Journal of Personality and Social Psychology, 81,* 436–447; Cota, A. A., Evans, C. R., Dion, K. L., Kilik, L., et al. (1995). The structure of group cohesion. *Personality and Social Psychology Bulletin, 21,* 572–580.

129. Jaffe & Yinon (1979), *supra* note 101.

130. Schachter, S. (1951). Deviation, rejection, and communication. *Journal of Abnormal and Social Psychology, 46,* 190–207. This is not to say that groups are always perfectly harmonious or egalitarian. Even when they are engaged in intergroup conflict, groups have internal stratification, hierarchical divisions, and even rivalry. Wit, A. P., & Kerr, N. L. (2002). "Me versus just us versus us all" categorization and cooperation in nested social dilemmas. *Journal of Personality and Social Psychology, 83,* 616–637. For the most part, though, these differences are tucked away within the groups, and tend not to be observable from across the intergroup divide.

131. See Blockars, C. B., Ikkovic, S. K., & Haberfeld, M. R. (2006). *Enhancing police integrity.* Dordrecht: Springer; Savitz, L. (1970). The dimensions of police loyalty. *American Behavioral Scientist, 13,* 693–704; Westley, W. A. (1970). *Violence and the police: A sociological study of law, custom, and morality.* Cambridge, MA: MIT Press. Selection to a detective unit usually entails a commitment to the team. One must prove oneself to be loyal, a team player. Loyalty affects retention and promotion in the units. See Manning (2006), *supra* note 7. The case of Richard Ceballos provides an example of group cohesion against a prosecutor who broke ranks. See *Garcetti v. Ceballos,* 547 U.S. 410 (2006).

132. For example, after the suspects in the Ford Heights case were charged with this high-profile murder, the Chicago police ignored witnesses who provided them with the correct evidence. Protess, D., & Warden, R. (1998). *A promise of justice,* chaps. 12, 14. New York: Hyperion.

133. The lead character in the Chinese film *King of Masks* (directed by Minglun Wei, 1996) captures this intuition: "The lightest breeze can blow you into jail, but the strongest ox cannot pull you out."

134. For prosecutorial resistance to admit investigative or prosecutorial errors, see Medwed, D. (2004). The zeal deal: Prosecutorial resistance to post-conviction claims of innocence. *Boston University Law Review, 84,* 125–183.

135. Against the advice of his lawyer Scott Borthwick (who worked on the case pro bono), James Ochoa pleaded guilty to the crime. He was subsequently exonerated after the biological evidence from the crime scene was matched to a

man who was imprisoned at the time for similar crimes. Reza, H. G. (2006). Innocent man grabs his freedom and leaves town. *Los Angeles Times,* November 2. See also Moxely, R. S. (2005). The case of the dog that couldn't sniff straight. *OC Weekly,* November 5; Innocence Project, profile, James Ochoa. http://www.innocenceproject.org/Content/James_Ochoa.php.

136. See Garrett, B. L. (2011). *Convicting the innocent: Where criminal prosecutions go wrong,* pp. 100–102. Cambridge, MA: Harvard University Press.

137. Cooper, C. L., & Grimley, P. J. (1983). Stress among police detectives. *Journal of Occupational Medicine, 25,* 534–540; Wright, A. (2007). Ethics and corruption. In T. Newburn, T. Williamson, & A. Wright, eds., *Handbook of criminal investigation,* pp. 586–609, p. 605. Portland, OR: Willan Publishing.

138. Investigative dilemmas have been described as ethical "minefields." Wright (2007), *supra* note 137, p. 605.

139. On the tendency to discount the significance of future events, see Ainslie, G., & Haslam, N. (1992). Hyperbolic discounting. In G. Loewenstein & J. Elster, eds., *Choice over time,* pp. 57–92. New York: Russell Sage Foundation. It has also been shown that events that are far in the future are typically represented more abstractly whereas proximate events are seen to be more concrete and detailed. Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review, 110,* 403–421.

140. For a review of the literature, see Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5,* 69–106.

141. See Ramsey, R. J., & Frank, J. (2007). Wrongful conviction: Perceptions of criminal justice professionals regarding the frequency of wrongful conviction and the extent of system errors. *Crime & Delinquency, 53,* 436–470; Zalman, M., Smith, B., & Kiser, A. (2008). Officials' estimates of the incidence of "actual innocence" convictions. *Justice Quarterly, 25,* 72–100.

142. See Leo, R. A. (2008). *Police interrogation and American justice*. Cambridge, MA: Harvard University Press. A number of English criminologists depict *case construction* as a set of biased practices that include prejudicial decision making and manipulation of the facts in order to achieve strong cases for the prosecution. See, e.g., McConville, M., Sanders, M., & Leng, R. (1991). *The case for the prosecution: Police suspects and the construction of criminality.* London: Routledge. For discussions, see Innes (2003), *supra* note 5, pp. 214–216; Bayley (1994), *supra* note 48, p. 27. For opposite views, see Smith, D. J. (1997). Case construction and the goals of criminal process. *British Journal of Criminology, 37,* 319–346.

143. *Johnson v. United States,* 333 U.S. 10, 13–14 (1948).

144. See National Academy of Science (2009). *Strengthening forensic science in the United States: A path forward*. Washington, DC: National Academies

Press; Garrett, B. L., & Neufeld, P. J. (2009). Invalid forensic science testimony and wrongful convictions. *Virginia Law Review, 95,* 1–97; Mnookin, J. L. (2010). The Courts, the NAS, and the Future of Forensic Science. *Brooklyn Law Review, 75*, 1209–1275; Giannelli (1997), *supra* note 55.

145. Anecdotal evidence suggests that high-profile crimes are particularly susceptible to adversarial pressures, and thus to guilt-prone error. Notable examples include the conviction of five youths for the notorious assault on the Central Park jogger (see Saulny, S. [2002]. Convictions and charges voided in '89 Central Park jogger attack. *New York Times,* December 20. http://www.nytimes.com/2002/12/20/nyregion/convictions-and-charges-voided-in-89-central-park-jogger-attack.html); the indictment of three members of the Duke University lacrosse team on sexual assault and kidnapping charges (Wilson D., & and Barstow, D. [2007]. All charges dropped in Duke case. *New York Times,* April 12. http://www.nytimes.com/2007/04/12/us/12duke.html); the relentless pursuit of Steven Hatfill, the army scientist who was suspected of being responsible for the deadly anthrax attacks in 2001 (Shane, S., & Lichtblau, E. [2008]. New details on F.B.I.'s false start in anthrax case. *New York Times,* November 25. http://www.nytimes.com/2008/11/26/washington/26anthrax.html?_r=1); the prolonged arrest of Wen Ho Lee, mistakenly suspected of spying for China (F.B.I. faulted in nuclear secrets investigation. *New York Times,* December 13, 2001. http://www.nytimes.com/2001/12/13/us/fbi-faulted-in-nuclear-secrets-investigation.html); the overturned conviction of members of the alleged Al Qaeda "sleeper operational combat cell" in Detroit (Hakim, D., & Lichtblau, E. [2004]. After convictions, the undoing of a U.S. terror prosecution. *New York Times,* October 7. http://www.nytimes.com/2004/10/07/national/07detroit.html); the false confession obtained from Abdallah Higazzi, implicating himself in the terrorist attacks on the World Trade Center's Twin Towers on September 11 (Dwyer, J. [2007]. Roots of false confession: Spotlight is now on the F.B.I. *New York Times,* October 31. http://www.nytimes.com/2007/10/31/nyregion/31about.html?ref=abdallahhigazy); and the prosecution of Alaska senator Ted Stevens (Lewis, N. A. [2009]. Tables turned on prosecution in Stevens case. *New York Times,* April 7. http://www.nytimes.com/2009/04/08/us/politics/08stevens.html).

146. As mentioned in Chapter 1, police misconduct was observed in almost half of DNA exoneration cases, and prosecutorial misconduct was found in 45 percent of cases. False or misleading forensic testimony was given in about one-quarter of the cases.

147. The term *testilying* was coined by officers who were involved in committing perjury. Commission to Investigate Allegations of Police Corruption and the Anti-Corruption Practices of the Police Department, Milton Mollen, Chair, July 7, 1994, at 36. See also Slobogin, C. (1996). Testilying: Police perjury and what to do about it. *Colorado Law Review, 67,* 1037–1060. Renowned criminologist Jerome Skolnick observes that for the police, "lying is a routine way of

managing legal impediments—whether to protect fellow officers or to compensate for what [the officer] views as limitations the courts have placed on his capacity to deal with criminals." Skolnick, J. H. (1982). Deception by police. *Criminal Justice Ethics,* Summer/Fall, 40–54.

148. Diane Vaughn's study of NASA activities leading up to the crash of the Challenger shuttle reveal such a culture shift. Working under extreme pressure, NASA scientists and engineers progressively deviated from the standard operating procedures, and gradually generated a culture that normalized faulty practices. Without engaging in willful misconduct, these practices led to deeply flawed decision making. Vaughn, D. (1996). *The Challenger launch decision: Risky technology, culture, and deviance at NASA.* Chicago: University of Chicago Press.

149. For the strengthening of the police officers' attitudes over time, see Wortley & Homel (1995), *supra* note 50; Gatto, J., Dambrun, M., Kerbrat, C., & De Olivera, P. (2010). Prejudice in the police: On the processes underlying the effects of selection and group socialization. *European Journal of Social Psychology, 40,* 252–269; Perrott & Taylor (1995), *supra* note 50.

150. According to this count, the prosecution's case contained 139 evidence items and the defense's case 199. Kadane, J. B., & Schum, D. A. (1996). *A probabilistic analysis of the Sacco and Vanzetti case,* pp. 80, 286–337. New York: John Wiley & Sons.

151. The term *mental model* is used here in the broad sense of a structured representation. See Markman, A. B. (1999). *Knowledge representation.* Mahwah, NJ: Lawrence Erlbaum.

152. For experimental results, see Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General, 128,* 3–31; Simon, D., Pham, L. B., Le, Q. A., & Holyoak, K. J. (2001). The emergence of coherence over the course of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1250–1260; Simon, D., Snow, C. J., & Read, S. J. (2004). The redux of cognitive consistency theories: Evidence judgments by constraint satisfaction. *Journal of Personality and Social Psychology, 86,* 814–837; Simon, D., Krawczyk, D. C., & Holyoak, K. J. (2004). Construction of preferences by constraint satisfaction. *Psychological Science, 15,* 331–336; Simon, D., Krawczyk, D. C., Bleicher, A., & Holyoak, K. J. (2008). The transience of constructed preferences. *Journal of Behavioral Decision Making, 21,* 1–14; Glöckner, A., & Betsch, T. (2008). Multiple-reason decision making based on automatic processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1055–1075; Glöckner, A., Betsch, T., & Schindler, N. (2010). Coherence shifts in probabilistic inference tasks. *Journal of Behavioral Decision Making, 23,* 439–462.

For reviews of the coherence effect, see Simon, D., & Holyoak, K. J. (2002). Structural dynamics of cognition: From consistency theories to constraint satisfaction. *Personality and Social Psychology Review, 6,* 283–294; Simon, D.

(2004). A third view of the black box: Cognitive coherence in legal decision making. *University of Chicago Law Review, 71,* 511–586.

For overviews of the underlying cognitive architecture, see Read, S. J., Vanman, E. J., & Miller, L. C. (1997). Connectionism, parallel constraint satisfaction processes, and Gestalt principles: (Re)introducing cognitive dynamics to social psychology. *Personality and Social Psychology Review, 1,* 26–53; Thagard, P. (2000). *Coherence in thought and action.* Cambridge, MA: MIT Press.

153. See Holyoak & Simon (1999), *supra* note 152, studies 2, 3.

154. Simon et al. (2001), *supra* note 152.

155. Simon, Snow, & Read (2004), *supra* note 152.

156. See Simon, Stenstrom, & Read (2008), *supra* note 65.

157. Ask & Granhag (2007a), *supra* note 44.

158. Simon, Stenstrom, & Read (2008), *supra* note 65.

159. By the same token, providing information that placed the defendant far from the scene led to more exculpatory evaluations of the rest of the evidence. Simon, Snow, & Read (2004), *supra* note 152, study 3. The effect of adding one piece of evidence on all the other evidence items was observed also in Holyoak & Simon (1999), *supra* note 152, study 3; Simon, Krawczyk, & Holyoak (2004), *supra* note 152, study 2.

160. Holyoak & Simon (1999), *supra* note 152, study 3.

161. Moreover, learning of a confession caused witnesses to change the responses they had given at a lineup conducted two days earlier. Hasel, L. E., & Kassin, S. M. (2009). On the presumption of evidentiary independence: Can confessions corrupt eyewitness identifications? *Psychological Science, 20,* 122–126.

162. Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83,* 360–376; Wells, G. L., Olson, E. A., & Charman, S. D. (2003). Distorted retrospective eyewitness reports as functions of feedback and delay. *Journal of Experimental Psychology: Applied, 9,* 42–52.

163. Likewise, simulated investigators rated a facial composite image to be more similar to the suspect when told that he had been identified by eyewitnesses. Lower similarity ratings were given when the investigators were told that the witnesses had not identified the suspect and when they were given no information about the witnesses' identification. In reality, the facial composite was not based at all on the suspect. Charman, S. D., Gregory, A. H., & Carlucci, M. (2009). Exploring the diagnostic utility of facial composites: Beliefs of guilt can bias perceived similarity between composite and suspect. *Journal of Experimental Psychology: Applied, 15,* 76–90.

164. Elaad, E., Ginton, A., & Ben-Shakhar, G. (1994). The effects of prior expectations and outcome knowledge on polygraph examiners' decisions. *Journal of Behavioral Decision Making, 7,* 279–292.

165. Unbeknownst to the participants, they had previously analyzed those very prints in a real-life case. The results showed that almost half of the experts were misled by the information, reaching conclusions that were opposite to their own prior judgments. Of the twelve relevant cases (where the matches were not easy, and where incorrect information was suggested), three judgments were reversed. Notably, experts reversed their previous findings also in two of the twelve difficult cases that contained no extraneous information. Dror & Charlton (2006), *supra* note 53.

166. Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley. The positive test strategies was intuited also by Francis Bacon: "it is the peculiar and perpetual error of the human intellect to be more moved and excited by affirmatives than be negatives" (Bacon, F. [1844]. *Novum organum or true suggestions for the interpretation of nature*, p. 21. London: William Pickering).

167. A similar strategy, the *hypothesis-preservation strategy,* involves asking questions that are likely to lead to the conclusion that the working hypothesis is true. For reviews, see Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information hypothesis testing. *Psychological Review, 94,* 211–228; Nickerson (1998), *supra* note 26.

168. Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content.* Cambridge, MA: Harvard University Press.

169. Klayman (1995), *supra note* 26, p. 399. According to Jonathan Baron, the phenomenon can be described in the following terms: "To test a hypothesis, think of a result that would be found if the hypothesis were true and then look for that result (and do not worry about other hypotheses that might yield the same result)." Baron has labeled this the *congruence heuristic.* Baron, J. (2000). *Thinking and deciding,* p. 162. New York: Cambridge University Press.

The research has identified two information-gathering strategies that people use when testing hypotheses in making social judgments: a *diagnostic strategy* asks questions whose answers permit the greatest distinction between the focal hypothesis and its alternatives. A *confirmation strategy* tends to rely on questions that confirm the hypothesis without much regard to their diagnosticity. Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology, 22,* 93–121.

170. Snyder, M., & Swann, W. B. (1978). Hypothesis testing processes in social interaction. *Journal of Personality and Social Psychology, 36,* 1202–1212.

171. Kassin, S. M., Goldstein, C. C., & Savitsky, K. (2003). Behavioral confirmation in the interrogation room: On the dangers of presuming guilt. *Law and Human Behavior, 27,* 187–203.

172. Selective exposure was one of the central themes in Leon Festinger's cognitive dissonance theory. Festinger (1957), *supra* note 112, chaps. 6, 7. See

also Frey, D. (1986). Recent research on selective exposure to information. In L. Berkowitz, ed., *Advances in experimental social psychology,* vol. 19, pp. 41–80. New York: Academic Press; Snyder & Swann (1978), *supra* note 170; Jonas, E., Schulz-Hardt, S., Frey, D., & Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: An expansion of dissonance theoretical research on selective exposure to information. *Journal of Personality and Social Psychology, 80,* 557–571.

173. Selective exposure was observed also during the Senate's Watergate hearings of 1973, which were followed more by supporters of the Democratic Party than by Republicans. Sweeney, P. D., & Gruber, K. L. (1984). Selective exposure: Voter information preferences and the Watergate affair. *Journal of Personality and Social Psychology, 46,* 1208–1221.

174. Ehrlich, D., Guttman, I., Schönbach, P., & Mills, J. (1957). Postdecision exposure to relevant information. *Journal of Abnormal and Social Psychology, 54,* 98–102.

175. Holton, B., & Pyszczynski, T. (1989). Biased information search in the interpersonal domain. *Personality and Social Psychology Bulletin, 15,* 42–51.

176. Fischer, P., Jonas, E., Frey, D., & Schulz-Hardt, S. (2005). Selective exposure to information: The impact of information limits. *European Journal of Social Psychology, 35,* 469–492.

177. Kunda, Z. & Sinclair, L. (1999). Motivated reasoning with stereotypes: Activation, application, and inhibition. *Psychological Inquiry, 10,* 12–22.

178. A study of lay people's judgments of judicial decisions found that when the participants agree with the outcome of the court's decision, they are indifferent to the type of reasoning offered by the court. When they disagree with the outcome, they react differently to different modes of reasoning. Simon, D., & Scurich, N. (2011). Lay judgments of judicial decision making. *Journal of Empirical Legal Studies, 8,* 709–727.

179. Edwards & Smith (1996), *supra* note 29. Similar findings were made by political scientists Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science, 50,* 755–769.

180. Wyer & Frey (1983), *supra* note 59. For similar findings, see Pyszczynski, T., Greenberg, J., & Holt, K. (1985). Maintaining consistency between self-serving beliefs and available data: A bias in information evaluation. *Personality and Social Psychology Bulletin, 11,* 179–190.

181. Ditto et al. (2003), *supra* note 58.

182. The small but indisputable flaw was noticed by 71 percent of the reviewers who disagreed with the study's results, but by only 25 percent of the reviewers who agreed with them. Mahoney (1977), *supra* note 31.

183. This mechanism is also labeled *biased assimilation;* Lord, Ross, & Lepper (1979), *supra* note 49.

184. Duncan, B. L. (1976). Differential social perception and attribution of intergroup violence: Testing the lower limits of stereotyping of blacks. *Journal of Personality and Social Psychology, 34,* 590–598; Cohen (1981), *supra* note 35.

185. Munro et al. (2002), *supra* note 60.

186. Hastorf & Cantril (1954), *supra* note 61.

187. Brownstein, Read, & Simon (2004), *supra* note 63.

188. Dror, Charlton, & Péron (2006), *supra* note 46; Dror & Charlton (2006), *supra* note 53.

189. Shaklee, H., & Fischhoff, B. (1982). Strategies of information search in causal analysis. *Memory & Cognition, 10,* 520–530; Saad, G., & Russo, J. E. (1996). Stopping criteria in sequential choice. *Organizational Behavior and Human Decision Processes, 67,* 258–270.

190. Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology, 63,* 568–584.

191. McGonigle, S., & and Emily, J. (2008). A blind faith in eyewitnesses: 18 of 19 local cases overturned by DNA relied heavily on unreliable testimony. *Dallas Morning News,* October 12, p. 1A.

192. For reviews, see Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125,* 255–275; Tetlock, P. E. (2002). Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review, 109,* 451–471.

193. Tetlock, P. E., & Boettger, R. (1989). Accountability: A social magnifier of the dilution effect. *Journal of Personality and Social Psychology, 57,* 388–398; Lerner & Tetlock (1999), *supra* note 192; Simonson & Nye (1992), *supra* note 121.

194. Wogalter, M. S., Malpass, R. S., & Mcquiston, D. E. (2004). A national survey of police on preparation and conduct of identification lineups. *Psychology, Crime & Law, 10,* 69–82.

195. In jurisdictions that have recently undergone a reform of identification procedures, 23 percent of officers videotape the procedures. Wise, R. A., Safer, M. A., & Maro, C. M. (2011). What U.S. law enforcement officers know and believe about eyewitness interviews and identification procedures. *Applied Cognitive Psychology, 25,* 488–500.

196. Incomplete records were mentioned in *Coleman v. Alabama,* 399 U.S. 1 (1970); *Gilbert v. California,* 388 U.S. 263 (1967); *Neil v. Biggers,* 409 U.S. 188 (1972); *Simmons v. United States,* 390 U.S. 377 (1968); *Stovall v. Denno,* 388 U.S. 263 (1967); and *United States v. Ash,* 413 U.S. 300 (1973).

197. Warren, A. R., & Woodall, C. E. (1999). The reliability of hearsay testimony: How well do interviewers recall their interviews with children? *Psychology, Public Policy, and Law, 5,* 355–371. The latter finding was observed also in a study in which mothers were asked about a conversation they had had some days earlier with their children. Only one of every six questions asked was recalled.

Bruck, M., Ceci, S. J., & Francoeur, E. (1999). The accuracy of mothers' memories of conversations with their preschool children. *Journal of Experimental Psychology: Applied, 5,* 89–106.

198. This study compared the notes taken with audio-tape recordings of the interviews. Lamb, M. E., Orbach, Y., Sternberg, K. J., Hershkowitz, I., & Horowitz, D. (2000). Accuracy of investigators' verbatim notes of their forensic interviews with alleged child abuse victims. *Law and Human Behavior, 24,* 699–708.

199. Gregory, A. H., Schreiber-Compo, N., Vertefeuille, L., & Zambrusky, G. (2011). A comparison of US police interviewers' notes with their subsequent reports. *Journal of Investigative Psychology and Offender Profiling, 8,* 203–215.

200. Moreover, under certain circumstances, accountability can actually increase bias. The construct's darker side appears when conformity, rather than preemptive self-criticism, is deemed the better way to gain the intended audience's approval. For example, when asked to explain their positions on issues such as affirmative action, university tuition increases, and nuclear armament, participants expressed more liberal views to the liberal audience and more conservative views to the conservative audience. Tetlock, P. E., Skitka, L., & Boettger, R. (1989). Social and cognitive strategies for coping with accountability: Conformity, complexity, and bolstering. *Journal of Personality and Social Psychology, 57,* 632–640. A criminal investigator wanting to curry favor with a heavy-handed superior or an overambitious prosecutor will be more likely to reach conclusions that comport with those preferences.

201. The FBI report: Stacey, R. B. (2004). A report on the erroneous fingerprint individualization in the Madrid train bombing case. *Journal of Forensic Identification, 54,* 706. The DOJ report: Department of Justice, Office of the Inspector General of the Oversight and Review Division (2006a). *A review of the FBI's handling of the Brandon Mayfield case, Executive Summary.* Washington, DC. http://www.usdoj.gov/oig/special/s0601/exec.pdf.

202. E.g., *United States v. Llera Plaza,* 188 F. Supp. 2d 549 (E. D. Pa. 2002). See Cole (2005), *supra* note 55.

203. The official reports of the Mayfield investigation conclude that the error was not driven by Mayfield's religion. Department of Justice (2006a), *supra* note 201, p. 18. Yet it seems inconceivable that investigators would have overlooked the fact that this former military man had embraced Islam and maintained contacts with suspected and convicted terrorists. One of the examiners admitted that if the person identified had been someone without Islamic characteristics, like the "Maytag Repairman," the laboratory might have treated the identification with greater skepticism. Ibid., p. 12.

204. Kershaw, S. (2004). Spain and U.S. at odds on mistaken terror arrest. *New York Times,* June 5, p. A1. http://www.nytimes.com/2004/06/05/us/spain-and-us-at-odds-on-mistaken-terror-arrest.html?scp=1&sq=kershaw%20sarah%20spain%20us%20at%20odds&st=cse.

205. Ibid.

206. Stacey (2004), *supra* note 201.

207. Ibid. The DOJ report found no evidence that the investigators were influenced by high profile nature of the case. Department of Justice (2006a), *supra* note 201, p. 11.

208. Department of Justice (2006a), *supra* note 201, p. 8.

209. "Level 3" details include tiny individual pores, incipient dots between ridges, ridge edges, and small between-ridge details. These details are controversial because they are small, and their appearance is highly variable, even between different prints made by the same finger. Ibid.

210. Ibid.

211. The examiners explained away the apparent mismatch of this region on the basis of a "double touch" theory, an explanation that was flatly rejected by the experts advising the inquiries. Ibid., p. 9.

212. Ibid., p. 8.

213. Ibid., p. 12.

214. Ibid., p. 7.

215. Stacey (2004), *supra* note 201.

216. Department of Justice (2006a), *supra* note 201, p. 10.

217. "Points," or "minutiae," are places where the individual ridges in the fingerprint end or split.

218. Kershaw (2004), *supra* note 204.

219. Ibid., quoting Mr. Corrales.

220. Department of Justice (2006a), *supra* note 201, p. 11.

221. Kershaw (2004), *supra* note 204. The judge stated: "I have no affidavit from any Spanish authorities as to questioning the fingerprint. The only information I have is that after consulting with the FBI, that they agreed with the 100 percent identification." Cited in Department of Justice, Office of the Inspector General of the Oversight and Review Division (2006b). *A review of the FBI's handling of the Brandon Mayfield case,* p. 80. Washington, DC. http://www.justice.gov/oig/special/s0601/Chapter2.pdf) The DOJ report described the inaccuracies in the affidavits as a "regrettable inattention to detail" (ibid., p. 268). The conduct of the attorneys was outside the purview of the DOJ inquiry.

222. Kershaw (2004), *supra* note 204.

223. Ibid.

224. See, e.g., De Bono, E. (1968). *New think: The use of lateral thinking in the generation of new ideas.* New York: Basic Books.

225. Detectives are encouraged to continually challenge the meaning and reliability of any material they gather. National Centre for Police Excellence (2005). *Practice advice on core investigative doctrine,* p. 62. Cambourne, UK: Association of Chief Police Officers. The English statute governing police investigations

(PACE) requires that all reasonable inquiries, both indicating and challenging the responsibility of the suspect, ought to be undertaken and recorded.

226. Canadian courts have ordered police officers to take into account "all the information available." Officers are entitled to disregard evidence only if they find it to be unreliable. *Dix v. AG Canada,* 2002, para. 357.

227. Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47,* 1231–1243; Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin, 26,* 1142–1150. Some research indicates that debiasing can occur when one considers any other hypothesis, not only the opposite one. Hirt & Markman (1995), *supra* note 33.

228. Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin, 110,* 486–498. See also Mussweiler, Strack, & Pfeiffer (2000), *supra* note 227.

229. For example, although the intervention succeeded in reducing students' beliefs in a random scenario they were asked to explain (the victory of a team in a random sporting event), it was unsuccessful in debiasing their beliefs when their motivation was implicated in the outcome (a victory of their own team). Markman & Hirt (2002), *supra* note 19, study 1.

230. See, e.g., Sanna, L. J., Schwarz, N., & Stocker, S. L. (2002). When debiasing backfires: Accessible content and accessibility experiences in debiasing hindsight. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 497–502; Hirt & Markman (1995), *supra* note 33, study 3.

231. A similar intervention involves designating a *devil's advocate,* which assigns the responsibility to lodge a critique of the focal hypothesis, without necessarily offering a countertheory.

232. For a review of the literature and a meta-analysis, see Schwenk, C. R. (1990). Effects of devil's advocacy and dialectical inquiry on decision making: A meta-analysis. *Organizational Behavior and Human Decision Processes, 47,* 161–176.

233. Greitemeyer, Schulz-Hardt, & Frey (2009), *supra* note 126; Nemeth, C., Brown, K., & Rogers, J. (2001). Devil's advocate versus authentic dissent: Stimulating quantity and quality. *European Journal of Social Psychology, 31,* 707–720. See also Gunia, B. C., Sivanathan, N., & Galinsky, A. D. (2009). Vicarious entrapment: Your sunk costs, my escalation of commitment. *Journal of Experimental Social Psychology, 45,* 1238–1244.

234. Kerstholt & Eikelbloom (2007), *supra* note 45.

235. As observed by Jacqueline Hodgson, the investigative magistrates *(juges d'instruction)* often tend to verify the evidence that was gathered by the police before being appointed to the case. Hodgson, J. (2005). *French criminal justice:*

A comparative account of the investigation and prosecution of crime in France, p. 247. Oxford: Hart Publishing.

236. Schachter (1951), *supra* note 130; Nemeth, Brown, & Rogers (2001), *supra* note 233.

237. Nemeth, C. J., Connell, J. B., Rogers, J. D., & Brown, K. S. (2001). Improving decision making by means of dissent. *Journal of Applied Social Psychology, 31,* 48–58; Nemeth, Brown, & Rogers (2001), *supra* note 233.

238. See http://www.dallasda.com/. Tellingly, the front page of the Summer 2011 issue of the newsletter of Watkins's office, *The Justice Report,* carries the story of an exoneration of a man convicted for aggravated rape by the office in 1984: http://dallascounty.org/department/da/media/Summer2011.pdf.

239. The high rate of exonerations is made possible by the fact that Dallas County has traditionally kept evidence from closed cases, which has enabled the presentation of compelling evidence for post-conviction review, at least in some cases.

240. On innocence commissions, see Chapter 8.

241. As noted in Chapter 1, Lloyd Weinreb proposed the establishment of an "investigating magistracy" (Weinreb [1977], *supra* note 54, p. 119). George Thomas proposed that criminal investigations and pretrial procedures be overseen by a "screening magistrate": Thomas, G. C. III (2008). *The supreme court on trial: How the American justice system sacrifices innocent defendants*, pp. 193–227. Ann Arbor: University of Michigan Press. Keith Findley has suggested a system that blends the strengths of the adversarial and inquisitorial systems: Findley, K. A. (in press). Adversarial inquisitions: Rethinking the search for the truth. *New York Law Review*.

242. See Kassin, S. M. (1998). Eyewitness identification procedures: The fifth rule. *Law and Human Behavior, 22,* 649–653.

243. The well-known RAND study of police investigations found that many investigative records are incomplete and casually maintained. Police files covered between 26 percent and 45 percent of the evidentiary questions considered essential by prosecutors. The authors posited that poor recordkeeping results in higher case dismissal rates and weakening of the prosecutors' plea-bargaining position. Greenwood, P. W., Chaiken, J. M., Petersilia, J., & Prusoff, L. L. (1975). *The criminal investigation process, Part III.* Santa Monica, CA: RAND. Likewise, experienced Canadian police officers concede that their note-taking habits result in case dismissals. Yuille, J. C. (1984). Research and teaching with police: A Canadian example. *International Review of Applied Psychology, 33,* 5–23.

SAUL KASSIN

Saul Kassin is a Distinguished Professor of Psychology at John Jay College of Criminal Justice in New York and Massachusetts Professor of Psychology at Williams College. He received his Ph.D. at the University of Connecticut. He later served as a postdoctoral research fellow at the University of Kansas; a U. S. Supreme Court Judicial Fellow, working at the Federal Judicial Center in D.C.; and a visiting professor at Stanford University. Dr. Kassin is an author of several college textbooks—including *Social Psychology*, now in its Ninth Edition; *Confessions in the Courtroom*, *The Psychology of Evidence and Trial Procedure*; and *The American Jury on Trial: Psychological Perspectives.*

Starting in the 1980's, Dr. Kassin pioneered the scientific study of false confessions by introducing a three-part taxonomy that is universally accepted today. He also developed experimental laboratory paradigms that are widely used to examine why people are targeted for interrogation, why they waive their rights, why they confess, the corruptive effects of confessions on other evidence, the consequences of confessions in court, and the use of video recording to alleviate these problems. Dr. Kassin has published numerous scientific articles and book chapters on the subject and was awarded an APA presidential citation for his research on false confessions. His work has been cited all over the world—including by the U.S. Supreme Court—and his research on the video recording of interrogations is currently funded by the National Science Foundation.

Dr. Kassin is Past President of the American Psychology-Law Society (AP-LS). He has served as a consultant in a number of high profile cases and has testified in several state, federal, and military courts. He is senior author of the 2010 AP-LS White Paper entitled "Police-Induced Confessions: Risk Factors and Recommendations." He lectures frequently to psychologists, judges, lawyers, law enforcement groups, and state criminal justice commissions and task forces; and he has appeared as a media consultant and guest for all major news networks and in a number of documentaries—including Ken Burns' 2012 film, *The Central Park Five*.

Law Hum Behav (2010) 34:3–38
DOI 10.1007/s10979-009-9188-6

ORIGINAL ARTICLE

Police-Induced Confessions: Risk Factors and Recommendations

Saul M. Kassin · Steven A. Drizin · Thomas Grisso · Gisli H. Gudjonsson · Richard A. Leo · Allison D. Redlich

Published online: 15 July 2009

Abstract Recent DNA exonerations have shed light on the problem that people sometimes confess to crimes they did not commit. Drawing on police practices, laws concerning the admissibility of confession evidence, core principles of psychology, and forensic studies involving multiple methodologies, this White Paper summarizes what is known about police-induced confessions. In this review, we identify suspect characteristics (e.g., adolescence; intellectual disability; mental illness; and certain personality traits), interrogation tactics (e.g., excessive interrogation time; presentations of false evidence; and minimization), and the phenomenology of innocence (e.g., the tendency to waive *Miranda* rights) that influence confessions as well as their effects on judges and juries. This article concludes with a strong recommendation for the mandatory electronic recording of interrogations and considers other possibilities for the reform of interrogation practices and the protection of vulnerable suspect populations.

Keywords Police interviews · Interrogations · Confessions

S. M. Kassin (✉)
John Jay College of Criminal Justice, City University of New York, New York, NY, USA
e-mail: skassin@jjay.cuny.edu

S. A. Drizin
Northwestern University School of Law and Center on Wrongful Convictions, Chicago, IL, USA

T. Grisso
University of Massachusetts Medical School, Worcester, MA, USA

G. H. Gudjonsson
Institute of Psychiatry, King's College, London, UK

R. A. Leo
University of San Francisco School of Law, San Francisco, CA, USA

A. D. Redlich
State University of New York at Albany, Albany, NY, USA

In recent years, a disturbing number of high-profile cases, such as the Central Park jogger case, have surfaced involving innocent people who had confessed and were convicted at trial, only later to be exonerated (Drizin & Leo, 2004; Gudjonsson, 1992, 2003; Kassin, 1997; Kassin & Gudjonsson, 2004; Lassiter, 2004; Leo & Ofshe, 1998). Although the precise incidence rate is not known, research suggests that false confessions and admissions are present in 15–20% of all DNA exonerations (Garrett, 2008; Scheck, Neufeld, & Dwyer, 2000; http://www.innocenceproject.org/). Moreover, because this sample does not include those false confessions that are disproved before trial, many that result in guilty pleas, those in which DNA evidence is not available, those given to minor crimes that receive no post-conviction scrutiny, and those in juvenile proceedings that contain confidentiality provisions, the cases that are discovered most surely represent the tip of an iceberg.

In this new era of DNA exonerations, researchers and policy makers have come to realize the enormous role that psychological science can play in the study and prevention of wrongful convictions. In cases involving wrongfully convicted defendants, the most common reason (found in three-quarters of the cases) has been eyewitness misidentification. Eyewitness researchers have thus succeeded at identifying the problems and proposing concrete reforms. Indeed, following upon an AP-LS White Paper on the subject (Wells et al., 1998), the U.S. Department of Justice assembled a working group of research

psychologists, prosecutors, police officers, and lawyers, ultimately publishing guidelines for law enforcement on how to minimize eyewitness identification error (Technical Working Group for Eyewitness Evidence, 1999; see Doyle, 2005; Wells et al., 2000). While other problems have been revealed—for example, involving flaws in various forensic sciences (see Faigman, Kaye, Saks, & Sanders, 2002), the number of cases involving confessions—long considered the "gold standard" in evidence—has proved surprising (http://www.innocenceproject.org/).

Wrongful convictions based on false confessions raise serious questions concerning a chain of events by which innocent citizens are judged deceptive in interviews and misidentified for interrogation; waive their rights to silence and to counsel; and are induced into making false narrative confessions that form a sufficient basis for subsequent conviction. This White Paper summarizes much of what we know about this phenomenon. It draws on core psychological principles of influence as well as relevant forensic psychology studies involving an array of methodologies. It identifies various risk factors for false confessions, especially in police interviewing, interrogation, and the elicitation of confessions. It also offers recommendations for reform.

Citing the impact on policy and practice of the eyewitness White Paper, Wiggins and Wheaton (2004) called for a similar consensus-based statement on confessions. Fulfilling this call, the objectives of this White Paper are threefold. The first is to review the state of the science on interviewing and interrogation by bringing together a multidisciplinary group of scholars from three perspectives: (1) clinical psychology (focused on individual differences in personality and psychopathology); (2) experimental psychology (focused on the influence of social, cognitive, and developmental processes); and (3) criminology (focused on the empirical study of criminal justice as well as criminal law, procedure, and legal practice). Our second objective is to identify the dispositional characteristics (e.g., traits associated with *Miranda* waivers, compliance, and suggestibility; adolescence; mental retardation; and psychopathology) and situational-interrogation factors (e.g., prolonged detention and isolation; confrontation; presentations of false evidence; and minimization) that influence the voluntariness and reliability of confessions. Our third objective is to make policy recommendations designed to reduce both the likelihood of police-induced false confessions and the number of wrongful convictions based on these confessions.

BACKGROUND

The pages of American legal history are rich in stories about false confessions. These stories date back to the Salem witch trials of 1692, during which about 50 women confessed to witchcraft, some, in the words of one observer, after being "tyed… Neck and Heels till the Blood was ready to come out of their Noses" (Karlsen, 1989, p. 101). Psychologists' interest as well can be traced to its early days as a science. One hundred years ago, in *On the Witness Stand*, Hugo Munsterberg (1908) devoted an entire chapter to the topic of "Untrue Confessions." In this chapter, he discussed the Salem witch trials, reported on a contemporary Chicago confession that he believed to be false, and sought to explain the causes of this phenomenon (e.g., he used such words as "hope," "fear," "promises," "threats," "suggestion," "calculations," "passive yielding," "shock," "fatigue," "emotional excitement," "melancholia," "auto-hypnosis," "dissociation," and "self-destructive despair").

DNA Exonerations and Discoveries in the U.S.

In 1989, Gary Dotson was the first wrongfully convicted individual to be proven innocent through the then-new science of DNA testing. Almost two decades later, more than 200 individuals have been exonerated by post-conviction DNA testing and released from prison, some from death row. In 15–20% of these cases, police-induced false confessions were involved (Garrett, 2008; www.innocenceproject.org). A disturbing number of these have occurred in high-profile cases, such as New York City's Central Park Jogger case, where five false confessions were taken within a single investigation. In that case, five teenagers confessed during lengthy interrogations to the 1989 brutal assault and rape of a young woman in Central Park. Each boy retracted his statement immediately upon arrest, saying he had confessed because he expected to go home afterward. All the boys were convicted and sent to prison, only to be exonerated in 2002 when the real rapist gave a confession, accurately detailed, that was confirmed by DNA evidence (*People of the State of New York v. Kharey Wise et al.*, 2002).

Post-conviction DNA tests and exonerations have offered a window into the causes of wrongful conviction. Researchers and legal scholars have long documented the problem and its sources of error (Borchard, 1932; Frank & Frank, 1957; see Leo, 2005 for a review). Yet criminal justice officials, commentators, and the public have tended until recently to be highly skeptical of its occurrence, especially in death penalty cases (Bedau & Radelet, 1987). The steady stream of post-conviction DNA exonerations in the last two decades has begun to transform this perception. Indeed, these cases have established the leading causes of error in the criminal justice system to be eyewitness misidentification, faulty forensic science, false informant testimony, and false confessions (Garrett, 2008).

The Problem of False Confessions

A false confession is an admission to a criminal act—usually accompanied by a narrative of how and why the crime occurred—that the confessor did not commit. False confessions are difficult to discover because neither the state nor any organization keeps records of them, and they are not usually publicized. Even if they are discovered, false confessions are hard to establish because of the difficulty of proving the confessor's innocence. The literature on wrongful convictions, however, shows that there are several ways to determine whether a confession is false. Confessions may be deemed false when: (1) it is later discovered that no crime was committed (e.g., the presumed murder victim is found alive, the autopsy on a "shaken baby" reveals a natural cause of death); (2) additional evidence shows it was physically impossible for the confessor to have committed the crime (e.g., he or she was demonstrably elsewhere at the time or too young to have produced the semen found on the victim); (3) the real perpetrator, having no connection to the defendant, is apprehended and linked to the crime (e.g., by intimate knowledge of nonpublic crime details, ballistics, or physical evidence); or (4) scientific evidence affirmatively establishes the confessor's innocence (e.g., he or she is excluded by DNA test results on semen, blood, hair, or saliva).

Drizin and Leo (2004) analyzed 125 cases of proven false confession in the U.S. between 1971 and 2002, the largest sample ever studied. Ninety-three percent of the false confessors were men. Overall, 81% of the confessions occurred in murder cases, followed by rape (8%) and arson (3%). The most common bases for exoneration were the real perpetrator was identified (74%) or that new scientific evidence was discovered (46%). With respect to personal vulnerabilities, the sample was younger than the total population of murderers and rapists: A total of 63% of false confessors were under the age of 25, and 32% were under 18; yet of all persons arrested for murder and rape, only 8 and 16%, respectively, are juveniles (Snyder, 2006). In addition, 22% were mentally retarded, and 10% had a diagnosed mental illness. Surprisingly, multiple false confessions to the same crime were obtained in 30% of the cases, wherein one false confession was used to prompt others. In total, 81% of false confessors in this sample whose cases went to trial were wrongfully convicted.

Although other researchers have also documented false confessions in recent years, there is no known incidence rate, and to our knowledge empirically based estimates have never been published. There are several reasons why an incidence rate cannot be determined. First, researchers cannot identify the universe of false confessions because no governmental or private organization keeps track of this information. As noted earlier, the sample of discovered cases is thus incomplete. Second, even if one could identify a nonrandom set of hotly contested and possibly false confessions, it is often difficult if not impossible as a practical matter to obtain the primary case materials (e.g., police reports; pretrial and trial transcripts; and electronic recordings of the interrogations) needed to determine "ground truth" with sufficient certainty to prove that the confessor is innocent. Also, it is important to note that although most case studies are based in the U.S. and England, proven false confessions have been documented in countries all over the world—including Canada (CBC News, August 10, 2005), Norway (Gudjonsson, 2003), Finland (Santtila, Alkiora, Ekholm, & Niemi, 1999), Germany (Otto, 2006), Iceland (Sigurdsson & Gudjonsson, 2004), Ireland (Inglis, 2004), The Netherlands (Wagenaar, 2002), Australia (Egan, 2006), New Zealand (Sherrer, 2005), China (Kahn, 2005), and Japan (Onishi, 2007).

For estimating the extent of the problem, self-report methods have also been used. Sigurdsson and Gudjonsson (2001) conducted two self-report studies of prison inmates in Iceland and found that 12% claimed to have made a false confession to police at some time in their lives, a pattern that the authors saw as part of the criminal lifestyle. In a more recent study of Icelandic inmates, the rate of self-reported false confessions had increased (Gudjonsson, Sigurdsson, Einarsson, Bragason, & Newton, 2008). Similar studies have been conducted in student samples within Iceland and Denmark. Among those interrogated by police, the self-reported false confession rates ranged from 3.7 to 7% among college and older university students (Gudjonsson, Sigurdsson, Asgeirsdottir, & Sigfusdottir, 2006; Gudjonsson, Sigurdsson, & Einarsson, 2004; Steingrimsdottir, Hreinsdottir, Gudjonsson, Sigurdsson, & Nielsen, 2007; Gudjonsson, Sigurdsson, Bragason, Einarsson, & Valdimarsdottir, 2004). In a North American survey of 631 police investigators, respondents estimated from their own experience that 4.78% of innocent suspects confess during interrogation (Kassin et al., 2007). Retrospective self-reports and observer estimates are subject to various cognitive and motivational biases and should be treated with caution as measures of a false confession rate. In general, however, they reinforce the wrongful conviction data indicating that a small but significant minority of innocent people confess under interrogation.

POLICE INTERROGATIONS IN CONTEXT

The practices of interrogation and the elicitation of confessions are subject to historical, cultural, political, legal, and other contextual influences. Indeed, although this article is focused on confessions to police within in a

criminal justice framework, it is important to note that similar processes occur, involving varying degrees of pressure, within the disparate frameworks of military intelligence gathering and corporate loss-prevention investigations. Focused on criminal justice, we examine American interrogation practices of the past and present; the role played by *Miranda* rights; the admissibility and use of confession evidence in the courts; and current practices not only in the U.S. but in other countries as well.

"Third-Degree" Practices of the Past

From the late nineteenth century through the 1930s, American police occasionally employed "third-degree" methods of interrogation—inflicting physical or mental pain and suffering to extract confessions and other types of information from crime suspects. These techniques ranged from the direct and explicit use of physical assaults to tactics that were both physically and psychologically coercive to lesser forms of duress. Among the most commonly used "third-degree" techniques were physical violence (e.g., beating, kicking, or mauling suspects); torture (e.g., simulating suffocation by holding a suspect's head in water, putting lighted cigars or pokers against a suspect's body); hitting suspects with a rubber hose (which seldom left marks); prolonged incommunicado confinement; deprivations of sleep, food, and other needs; extreme sensory discomfort (e.g., forcing a suspect to stand for hours on end, shining a bright, blinding light on the suspect); and explicit threats of physical harm (for a review, see Leo, 2004). These methods were varied and commonplace (Hopkins, 1931), resulting in large numbers of coerced false confessions (Wickersham Commission Report, 1931).

The use of third-degree methods declined precipitously from the 1930s through the 1960s. They have long since become the exception rather than the rule in American police work, having been replaced by interrogation techniques that are more professional and psychologically oriented. The twin pillars of modern interrogation are behavioral lie-detection methods and psychological interrogation techniques, both of which have been developed and memorialized in interrogation training manuals. By the middle of the 1960s, police interrogation practices had become entirely psychological in nature (Wald, Ayres, Hess, Schantz, & Whitebread, 1967). The President's Commission on Criminal Justice and the Administration of Justice declared in 1967: "Today the third degree is virtually non-existent" (Zimring & Hawkins, 1986, p. 132). Still, as the United States Supreme Court recognized in *Miranda v. Arizona* (1966), psychological interrogation is inherently compelling, if not coercive, to the extent that it relies on sustained pressure, manipulation, trickery, and deceit.

Current Law Enforcement Objectives and Practices in the U.S.

American police typically receive brief instruction on interrogation in the academy and then more sustained and specialized training when promoted from patrol to detective. Interrogation is an evidence-gathering activity that is supposed to occur after detectives have conducted an initial investigation and determined, to a reasonable degree of certainty, that the suspect to be questioned committed the crime.

Sometimes this determination is reasonably based on witnesses, informants, or tangible evidence. Often, however, it is based on a clinical hunch formed during a pre-interrogation interview in which special "behavior-provoking" questions are asked (e.g., "What do you think should happen to the person who committed this crime?") and changes are observed in aspects of the suspect's behavior that allegedly betray lying (e.g., gaze aversion, frozen posture, and fidgety movements). Yet in laboratories all over the world, research has consistently shown that most commonsense behavioral cues are not diagnostic of truth and deception (DePaulo et al., 2003). Hence, it is not surprising as an empirical matter that laypeople on average are only 54% accurate at distinguishing truth and deception; that training does not produce reliable improvement; and that police investigators, judges, customs inspectors, and other professionals perform only slightly better, if at all—albeit with high levels of confidence (for reviews, see Bond & DePaulo, 2006; Meissner & Kassin, 2002; Vrij, 2008).

The purpose of interrogation is therefore not to discern the truth, determine if the suspect committed the crime, or evaluate his or her denials. Rather, police are trained to interrogate only those suspects whose culpability they "establish" on the basis of their initial investigation (Gordon & Fleisher, 2006; Inbau, Reid, Buckley, & Jayne, 2001). For a person under suspicion, this initial impression is critical because it determines whether police proceed to interrogation with a strong presumption of guilt which, in turn, predisposes an inclination to ask confirmatory questions, use persuasive tactics, and seek confessions (Hill, Memon, & McGeorge, 2008; Kassin, Goldstein, & Savitsky, 2003). In short, the single-minded purpose of interrogation is to elicit incriminating statements, admissions, and perhaps a full confession in an effort to secure the conviction of offenders (Leo, 2008).

Designed to overcome the anticipated resistance of individual suspects who are presumed guilty, police interrogation is said to be stress-inducing by design—structured to promote a sense of isolation and increase the anxiety and despair associated with denial relative to confession. To achieve these goals, police employ a number of tactics. As

described in Inbau et al.'s (2001) *Criminal Interrogation and Confessions*, the most influential approach is the so-called Reid technique (named after John E. Reid who, along with Fred Inbau, developed this approach in the 1940s and published the first edition of their manual in 1962). First, investigators are advised to isolate the suspect in a small private room, which increases his or her anxiety and incentive to escape. A nine-step process then ensues in which an interrogator employs both negative and positive incentives. On one hand, the interrogator confronts the suspect with accusations of guilt, assertions that may be bolstered by evidence, real or manufactured, and refuses to accept alibis and denials. On the other hand, the interrogator offers sympathy and moral justification, introducing "themes" that minimize the crime and lead suspects to see confession as an expedient means of escape. The use of this technique has been documented in naturalistic observational studies (Feld, 2006b; Leo, 1996b; Simon, 1991; Wald et al., 1967) and in recent surveys of North American investigators (Kassin et al., 2007; Meyer & Reppucci, 2007).

Miranda Warnings, Rights, and Waivers

One of the U.S. legal system's greatest efforts to protect suspects from conditions that might produce involuntary and unreliable confessions is found in the U.S. Supreme Court decision in *Miranda v. Arizona* (1966). The Court was chiefly concerned with cases in which the powers of the state, represented by law enforcement, threatened to overbear the will of citizen suspects, thus threatening their Constitutional right to avoid self-incrimination.

In *Miranda*, the Court offered a remedy, requiring that police officers had to inform suspects of their rights to remain silent and to the availability of legal counsel prior to confessions. This requirement aimed to strike a balance against the inherently threatening power of the police in relation to the disadvantaged position of the suspect, thus reducing coercion of confessions. In cases involving challenges to the validity of the waiver of rights, courts were to apply a test regarding the admissibility of the confession at trial. Statements made by defendants would be inadmissible if a waiver of the rights to silence and counsel was not made "voluntarily, knowingly, and intelligently." One year after the *Miranda* decision, *In re Gault* (1967) extended these rights and procedures to youth when they faced delinquency allegations in juvenile court.

Forty years later, there is no research evidence that *Miranda* and *Gault* achieved their ultimate objective. Police officers routinely offer the familiar warnings to suspects prior to taking their statements. But research has not unequivocally determined whether confessions became more or less likely, are any more or less reliable, or are occurring in ways that are more or less "voluntary, knowing, and intelligent" than in the years prior to *Miranda*. Several years ago, Paul Cassell, an outspoken critic of *Miranda*, had maintained (based on pre–post studies as well as international comparisons) that the confession and conviction rates have dropped significantly as a direct result of the warning and waiver requirements, thus triggering the release of dangerous criminals (Cassell, 1996a, 1996b; Cassell & Hayman, 1996). Yet others countered that his analysis was based on selective data gathering methods and unwarranted inferences (Donahue, 1998; Feeney, 2000; Thomas & Leo, 2002); that these declines, if real, were insubstantial (Schulhofer, 1996); that four out of five suspects waive their rights and submit to questioning (Leo, 1996a, 1996b); and that the costs to law enforcement were outweighed by social benefits—for example, that *Miranda* has had a civilizing effect on police practices and has increased public awareness of constitutional rights (Leo, 1996c; Thomas, 1996).

In recent years, the U.S. Supreme Court has upheld the basic warning-and-waiver requirement (*Dickerson v. United States*, 2000)—for example, refusing to accept confessions given after a warning that was tactically delayed to produce an earlier inadmissible statement (*Missouri v. Seibert*, 2004). Practically speaking, however, research has suggested that the Court's presumption concerning the protections afforded by *Miranda* warnings is questionable. At minimum, a valid waiver of rights requires that police officers provide suspects an understandable description of their rights and that suspects must understand these warnings to waive them validly. What empirical evidence do we have that *Miranda's* procedural safeguards produce these conditions?

First, the *rights* of which suspects must be informed were clearly defined in Miranda, but the *warnings* were not. The *Miranda* decision included an appendix wherein the Court offered an example of the warnings that were suggested, but police departments were free to devise their own warnings. A recent study examined 560 *Miranda* warning forms used by police throughout the U.S. (Rogers, Harrison, Shuman, Sewell, & Hazelwood, 2007). A host of variations in content and format were identified, and metric analysis of their wording revealed reading-level requirements ranging from third-grade level to the verbal complexity of postgraduate textbooks (see Kahn, Zapf, & Cooper, 2006, for similar results; also see Rogers, Hazelwood, Sewell, Harrison, & Shuman, 2008). Moreover, *Miranda* warning forms varied considerably in what they conveyed. For example, only 32% of the forms told suspects that legal counsel could be obtained without charge. Thus, many warning forms raise serious doubts about the knowing and intelligent waiver of rights by almost any suspect who is "informed" by them.

Second, studies have repeatedly shown that a substantial proportion of adults with mental disabilities, and "average" adolescents below age 16 have impaired understanding of *Miranda* warnings when they are exposed to them. Even adults and youth who understand them sometimes do not grasp their basic implications. Many of these studies have examined actual adult or juvenile defendants, using reliable procedures that allow the quality of an individual's understanding to be scored according to specified criteria. For example, do people after warnings factually understand that "I don't have to talk" and that "I can get an attorney to be here now and during any questioning by police?" To answer this question, respondents have been examined in the relatively benign circumstance of a testing session with a researcher rather than in the context of an accusatory, highly stressful interrogation using standardized *Miranda* warnings that have about an average sixth- to seventh-grade reading level. Thus, the results obtained in these studies represent people's grasp of the *Miranda* warnings under relatively favorable circumstances. Under these conditions, average adults exhibit a reasonably good understanding of their rights (Grisso, 1980, 1981). But studies of adults with serious psychological disorders (Cooper & Zapf, 2008; Rogers, Harrison, Hazelwood, & Sewell, 2007) or with mental retardation (Clare & Gudjonsson, 1991; Everington & Fulero, 1999; Fulero & Everington, 1995; O'Connell, Garmoe, & Goldstein, 2005) have found substantial impairments in understanding of *Miranda* warnings compared to nonimpaired adult defendants.

Many studies have examined adolescents' understanding of *Miranda* warnings, and the results have been very consistent (Abramovitch, Higgins-Biss, & Biss, 1993; Abramovitch, Peterson-Badali, & Rohan, 1995: Colwell et al., 2005; Goldstein, Condie, Kalbeitzer, Osman, & Geier, 2003; Grisso, 1980, 1981; Redlich, Silverman, & Steiner, 2003; Viljoen, Klaver, & Roesch, 2005; Viljoen & Roesch, 2005; Wall & Furlong, 1985). In one comprehensive study, 55% of 430 youth of ages 10–16 misunderstood one or more of the *Miranda* warnings (for example, "That means I can't talk until they tell me to"). Across these studies, the understanding of adolescents ages 15–17 with near-average levels of verbal intelligence tends not to have been inferior to that of adults. But youth of that age with IQ scores below 85, and average youth below age 14, performed much poorer, often misunderstanding two or more of the warnings.

Some studies have shown that many defendants, especially adolescents, who seem to have an adequate factual understanding of *Miranda* warnings, do not grasp their relevance to the situation they are in (e.g., Grisso, 1980, 1981; Viljoen, Zapf, & Roesch, 2007). For example, one may factually understand that "I can have an attorney before and during questioning" yet not know what an attorney is or what role an attorney would play. Others may understand the attorney's role but disbelieve that it would apply in their own situation—as when youth cannot imagine that an adult would take their side against other adults, or when a person with paranoid tendencies believes that any attorney, even his own, would oppose him.

The ability to grasp the relevance of the warnings beyond having a mere factual understanding of what they say is sometimes referred to as having a "rational understanding" or "appreciation" of the warnings. Many states, however, require only a factual understanding of *Miranda* rights for a "knowing and intelligent" waiver (e.g., *People v. Daoud*, 2000). In those states that apply a strict factual understanding standard, youth who technically understand the warnings (e.g., "I can have an attorney to talk to" or "I can stay silent") but harbor faulty beliefs that may distort the significance of these warnings ("An attorney will tell the court whatever I say" or "You have to tell the truth in court, so eventually I'll have to talk if they want me to") are considered capable of having made a valid waiver, even if they have no recognition of the meanings of the words or a distorted view of their implications.

Even among those with adequate understanding, suspects will vary in their capacities to "think" and "decide" about waiving their rights. Whether decision-making capacities are deemed relevant for a "voluntary, knowing, and intelligent" waiver will depend on courts' interpretations of "intelligent" or "voluntary." Several studies have thus examined the decision-making process of persons faced with hypothetical *Miranda* waiver decisions.

Studies of adolescents indicate that youth under age 15 on average perform differently from older adolescents and adults. They are more likely to believe that they should waive their rights and tell what they have done, partly because they are still young enough to believe that they should never disobey authority. Studies have also shown that they are more likely to decide about waiver on the basis of the potential for immediate negative consequences—for example, whether they will be permitted to go home if they waive their rights—rather than considering the longer-range consequences associated with penalties for a delinquency adjudication (Grisso, 1981; Grisso et al., 2003). Young adolescents presented with hypothetical waiver decisions are less likely than older adolescents to engage in reasoning that involves adjustment of their decisions based on the amount of evidence against them or the seriousness of the allegations (Abramovitch, Peterson-Badali, & Rohan, 1995). These results regarding the likelihood of immature decision-making processes are consistent with research on the development of psychosocial abilities of young adolescents in everyday circumstances (Steinberg & Cauffman, 1996) and other

legal contexts (Grisso & Schwartz, 2000; Owen-Kostelnik, Reppucci, & Meyer, 2006).

Other *Miranda* decision-making studies have examined the suggestibility of persons with disabilities (Clare & Gudjonsson, 1995: Everington & Fulero, 1999; O'Connell, Garmoe, & Goldstein, 2005) and adolescents (Goldstein et al., 2003; Redlich et al., 2003; Singh & Gudjonsson, 1992). Suggestibility refers to a predisposition to accept information communicated by others and to incorporate that information into one's beliefs and memories. In general, these studies indicate that persons with mental retardation and adolescents in general are more susceptible to suggestion in the context of making hypothetical waiver decisions, and that greater suggestibility is related to poorer comprehension of the warnings. These results take on special significance in light of observational studies of police behavior when obtaining *Miranda* waiver decisions from adolescents (Feld, 2006a, 2006b) and adults (Leo, 1996b). As described elsewhere in this article, police officers often approach suspects with "friendly" suggestions regarding both the significance of the *Miranda* waiver procedure and their decision. In either case, results indicate that adults with disabilities and adolescents in general are prone to adjust their behaviors and decisions accordingly.

In a formal sense, whether one waives his or her rights voluntarily, knowingly, and intelligently does not have a direct bearing on the likelihood of false confessions (Kassin, 2005; White, 2001). The decision to waive one's rights in a police interrogation does not necessarily lead to a confession, much less a false confession. Nevertheless, research cited earlier regarding the lack of attentiveness of persons with disabilities and adolescents to long-range consequences suggests an increased risk that they would also comply with requests for a confession—whether true of false—to obtain the presumed short-term reward (e.g., release to go home). In addition, some studies have found that poor comprehension of *Miranda* warnings is itself predictive of a propensity to give false confessions (Clare & Gudjonsson, 1995; Goldstein et al., 2003). Sometimes this stems from low intelligence or a desire to comply; at other times it appears to be related to a naïve belief that one's actual innocence will eventually prevail—a belief that is not confined to adolescents or persons with disabilities (Kassin & Norwick, 2004).

Finally, many states require the presence of a parent or other interested adult when youth make decisions about their *Miranda* rights (Oberlander, Goldstein, & Goldstein, 2003). These rules are intended to offer youth assistance in thinking through the decision while recognizing that caretakers cannot themselves waive their children's rights in delinquency or criminal investigations. Studies have shown, however, that the presence of parents at *Miranda* waiver events typically does not result in any advice at all or, when it does, provides added pressure for the youth to waive rights and make a statement (Grisso & Ring, 1979). The presence of parents may be advisable, but it does not offer a remedy for the difficulties youth face in comprehending or responding to requests for a waiver of their rights.

In summary, research suggests that adults with mental disabilities, as well as adolescents, are particularly at risk when it comes to understanding the meaning of *Miranda* warnings. In addition, they often lack the capacity to weigh the consequences of rights waiver, and are more susceptible to waiving their rights as a matter of mere compliance with authority.

Overview of Confession Evidence in the Courts

American courts have long treated confession evidence with both respect and skepticism. Judicial respect for confessions emanates from the power of confession evidence and the critical role that confessions play in solving crimes. The U.S. Supreme Court has recognized that confession evidence is perhaps the most powerful evidence of guilt admissible in court (*Miranda v. Arizona*, 1966)—so powerful, in fact, that "the introduction of a confession makes the other aspects of a trial in court superfluous, and the real trial, for all practical purposes, occurs when the confession is obtained" (*Colorado v. Connelly*, 1986, p. 182 citing McCormick, 1972, p. 316).

Judicial skepticism of confession evidence stems from the historical fact that some law enforcement officers, aware that confession evidence can assure conviction, have abused their power in the interrogation room. As the U.S. Supreme Court stated in *Escobedo v. Illinois* (1964): "We have learned the lesson of history, ancient and modern, that a system of criminal law enforcement which comes to depend on the 'confession' will, in the long run, be less reliable and more subject to abuses than a system which depends on extrinsic evidence independently secured through skillful investigation" (pp. 488–489).

Judicial concern with juror over-reliance on confession evidence gave rise to a series of evolving rules designed to curb possible abuses in the interrogation room, exclude unreliable confessions from trial, and prevent wrongful convictions. These doctrines, which developed both in the common law of evidence and under the Constitution as interpreted by the U.S. Supreme Court, fell into two distinct sets of legal rules: corroboration rules and the voluntariness rules (Ayling, 1984; Leo, Drizin, Neufeld, Hal, & Vatner, 2006).

Corroboration Rules

The corroboration rule, which requires that confessions be corroborated by independent evidence, was the

American take on the English rule known as the *corpus delicti* rule. *Corpus delicti* literally means "body of the crime"—that is, the material substance upon which a crime has been committed" (Garner, 2004, p. 310). The rule was founded at common law in England in the wake of *Perry's Case*, a seventeenth-century case in which a mother and two brothers were convicted and executed based upon a confession to a murder that was later discovered to be false when the supposed murder victim turned up alive (Leo et al., 2006). America's version of *Perry's Case* is the infamous 1819 case of Stephen and Jesse Boorn, two brothers who were convicted and sentenced to death in Manchester, Vermont for the murder of their brother-in-law Russell Colvin. Fortunately for the two men, both of whom had confessed to the killing under intense pressure from authorities, their lawyers located Colvin alive before their hangings took place (Warden, 2005).

In American homicide cases, in response to *Boorn*, the rule came to mean that no individual can be convicted of a murder without proof that a death occurred, namely the existence of a "dead body." As the rule evolved in the courts over time, it was applied to all crimes and required that before a confession could be admitted to a jury, prosecutors had to prove: (1) that a death, injury, or loss had occurred and (2) that criminal agency was responsible for that death, injury, or loss (Leo et al., 2006). The rule was designed to serve three purposes: to prevent false confessions, to provide incentives to police to continue to investigate after obtaining a confession, and to safeguard against the tendency of juries to view confessions as dispositive of guilt regardless of the circumstances under which they were obtained (Ayling, 1984).

The *corpus delicti* rule does not require corroboration that the defendant committed the crime, nor does it demand any proof of the requisite mental state or any other elements of the crime. Moreover, the rule only requires corroboration of the fact that a crime occurred; it does not require that the facts contained in the confession be corroborated. Given the relative ease of establishing the *corpus delicti* in most criminal cases (e.g., producing a dead body in a homicide case and showing that death was not self-inflicted or the result of an accident), and the weight that most jurors attach to confession evidence, prosecutors can still obtain many convictions from unreliable confessions. The rule thus makes it easier in some cases for prosecutors to convict both the guilty and the innocent (Leo et al., 2006).

At the same time, in a certain class of cases, the *corpus delicti* rule may bar the admission of reliable confessions. Because the rule requires that prosecutors prove that there be death or injury resulting from a criminal act, prosecutors may have a hard time getting confessions admitted when the evidence is unclear as to whether any injury had occurred (e.g., child molestation without physical evidence) or whether it resulted from an accident or natural causes as opposed to a criminal act (e.g., child death by smothering or Sudden Infant Death Syndrome; see Taylor, 2005).

For these reasons and others, the rule has been severely criticized. In *Smith v. United States* (1954), the U.S. Supreme Court criticized the *corpus delicti* rule for "serv[ing] an extremely limited function" (p. 153). The Court noted that the rule was originally designed to protect individuals who had confessed to crimes that never occurred but that it does little to protect against the far more frequent problem wherein a suspect confesses to a crime committed by someone else. In short, the rule did "nothing to ensure that a particular defendant was the perpetrator of a crime" (*State v. Mauchley*, 2003, p. 483).

In place of the *corpus delicti* rule, the Supreme Court, in two decisions released on the same day—*Smith* and *Opper v. United States* (1954)—announced a new rule, dubbed the trustworthiness rule, which requires corroboration of the confession itself rather than the fact that a crime occurred. Under the trustworthiness rule, which was adopted by several states, the government may not introduce a confession unless it provides "substantial independent evidence which would tend to establish the trustworthiness of the confession" (*State v. Mauchley*, 2003, p. 48; citing Opper).

In theory, the trustworthiness standard is a marked improvement on the *corpus delicti* rule in its ability to prevent false confessions from entering the stream of evidence at trial. In practice, however, the rule has not worked to screen out false confessions. Because investigators sometimes suggest and incorporate crime details into a suspect's confession, whether deliberately or inadvertently, many false confessions appear highly credible to the secondhand observer. Without an electronic recording of the entire interrogation process, courts are thus left to decide a swearing contest between the suspect and the detective over the source of the details contained within the confession. Moreover, the quantum of corroboration in most jurisdictions that apply the trustworthiness doctrine is very low, allowing many unreliable confessions to go before the jury (Leo et al., 2006).

Rules Prohibiting Involuntary Confession

Until the late eighteenth century, out-of-court confessions were admissible as evidence even if they were the involuntary product of police coercion. In 1783, however, in *The King v. Warrickshall*, an English Court recognized the inherent lack of reliability of involuntary confessions and established the first exclusionary rule:

> Confessions are received in evidence, or rejected as inadmissible, under a consideration whether they are or are not intitled [sic] to credit. A free and voluntary confession is deserving of the highest credit, because it is presumed to flow from the strongest sense of guilt …but a confession forced from the mind by the flattery of hope, or by the torture of fear, comes in so questionable a shape…that no credit ought to be given it; and therefore it should be rejected (*King v. Warrickshall*, 1783, pp. 234–235).

The basis for excluding involuntary confessions in *Warrickshall* was a concern that confessions procured by torture or other forms of coercion must be prohibited because of the risk that such tactics could cause an innocent person to confess. In other words, involuntary confessions were to be prohibited because they were unreliable. Following *Warrickshall*, in the late 1800s, the U.S. Supreme Court adopted this reliability rationale for excluding involuntary confessions in a series of decisions (*Hopt v. Utah*, 1884; *Pierce v. United States*, 1896; *Sparf v. United States*, 1895; *Wilson v. United States*, 1896).

The Supreme Court adopted a second rationale for excluding involuntary confessions in 1897, in *Bram v. United States.* In *Bram*, the Court for the first time linked the voluntariness doctrine to the Fifth Amendment's provision that "no person shall be compelled in any criminal case to be a witness against himself." This privilege against self-incrimination was not rooted in a concern about the reliability of confessions. Rather, its origins were grounded in the rule of *nemo teneturseipsum prodere* ("no one is bound to inform on himself"), a rule dating back to the English ecclesiastical courts which sought to protect individual free will from state intrusion (Leo et al., 2006). The rule of *nemo tenetur*, which was adopted in the colonies and incorporated into the Fifth Amendment, applied only to self-incriminating statements in court, and had never been applied to extrajudicial confessions. By mixing two unrelated voluntariness doctrines, *Bram* rewrote history and provoked considerable confusion by courts and academics alike (Wigmore, 1970). Still, it gave birth to a new basis for excluding involuntary confession evidence—the protection of individual free will.

A third basis for excluding involuntary confessions began to emerge in 1936, in the case of *Brown v. Mississippi*, to deter unfair and oppressive police practices. In *Brown*, three black tenant farmers who had been accused of murdering a white farmer were whipped, pummeled, and tortured until they provided detailed confessions. The Court unanimously reversed the convictions of all three defendants, holding that confessions procured by physical abuse and torture were involuntary. The Court established the Fourteenth Amendment's due process clause as the constitutional test for assessing the admissibility of confessions in state cases. In addition to common law standards, trial judges would now have to apply a federal due process standard when evaluating the admissibility of confession evidence, looking to the "totality of the circumstances" to determine if the confession was 'made freely, voluntarily and without compulsion or inducement of any sort'"(*Haynes v. Washington*, 1963, quoting *Wilson v. United States*, 1896). As such, the Court proposed to consider personal characteristics of the individual suspect (e.g., age, intelligence, mental stability, and prior contact with law enforcement) as well as the conditions of detention and interrogation tactics that were used (e.g., threats, promises, and lies).

This deterrence rationale, implied in *Brown*, was made even more explicit in *Haley v. Ohio*, a case involving a 15-year-old black boy who was questioned throughout the night by teams of detectives, isolated for 3 days, and repeatedly denied access to his lawyer (*Haley v. Ohio*, 1948). While the majority held that the confession was obtained "by means which the law should not sanction" (pp. 600–601), Justice Frankfurter, in his concurrence, went a step further, stating that the confession must be held inadmissible "[t]o remove the inducement to resort to such methods this Court has repeatedly denied use of the fruits of illicit methods" (p. 607).

As these cases suggest, the Supreme Court relied on different and sometimes conflicting rationales for excluding involuntary confessions throughout the twentieth century (Kamisar, 1963; White, 1998). It was not always clear which of the three justifications the Court would rely on when evaluating the voluntariness of a confession. Nevertheless, the Court did appear to designate certain interrogation methods—including physical force, threats of harm or punishment, lengthy or incommunicado questioning, solitary confinement, denial of food or sleep, and promises of leniency—as presumptively coercive and therefore unconstitutional (White, 2001). The Court also considered the individual suspect's personal characteristics, such as age, intelligence, education, mental stability, and prior contact with law enforcement, in determining whether a confession was voluntary. The template of the due process voluntariness test thus involved a balancing of whether police interrogation pressures, interacting with a suspect's personal dispositions, were sufficient to render a confession involuntary (Schulhofer, 1981).

The "totality of the circumstances" test, while affording judges flexibility in practice, has offered little protection to suspects. Without bright lines for courts to follow, and without a complete and accurate record of what transpired during the interrogation process, the end result has been largely unfettered and unreviewable discretion by judges. In practice, when judges apply the test, "they exclude only the most egregiously obtained confessions and then only

haphazardly" (Feld, 1999, p. 118). The absence of a litmus test has also encouraged law enforcement officers to push the envelope with respect to the use of arguably coercive psychological interrogation techniques (Penney, 1998). Unlike its sweeping condemnation of *physical* abuse in *Brown v. Mississippi*, the Court's overall attitude toward *psychological* interrogation techniques has been far less condemnatory. In particular, the Court's attitudes toward the use of maximization and minimization (Kassin & McNall, 1991) and the false evidence ploy and other forms of deception (Kassin & Kiechel, 1996)—techniques that have frequently been linked to false confessions (Kassin & Gudjonsson, 2004)—has been largely permissive. A discussion of some of these cases follows.

Cases Addressing Interrogation Tactics: Maximization and Minimization

Today's interrogators seek to manipulate a suspect into thinking that it is in his or her best interest to confess. To achieve this change in perceptions of subjective utilities, they use a variety of techniques, referred to broadly as "maximization" and "minimization" (Kassin & McNall, 1991). Maximization involves a cluster of tactics designed to convey the interrogator's rock-solid belief that the suspect is guilty and that all denials will fail. Such tactics include making an accusation, overriding objections, and citing evidence, real or manufactured, to shift the suspect's mental state from confident to hopeless. Toward this end, it is particularly common for interrogators to communicate as a means of inducement, implicitly or explicitly, a threat of harsher consequences in response to the suspect's denials (Leo & Ofshe, 2001).

In contrast, minimization tactics are designed to provide the suspect with moral justification and face-saving excuses for having committed the crime in question. Using this approach, the interrogator offers sympathy and understanding; normalizes and minimizes the crime, often suggesting that he or she would have behaved similarly; and offers the suspect a choice of alternative explanations—for example, suggesting to the suspect that the murder was spontaneous, provoked, peer-pressured, or accidental rather than the work of a cold-blooded premeditated killer. As we will see later, research has shown that this tactic communicates by implication that leniency in punishment is forthcoming upon confession.

As the 1897 case of *Bram v. United States* demonstrates, minimization has been part of the arsenal of police interrogation tactics for over a century. In *Bram*, the authorities induced the defendant to confess based on the kind of unspoken promise that anchors the modern psychological interrogation: "Bram, I am satisfied that you killed the captain. But some of us here think you could not have done the crime alone. If you had an accomplice, you should say so, and not have the blame of this horrible crime on your own shoulders" (*Bram v. United States*, 1897, p. 539). This statement contained no direct threats or promises; rather, it combined elements of maximization (the interrogator's stated certainty in the suspect's guilt) and minimization (the suggestion that he will be punished less severely if he confesses and names an accomplice). Using language that condemns the latter, the Supreme Court reversed Bram's conviction, holding that a confession "must not be extracted by any sort of threats or violence, nor obtained by any direct or implied promises, however slight" (pp. 542–543).

Although a strict interpretation of *Bram* seemed to suggest a ban on minimization, courts throughout the twentieth century followed a practice of evading, contradicting, disregarding, and ultimately discarding *Bram* (Hirsch, 2005a). Briefly in the 1960s, it appeared that the Supreme Court was ready to revitalize *Bram* and to apply it broadly to the psychological interrogation techniques taught by such legendary police reformers as Chicago's Fred Inbau and John Reid. Indeed, the landmark case of *Miranda v. Arizona* (1966), described earlier, cited *Bram* and condemned the Reid technique and other tactics that "are designed to put the subject in a psychological state where his story is but an elaboration of what the police purport to know already—that he is guilty" (p. 450). This newfound concern with the impact of psychological interrogation tactics, however, was short lived. In the immediate aftermath of *Miranda*, the Supreme Court adopted a more deferential attitude toward law enforcement in its confession jurisprudence. In particular, *Arizona v. Fulminante* (1991) in dicta may have sounded the death knell for *Bram.* Responding to a party's invocation of *Bram*, the Court casually remarked that "under current precedent [Bram] does not state the standard for determining the voluntariness of a confession" (p. 286). However, White (1997) noted that "as Fulminante's holding indicates, some promises may be sufficient in and of themselves to render a confession involuntary; other promises may or may not be permissible depending upon the circumstances" (p. 150).

Cases Addressing Interrogation Tactics: Trickery and Deception

The false evidence ploy is a controversial tactic occasionally used by police. Not all interrogation trainers approve of this practice (Gohara, 2006), the use of which has been implicated in the vast majority of documented police-induced false confessions (Kassin, 2005). In several pre-*Miranda* voluntariness cases, the U.S. Supreme Court recognized that deception can induce involuntary confessions, although the Court never held that such tactics would automatically invalidate a confession. In *Leyra v. Denno* (1954), for

example, Leyra asked to see a physician because he was suffering from sinus problems and police brought in a psychiatrist who posed as a general physician. The Supreme Court held that the "subtle and suggestive" questioning by the psychiatrist amounted to a continued interrogation of the suspect without his knowledge. This deception and other circumstances of the interrogation rendered Leyra's confession involuntary. Similarly, in *Spano v. New York* (1959), the suspect considered one of the interrogating officers to be a friend. The Court held that the officer's false statements, in which he suggested that the suspect's actions might cost the officer his job, were a key factor in rendering the resulting confession involuntary. In *Miranda v. Arizona* (1966), the Supreme Court discussed the use of trickery and deception and noted that the deceptive tactics recommended in standard interrogation manuals fostered a coercive environment. Again, the Court did not specifically prohibit such tactics, choosing instead to offer suspects some relief from the coercive effect by empowering them with rights which could be used to bring interrogation to a halt. The criticism of deception may have fanned hopes that the Court would deal a more direct blow to this controversial tactic in future cases. But such hopes were quickly quashed.

Three years later, in *Frazier v. Cupp* (1969), the Supreme Court addressed interrogation trickery and issued a decision that to this day has been interpreted by police and the courts as a green light to deception. In *Frazier*, police used a standard false evidence ploy—telling Frazier that another man whom he and the victim had been seen with on the night of the crime had confessed to their involvement. The investigating detective also used minimization, suggesting to Frazier that he had started a fight with the victim because the victim made homosexual advances toward him. Despite the use of these deceptive tactics, the Court held that Frazier's confession was voluntary. This ruling established that police deception by itself is not sufficient to render a confession involuntary. Rather, according to *Frazier*, deception is but one factor among many that a court should consider. Some state courts have distinguished between mere false assertions, which are permissible, and the fabrication of reports, tapes, and other evidence—which is not. In the Florida case of *State v. Cayward* (1989), the defendant's confession was suppressed because police had typed up a phony crime laboratory report that placed Cayward's DNA on the victim. However, the court's concern was not that the manufactured evidence might prompt an innocent person to confess but that it might find its way into court as evidence. Similarly, New Jersey confessions were suppressed when produced by a fake, staged audiotape of an alleged eyewitness account (*State v. Patton*, 1993) and a fake crime lab report identifying the suspect's DNA at the crime scene (*State v. Chirokovskcic*, 2004). This is where the law remains today despite numerous cautionary notes from academics and researchers on the use of deception (Gohara, 2006; Gudjonsson, 2003; Kassin, 2005; Kassin & Gudjonsson, 2004; Skolnick & Leo, 1992; but see Grano, 1994; Slobogin, 2007).

Practices in England

Interrogations and confession evidence are regulated in England and Wales by the Police and Criminal Evidence Act of 1984 (PACE; Home Office, 1985), which became effective in January 1986. The Act is supplemented by five Codes of Practice, referred to as Codes A (on stop and search), B (entry and searches of premises), C (detention and questioning of suspects), D (on identification parades), and E (tape recording of interviews). The Codes provide guidance to police officers concerning procedures and the appropriate treatment of suspects. Code C is particularly relevant to issues surrounding "fitness to be interviewed," as it provides guidance "on practice for the detention, treatment and questioning of persons by police officers" (Home Office, 2003, p. 47).

The most important interview procedures set out in PACE and its Codes of Practice are that: Suspects who are detained at a police station must be informed of their legal rights; in any 24-h period the detainee must be allowed a continuous period of rest of at least 8 hours; detainees who are vulnerable in terms of their age or mental functioning should have access to a responsible adult (known as an 'appropriate adult'), whose function is to give advice, further communication, and ensure that the interview is conducted properly and fairly; and all interviews shall be electronically recorded.

Compared to the approach typically taken in the U.S. (e.g., using the Reid technique), investigative interview practices in England are less confrontational. Williamson (2007) discussed in detail how psychological science has influenced the training of police officers and their interviewing practice, making it fairer and more transparent. Prior to 1992, investigators in Britain received no formal training and the chief purpose of interviewing suspects was to obtain confessions. Following some high-profile miscarriages of justice, such as the "Guildford Four" and "Birmingham Six," the Association of Chief Police Officers for England and Wales (ACPO) published the first national training program for police officers interviewing both suspects and witnesses. This new approach was developed through a collaboration of police officers, psychologists, and lawyers. The mnemonic PEACE was used to describe the five distinct parts of the new interview approach ("Preparation and Planning," "Engage and Explain," "Account," "Closure," and "Evaluate"). The theory underlying this approach, particularly in cases of witnesses, victims, and cooperative suspects, can be traced

to Fisher and Geiselman's (1992) work on the "Cognitive Interview" (Milne & Bull, 1999; for research evidence, see Clarke & Milne, 2001; Williamson, 2006). Recent analyses of police–suspect interviews in England have revealed that the confrontation-based tactics of maximization and minimization are in fact seldom used (Soukara, Bull, Vrij, Turner, & Cherryman, in press; Bull & Soukara, 2009).

POLICE-INDUCED FALSE CONFESSIONS

As described earlier, the process of interrogation is designed to overcome the anticipated resistance of individual suspects who are presumed guilty and to obtain legally admissible confessions. The single-minded objective, therefore, is to increase the anxiety and despair associated with denial and reduce the anxiety associated with confession. To achieve these goals, police employ a number of tactics that involve isolating the suspect and then employing both negative and positive incentives. On the negative side, interrogators confront the suspect with accusations of guilt, assertions that are made with certainty and often bolstered by evidence, real or manufactured, and a refusal to accept alibis and denials. On the positive side, interrogators offer sympathy and moral justification, introducing "themes" that normalize and minimize the crime and lead suspects to see confession as an expedient means of escape. In this section, we describe some core principles of psychology relevant to understanding the suspect's decision making in this situation; then we describe the problem of false confessions and the situational and dispositional factors that put innocent people at risk.

Types of False Confessions

Although it is not possible to calculate a precise incidence rate, it is clear that false confessions occur in different ways and for different reasons. Drawing on the pages of legal history, and borrowing from social-psychological theories of influence, Kassin and Wrightsman (1985) proposed a taxonomy that distinguished among three types of false confession: voluntary, coerced-compliant, and coerced-internalized (see also Kassin, 1997; Wrightsman & Kassin, 1993). This classification scheme has provided a useful framework for the study of false confessions and has since been used, critiqued, extended, and refined by others (Gudjonsson, 2003; Inbau et al., 2001; McCann, 1998; Ofshe & Leo, 1997a, 1997b).

Voluntary False Confessions

Sometimes innocent people have claimed responsibility for crimes they did not commit without prompting or pressure from police. This has occurred in several high-profile cases. After Charles Lindbergh's infant son was kidnapped in 1932, 200 people volunteered confessions. When "Black Dahlia" actress Elizabeth Short was murdered and her body mutilated in 1947, more than 50 men and women confessed. In the 1980s, Henry Lee Lucas in Texas falsely confessed to hundreds of unsolved murders, making him the most prolific serial confessor in history. In 2006, John Mark Karr volunteered a confession, replete with details, to the unsolved murder of young JonBenet Ramsey. There are a host of reasons why people have volunteered false confessions—such as a pathological desire for notoriety, especially in high-profile cases reported in the news media; a conscious or unconscious need for self-punishment to expiate feelings of guilt over prior transgressions; an inability to distinguish fact from fantasy due to a breakdown in reality monitoring, a common feature of major mental illness; and a desire to protect the actual perpetrator—the most prevalent reason for false admissions (Gudjonsson et al., 2004; Sigurdsson & Gudjonsson, 1996, 1997, 2001). Radelet, Bedau, and Putnam (1992) described one case in which an innocent man confessed to a murder to impress his girlfriend. Gudjonsson (2003) described another case in which a man confessed to murder because he was angry at police for a prior arrest and wanted to mislead them in an act of revenge.

Compliant False Confessions

In contrast to voluntary false confessions, compliant false confessions are those in which suspects are induced through interrogation to confess to a crime they did not commit. In these cases, the suspect acquiesces to the demand for a confession to escape a stressful situation, avoid punishment, or gain a promised or implied reward. Demonstrating the form of influence observed in classic studies of social influence (e.g., Asch, 1956; Milgram, 1974), this type of confession is an act of mere public compliance by a suspect who knows that he or she is innocent but bows to social pressure, often coming to believe that the short-term benefits of confession relative to denial outweigh the long-term costs. Based on a review of a number of cases, Gudjonsson (2003) identified some very specific incentives for this type of compliance—such as being allowed to sleep, eat, make a phone call, go home, or, in the case of drug addicts, feed a drug habit. The desire to bring the interview to an end and avoid additional confinement may be particularly pressing for people who are young, desperate, socially dependent, or phobic of being locked up in a police station. The pages of legal history are filled with stories of compliant false confessions. In the 1989 Central Park jogger case described earlier, five

teenagers confessed after lengthy interrogations. All immediately retracted their confessions but were convicted at trial and sent to prison—only to be exonerated 13 years later (*People of the State of New York v. Kharey Wise et al.*, 2002).

Internalized False Confessions

In the third type of false confession, innocent but malleable suspects, told that there is incontrovertible evidence of their involvement, come not only to capitulate in their behavior but also to believe that they may have committed the crime in question, sometimes confabulating false memories in the process. Gudjonsson and MacKeith (1982) argued that this kind of false confession occurs when people develop such a profound distrust of their own memory that they become vulnerable to influence from external sources. Noting that the innocent confessor's belief is seldom fully internalized, Ofshe and Leo (1997a) have suggested that the term "persuaded false confession" is a more accurate description of the phenomenon. The case of 14-year-old Michael Crowe, whose sister Stephanie was stabbed to death in her bedroom, illustrates this type of persuasion. After a series of interrogation sessions, during which time police presented Crowe with compelling false physical evidence of his guilt, he concluded that he was a killer, saying: "I'm not sure how I did it. All I know is I did it." Eventually, he was convinced that he had a split personality—that "bad Michael" acted out of a jealous rage while "good Michael" blocked the incident from memory. The charges against Crowe were later dropped when a drifter in the neighborhood that night was found with Stephanie's blood on his clothing (Drizin & Colgan, 2004).

Relevant Core Principles of Psychology

Earlier we reviewed the tactics of a modern American interrogation and the ways in which the U.S. Supreme Court has treated these tactics with respect to the voluntariness and admissibility of the confessions they elicit. As noted, the goal of interrogation is to alter a suspect's decision making by increasing the anxiety associated with denial and reducing the anxiety associated with confession (for an excellent description of a suspect's decision-making process in this situation, see Ofshe & Leo, 1997b).

Long before the first empirical studies of confessions were conducted, the core processes of relevance to this situation were familiar to generations of behavioral scientists. Dating back to Thorndike's (1911) law of effect, psychologists have known that people are highly responsive to reinforcement and subject to the laws of conditioning, and that behavior is influenced more by perceptions of short-term than long-term consequences. Of distal relevance to a psychological analysis of interrogation are the thousands of operant animal studies of reinforcement schedules, punishment, appetitive, avoidance, and escape learning, as well as behavioral modification applications in clinics, schools, and workplaces. Looking through this behaviorist lens, it seems that interrogators have sometimes shaped suspects to confess to particular narrative accounts of crimes like they were rats in a Skinner box (see Herrnstein, 1970; Skinner, 1938).

More proximally relevant to an analysis of choice behavior in the interrogation room are studies of human decision making in a behavioral economics paradigm. A voluminous body of research has shown that people make choices that they think will maximize their well-being given the constraints they face, making the best of the situation they are in—what Herrnstein has called the "matching law" (Herrnstein, Rachlin, & Laibson, 1997). With respect to a suspect's response to interrogation, studies on the discounting of rewards and costs show that people tend to be impulsive in their orientation, preferring outcomes that are immediate rather than delayed, with delayed outcomes depreciating over time in their subjective value (Rachlin, 2000). In particular, animals and humans clearly prefer delayed punishment to immediate aversive stimulation (Deluty, 1978; Navarick, 1982). These impulsive tendencies are especially evident in juvenile populations and among cigarette smokers, alcoholics, and other substance users (e.g., Baker, Johnson, & Bickel, 2003; Bickel & Marsch, 2001; Bickel, Odum, & Madden, 1999; Kollins, 2003; Reynolds, Richards, Horn, & Karraker, 2004).

Rooted in the observation that people are inherently social beings, a second set of core principles is that individuals are highly vulnerable to influence from change agents who seek their compliance. Of direct relevance to an analysis of interrogation are the extensive literatures on attitudes and persuasion (Petty & Cacioppo, 1986), informational and normative influences (e.g., Asch, 1956; Sherif, 1936), the use of sequential request strategies, as in the foot-in-the-door effect (Cialdini, 2001), and the gradual escalation of commands, issued by figures of authority, to effectively obtain self- and other-defeating acts of obedience (Milgram, 1974). Conceptually, Latane's (1981) social impact theory provides a predictive mathematical model that can account for the influence of police interrogators—who bring *power*, *proximity*, and *number* to bear on their exchange with suspects (for a range of social psychological perspectives on interrogation, see Bem, 1966; Davis & O'Donahue, 2004; Zimbardo, 1967).

A third set of core principles consists of the "seven sins of memory" that Schacter (2001) identified from cognitive and neuroscience research—a list that includes memory transience, misattribution effects, suggestibility, and bias.

When Kassin and Wrightsman (1985) first identified coerced-internalized or coerced-persuaded false confessions, they were puzzled. At the time, existing models of memory could not account for the phenomenon whereby innocent suspects would come to internalize responsibility for crimes they did not commit and confabulate memories about these nonevents. These cases occur when a suspect is dispositionally or situationally rendered vulnerable to manipulation and the interrogator then misrepresents the evidence, a common ploy. In light of a now extensive research literature on misinformation effects and the creation of illusory beliefs and memories (e.g., Loftus, 1997, 2005), experts can now better grasp the process by which people come to accept guilt for a crime they did not commit as well as the conditions under which this may occur (see Kassin, 2008).

Situational Risk Factors

Among the situational risk factors associated with false confessions, three will be singled out: interrogation time, the presentation of false evidence, and minimization. These factors are highlighted because of the consistency in which they appear in cases involving proven false confessions.

Physical Custody and Isolation

To ensure privacy and control, and to increase the stress associated with denial in an incommunicado setting, interrogators are trained to remove suspects from their familiar surroundings and question them in the police station—often in a special interrogation room. Consistent with guidelines articulated by Inbau et al. (2001), most interrogations are brief. Observational studies in the U.S. and Britain have consistently shown that the vast majority of interrogations last approximately from 30 minutes up to 2 hours (Baldwin, 1993; Irving, 1980; Leo, 1996b; Wald et al., 1967). In a recent self-report survey, 631 North American police investigators estimated from their experience that the mean length of a typical interrogation is 1.60 hours. Consistent with cautionary advice from Inbau et al. (2001) against exceeding 4 hours in a single session, these same respondents estimated on average that their longest interrogations lasted 4.21 hours (Kassin et al., 2007). Suggesting that time is a concern among practitioners, one former Reid technique investigator has defined interrogations that exceed 6 hours as "coercive" (Blair, 2005). In their study of 125 proven false confessions, Drizin and Leo (2004) thus found, in cases in which interrogation time was recorded, that 34% lasted 6–12 hours, that 39% lasted 12–24 hours, and that the mean was 16.3 hours.

It is not particularly surprising that false confessions tend to occur after long periods of time—which indicates a dogged persistence in the face of denial. The human needs for belonging, affiliation, and social support, especially in times of stress, are a fundamental human motive (Baumeister & Leary, 1996). People under stress seek desperately to affiliate with others for the psychological, physiological, and health benefits that social support provides (Rofe, 1984; Schachter, 1959; Uchino, Cacioppo, & Kiecolt-Glaser, 1996). Hence, prolonged isolation from significant others in this situation constitutes a form of deprivation that can heighten a suspect's distress and incentive to remove himself or herself from the situation. Depending on the number of hours and conditions of interrogation, sleep deprivation may also become a source of concern. Controlled laboratory experiments have shown that sleep deprivation, which may accompany prolonged periods of isolation, can heighten susceptibility to influence and impair decision-making abilities in complex tasks. The range of effects is varied, with studies showing that sleep deprivation markedly impairs the ability to sustain attention, flexibility of thinking, and suggestibility in response to leading questions (Blagrove, 1996; for a review, see Harrison & Horne, 2000). This research literature is not all based in the laboratory. For example, performance decrements have been observed in medical interns (e.g., Veasey, Rosen, Barzansky, Rosen, & Owens, 2002; Weinger & Ancoli-Israel, 2002)—as when sleep deprivation increased the number of errors that resident surgeons made in a virtual reality surgery simulation (Taffinder, McManus, Gul, Russell, & Darzi, 1998). Also demonstrably affected are motorists (Lyznicki, Doege, Davis, & Williams, 1998) and F-117 fighter pilots (Caldwell, Caldwell, Brown, & Smith, 2004). Combining the results in a meta-analysis, Pilcher and Huffcut (1996) thus concluded that: "overall sleep deprivation strongly impairs human functioning." The use of sleep deprivation in interrogation is hardly a novel idea. In *Psychology and Torture*, Suedfeld (1990) noted that sleep deprivation is historically one of the most potent methods used to soften up prisoners of war and extract confessions from them. Indeed, Amnesty International reports that most torture victims interviewed report having been deprived of sleep for 24 hours or more.

Presentations of False Evidence

Once suspects are isolated, interrogators, armed with a strong presumption of guilt, seek to communicate that resistance is futile. This begins the confrontation process, during which interrogators exploit the psychology of inevitability to drive suspects into a state of despair. Basic research shows that once people see an outcome as inevitable, cognitive and motivational forces conspire to

promote their acceptance, compliance with, and even approval of the outcome (Aronson, 1999). In the case of interrogation, this process also involves interrupting the suspect's denials, overcoming objections, and refuting alibis. At times, American police will overcome a suspect's denials by presenting supposedly incontrovertible evidence of his or her guilt (e.g., a fingerprint, blood or hair sample, eyewitness identification, or failed polygraph)—even if that evidence does not exist. In the U.S., it is permissible for police to outright lie to suspects about the evidence (*Frazier v. Cupp*, 1969)—a tactic that is recommended in training (Inbau et al., 2001), and occasionally used (Kassin et al., 2007; Leo, 1996b).

Yet basic psychological research warns of the risk of this manipulation. Over the years, across a range of subdisciplines, basic research has revealed that misinformation renders people vulnerable to manipulation. To cite but a few highly recognized classics in the field, experiments have shown that presentations of false information—via confederates, witnesses, counterfeit test results, bogus norms, false physiological feedback, and the like—can substantially alter subjects' visual judgments (Asch, 1956; Sherif, 1936), beliefs (Anderson, Lepper, & Ross, 1980), perceptions of other people (Tajfel, Billig, Bundy, & Flament, 1971), behaviors toward other people (Rosenthal & Jacobson, 1968), emotional states (Schachter & Singer, 1962), physical attraction (Valins, 1966), self-assessments (Crocker, Voelkl, Testa, & Major, 1991), memories for observed and experienced events (Loftus, 2005), and even certain medical outcomes, as seen in studies of the placebo effect (Brown, 1998; Price, Finniss, & Benedetti, 2008). Scientific evidence for human malleability in the face of misinformation is broad and pervasive.

The forensic literature on confessions reinforces and extends this classic point, indicating that presentations of false evidence can lead people to confess to crimes they did not commit. This literature is derived from two sources of information. First, studies of actual cases reveal that the false evidence ploy, which is not permitted in Great Britain and most other European nations, is found in numerous wrongful convictions in the U.S., including DNA exonerations, in which there were confessions in evidence (Drizin & Leo, 2004; Leo & Ofshe, 1998). That this tactic appears in proven false confession cases makes sense. In self-report studies, actual suspects state that the reason they confessed is that they perceived themselves to be trapped by the weight of evidence (Gudjonsson & Sigurdsson, 1999; Moston, Stephenson, & Williamson, 1992).

Concerns about the polygraph are illustrative in this regard. Although it is best known for its use as a lie-detector test, and has value as an investigative tool, posttest "failure" feedback is often used to pressure suspects and can prompt false confessions. This problem is so common that Lykken (1998) coined the term "fourth degree" to describe the tactic (p. 235), and the National Research Council Committee to Review the Scientific Evidence on the Polygraph (2003) warned of the risk of polygraph-induced false confessions. In a laboratory demonstration that illustrates the point, Meyer and Youngjohn (1991) elicited false confessions to the theft of an experimenter's pencil from 17% of subjects told that they had failed a polygraph test on that question.

The second source of evidence is found in laboratory experiments that have tested the causal hypothesis that false evidence leads innocent people to confess to prohibited acts they did not commit. In one study, Kassin and Kiechel (1996) accused college students typing on a keyboard of causing the computer to crash by pressing a key they were instructed to avoid. Despite their innocence and initial denials, subjects were asked to sign a confession. In some sessions but not others, a confederate said she witnessed the subject hit the forbidden key. This false evidence nearly doubled the number of students who signed a written confession, from 48 to 94%.

Follow-up studies have replicated this effect to the extent that the charge was plausible (Horselenberg et al., 2006; Klaver, Lee, & Rose, 2008), even when the confession was said to bear a financial or other consequence (Horselenberg, Merckelbach, & Josephs, 2003; Redlich & Goodman, 2003), and even among informants who are pressured to report on a confession allegedly made by another person (Swanner, Beike, & Cole, in press). The effect has been particularly evident among stress-induced males (Forrest, Wadkins, & Miller, 2002) and children and juveniles who tend to be both more compliant and suggestible than adults (Candel, Merckelbach, Loyen, & Reyskens, 2005; Redlich & Goodman, 2003). Using a completely different paradigm, Nash and Wade (2009) used digital editing software to fabricate video evidence of participants in a computerized gambling experiment "stealing" money from the "bank" during a losing round. Presented with this false evidence, all participants confessed—and most internalized the belief in their own guilt.

One needs to be cautious in generalizing from laboratory experiments. Yet numerous false confession cases have featured the use and apparent influence of the false evidence ploy. In one illustrative case, in 1989, 17-year-old Marty Tankleff was accused of murdering his parents despite the complete absence of evidence against him. Tankleff vehemently denied the charges for several hours—until his interrogator told him that his hair was found within his mother's grasp, that a "humidity test" indicated he had showered (hence, the presence of only one spot of blood on his shoulder), and that his hospitalized father had emerged from his coma to say that Marty was his assailant—all of which were untrue (the father never

regained consciousness and died shortly thereafter). Following these lies, Tankleff became disoriented and confessed. Solely on the basis of that confession, Tankleff was convicted, only to have his conviction vacated and the charges dismissed 19 years later (Firstman & Salpeter, 2008; Lambert, 2008).

Minimization: Promises Implied But Not Spoken

In addition to thrusting the suspect into a state of despair by the processes of confrontation, interrogators are trained to minimize the crime through "theme development," a process of providing moral justification or face-saving excuses, making confession seem like an expedient means of escape. Interrogators are thus trained to suggest to suspects that their actions were spontaneous, accidental, provoked, peer-pressured, drug-induced, or otherwise justifiable by external factors. In the Central Park jogger case, every boy gave a false confession that placed his cohorts at center stage and minimized his own involvement (e.g., 16-year-old Kharey Wise said he felt pressured by peers)—and each said afterward that he thought he would go home after confessing based on statements made by police.

Minimization tactics that imply leniency may well lead innocent people who feel trapped to confess. Two core areas of psychology compel this conclusion. The first concerns the principle of reinforcement. As noted earlier, generations of basic behavioral scientists, dating back to Thorndike (1911), and formalized by Skinner (1938), have found that people are highly responsive to reinforcement and the perceived consequences of their behavior. More recent studies of human decision making have added that people are particularly influenced by outcomes that are immediate rather than delayed, the latter depreciating over time in their subjective value (Rachlin, 2000). The second core principle concerns the cognitive psychology of pragmatic implication. Over the years, researchers have found that when people read text or hear speech, they tend to process information "between the lines" and recall not what was stated per se, but what was *pragmatically implied.* Hence, people who read that "The burglar goes to the house" often mistakenly recall later that the burglar actually broke into the house; those who hear that "The flimsy shelf weakened under the weight of the books" often mistakenly recall that the shelf actually broke (Chan & McDermott, 2006; Harris & Monaco, 1978; Hilton, 1995). These findings indicate that pragmatic inferences can change the meaning of a communication, leading listeners to infer something that is "neither explicitly stated nor necessarily implied" (Brewer, 1977).

Taken together, basic research showing that people are highly influenced by perceived reinforcements and that people process the pragmatic implications of a communication suggests the possibility that suspects infer leniency in treatment from minimizing remarks that depict the crime as spontaneous, accidental, pressured by others, or otherwise excusable—even in the absence of an explicit promise. To test this hypothesis, Kassin and McNall (1991) had subjects read a transcript of an interrogation of a murder suspect (the text was taken from an actual New York City interrogation). The transcripts were edited to produce three versions in which the detective made a contingent explicit promise of leniency, used the technique of minimization by blaming the victim, or did not use either technique. Subjects read one version and then estimated the sentence that they thought would be imposed on the suspect. The result: As if explicit promises had been made, minimization lowered sentencing expectations compared to conditions in which no technique was used.

More recently, researchers have found that minimization can also lead innocent people to confess. Using the computer crash paradigm described earlier, Klaver, Lee, and Rose (2008) found that minimization remarks significantly increased the false confession rate when the accusation concerning the forbidden key press was plausible. Russano, Meissner, Kassin, and Narchet (2005) devised a newer laboratory paradigm to not only assess the behavioral effects of minimization but to assess the diagnosticity of the resulting confession (a technique has "diagnosticity" to the extent that it increases the ratio of true to false confessions). In their study, subjects were paired with a confederate for a problem-solving study and instructed to work alone on some problems and jointly on others. In the *guilty* condition, the confederate sought help on a problem that was supposed to be solved alone, inducing a violation of the experimental prohibition. In the *innocent* condition, the confederate did not make this request to induce the crime. The experimenter soon "discovered" a similarity in their solutions, separated the subject and confederate, and accused the subject of cheating. The experimenter tried to get the subject to sign an admission by overtly promising leniency (a deal in which research credit would be given in exchange for a return session without penalty), making minimizing remarks ("I'm sure you didn't realize what a big deal it was"), using both tactics, or using no tactics. Overall, the confession rate was higher among guilty subjects than innocent, when leniency was promised than when it was not, and when minimization was used than when it was not. Importantly, diagnosticity—defined as the rate of true confessions to false confessions—was highest at 7.67 when no tactics were used (46% of guilty suspects confessed vs. only 6% of innocents) and minimization—just like an explicit offer of leniency—reduced diagnosticity to 4.50 by increasing not only the rate of true confessions (from 46 to 81%) but even more so the rate of false confessions (which tripled from 6 to 18%). In short,

minimization provides police with a loophole in the rules of evidence by serving as the implicit but functional equivalent to a promise of leniency (which itself renders a confession inadmissible). The net result is to put innocents at risk to make false confessions.

It is important to note that minimization and the risk it engenders is not a mere laboratory phenomenon. Analyzing more than 125 electronically recorded interrogations and transcripts, Ofshe and Leo (1997a, 1997b) found that police often use techniques that serve to communicate promises and threats through pragmatic implication. These investigators focused specifically on what they called *high-end inducements*—appeals that communicate to a suspect that he or she will receive less punishment, a lower prison sentence, or some form of prosecutorial or judicial leniency upon confession and/or a higher charge or longer prison sentence in the absence of confession. In some homicide cases, for example, interrogators suggested that if the suspect admits to the killing it would be framed as unintentional, as an accident, or as an act of justifiable self-defense—not as premeditated cold-blooded murder, the portrayal that would follow from continued denial. This is a variant of the "maximization"/"minimization" technique described by Kassin and McNall (1991), which communicates through pragmatic implication that the suspect will receive more lenient treatment if he or she confesses but harsher punishment if he or she does not.

Dispositional Risk Factors

In any discussion of dispositional risk factors for false confession, the two most commonly cited concerns are a suspect's age (i.e., juvenile status) and mental impairment (i.e., mental illness, mental retardation). These common citations are because of the staggering overrepresentation of these groups in the population of proven false confessions. For example, of the first 200 DNA exonerations in the U.S., 35% of the false confessors were 18 years or younger and/or had a developmental disability. In their sample of wrongful convictions, Gross, Jacoby, Matheson, Montgomery, and Patel (2005) found that 44% of the exonerated juveniles and 69% of exonerated persons with mental disabilities were wrongly convicted because of false confessions.

Adolescence and Immaturity

There is strong evidence that juveniles are at risk for involuntary and false confessions in the interrogation room (for reviews see Drizin & Colgan, 2004; Owens-Kostelnik, Reppucci, & Meyer, 2006; Redlich, 2007; Redlich & Drizin, 2007; Redlich, Silverman, Chen, & Steiner, 2004). Juveniles are over represented in the pool of identified false confession cases: 35% of the proven false confessors in the Drizin and Leo (2004) sample were younger than age 18, and within this sample of juveniles, 55% were aged 15 or younger. Comparatively, of all persons arrested for murder and rape, only 8 and 16%, respectively, are juveniles (Snyder, 2006). Numerous high-profile cases, such as the Central Park Jogger case (Kassin, 2002), have demonstrated the risks of combining young age, and the attributes that are associated with it (e.g., suggestibility, heightened obedience to authority, and immature decision-making abilities), and the psychologically oriented interrogation tactics described earlier. Hence, Inbau et al. (2001) concede that minors are at special risk for false confession and advise caution when interrogating a juvenile. Referring to the presentation of fictitious evidence, for example, they note: "This technique should be avoided when interrogating a youthful suspect with low social maturity" (p. 429).

The field of developmental psychology was born over a century ago in the influential writings of James Baldwin, Charles Darwin, G. Stanley Hall, and William Stern (see Parke, Ornstein, Rieser, & Zahn-Waxler, 1994). Since that time, basic research has shown that children and adolescents are cognitively and psychosocially less mature than adults—and that this immaturity manifests in impulsive decision making, decreased ability to consider long-term consequences, engagement in risky behaviors, and increased susceptibility to negative influences. Specifically, this body of research indicates that early adolescence marks the onset of puberty, heightening emotional arousability, sensation seeking, and reward orientation; that mid-adolescence is a period of increased vulnerability to risk-taking and problems in affect and behavior; and that late adolescence is a period in which the frontal lobes continue to mature, facilitating regulatory competence and executive functioning (for reviews, see Steinberg, 2005; Steinberg & Morris, 2001). Recent neurological research on brain development dovetails with findings from behavioral studies. Specifically, these studies have shown continued maturation during adolescence in the limbic system (emotion regulation) and in the prefrontal cortex (planning and self-control), with gray matter thinning and white matter increasing (Steinberg, 2007).

The developmental capabilities and limitations of adolescents are highly relevant to behavior in the interrogation room. In *Roper v. Simmons* (2005), Justice Kennedy cited three general differences between juveniles and adults in support of the Court's reasoning for abolishing the death penalty for juveniles. First, he addressed the lessened maturity and responsibility of juveniles compared to adults with specific mention to the 18-year bright-line requirements for marriage without parental consent, jury duty, and voting. Second, Justice Kennedy noted that "juveniles are

more vulnerable or susceptible to negative influences and outside pressures, including peer pressure" (p. 15). Consistent with this portrait, Drizin and Leo (2004) found in their sample of false confessions that several involved two or more juveniles (out of 38 multiple false confession cases, half involved juveniles). In recommending that police "play one [suspect] against the other," Inbau et al. (2001) note that this tactic may be especially effective on young, first-time offenders (pp. 292–293). Third, Justice Kennedy recognized that juveniles' personality or "character" is not as well developed as adults. In light of the volatility of adolescence, it is interesting that Inbau et al. (2001) also suggest "themes" for confession that exploit a juvenile's restless energy, boredom, low resistance to temptation, and lack of supervision.

Drawing on basic principles of developmental psychology, there is now a wealth of forensically oriented research indicating that juveniles—suspects, defendants, and witnesses—have age-related limitations of relevance to the legal system in comparison to adults. For example, individuals younger than 16 years generally have impairments in adjudicative competence (e.g., the ability to help in one's own defense) and comprehension of legal terms (Grisso et al., 2003; Saywitz, Nathanson, & Snyder, 1993). In a subset of studies particularly germane to interrogations, several researchers employing a range of methodologies have shown that the risk of false confession is heightened during childhood and adolescence relative to adulthood. Of particular note, as described earlier, juveniles are more likely than adults to exhibit deficits in their understanding and appreciation of the *Miranda* rights that were explicitly put into place to protect people subject to "inherently coercive" interrogations (see Grisso, 1981; Redlich et al., 2003).

In the first set of studies, laboratory-based experiments have examined juveniles' responses in mock crimes and interrogations. Using the Kassin and Kiechel (1996) computer crash paradigm, Redlich and Goodman (2003) found that juveniles aged 12- and 13-years-old, and 15- and 16-years-old, were more likely to confess than young adults (aged 18–26 years), especially when confronted with false evidence of their culpability. In fact, a majority of the younger participants, in contrast to adults, complied with the request to sign a false confession without uttering a word. In another laboratory experiment, researchers examined the effect of positive and negative reinforcement on children aged 5 through 8 years (Billings et al., 2007). Reinforcement strongly affected children's likelihood of making false statements: Of those in the reinforcement condition, 52% made false admissions of guilty knowledge and 30% made false admissions of having witnessed the crime (within a span of 3.5 minutes!). In contrast, of children in the control condition, only 36 and 10% made false guilty knowledge and admissions, respectively. These findings mirror the vast majority of studies on the interview-relevant abilities of child-victim/witnesses (e.g., Garven, Wood, & Malpass, 2000).

In a second set of studies, youths have made decisions in response to hypothetical scenarios. Goldstein et al. (2003) investigated male juvenile offenders' self-reported likelihood of providing false confessions across different interrogation situations and found that younger age significantly predicted false confessions (25% surmised that they would definitely confess despite innocence to at least one of the situations). Similarly, Grisso et al. (2003) examined juveniles' and young adults' responses to a hypothetical mock-interrogation situation—specifically, whether they would confess to police, remain silent, or deny the offense. Compared to individuals aged 16 and older, those between 11 and 15 were significantly more likely to report that they would confess.

In a third set of studies, juveniles have been asked to self-report on actual interrogation experiences. In a sample of 114 justice-involved juveniles, Viljoen, Klaver, and Roesch (2005) found that suspects who were 15-years old and younger, compared to those who were 16- and 17-years old, were significantly more likely to waive their right to counsel and to confess. Overall, only 11 (less than 10%) said they had asked for an attorney during police questioning (see also Redlich et al., 2004) and 9 (6%) said they had at some point falsely confessed. A survey of over 10,000 Icelandic students aged 16–24 years similarly revealed that of those with interrogation experiences, 7% claimed to have falsely confessed, with the rates being higher among those with more than one interrogation experience (Gudjonsson, Sigurdsson, Asgeirsdottir, & Sigfusdottir, 2006). In a massive and more recent effort, more than 23,000 juveniles from grades 8, 9, and 10 (average age of 15.5 years) were surveyed from seven countries—Iceland, Norway, Finland, Latvia, Lithuania, Russia, and Bulgaria. Overall, 11.5% (2,726) reported having been interrogated by police. Within this group, 14% reported having given a false confession (Gudjonsson, Sigurdsson, Asgeirsdottir, & Sigfusdottir, in press).

Cognitive and Intellectual Disabilities

Much of what is true of juveniles is similarly true for persons with intellectual disabilities—another group that is over-represented in false confession cases (see Gudjonsson, 2003; Gudjonsson & MacKeith, 1994). Hence, in *Atkins v. Virginia* (2002), the U.S. Supreme Court explicitly cited the possibility of false confession as a rationale underlying their decision to exclude this group categorically from capital punishment. The case of Earl Washington is illustrative of the problem. Reported to have an IQ ranging

from 57 to 69 and interrogated over the course of 2 days, Washington "confessed" to five crimes, one being the rape and murder of a woman (charges resulting from the other four confessions were dismissed because of inconsistencies). Although he could not provide even basic details (e.g., that the victim was raped or her race) and although much of his statement was inconsistent with the evidence, Washington—who was easily led by suggestive questions and deferred to authority figures—was convicted, sentenced to death, and incarcerated for 18 years before being exonerated (Hourihan, 1995).

Mental retardation represents a constellation of symptoms, disorders, and adaptive functioning. The condition is defined by an IQ score of 70 or below and a range of impairments, such as adapting to societal norms, communication, social and interpersonal skills, and self-direction (American Psychiatric Association, 1994). In training police recruits, Perske (2004) identifies from research a number of tendencies exhibited by people who are mentally retarded. Collectively suggesting a heightened susceptibility to influence, the list includes the tendencies to rely on authority figures for solutions to everyday problems; please persons in authority; seek out friends; feign competence; exhibit a short attention span; experience memory gaps; lack impulse control; and accept blame for negative outcomes.

Some researchers have provided evidence for the diminished capacity of persons with cognitive disabilities in studies pertaining to interrogation (Fulero & Everington, 2004). Across four studies of *Miranda* comprehension, findings are quite consistent in showing that persons with mental retardation have significant deficits in their understanding and appreciation of *Miranda* warnings (Cloud, Shepard, Barkoff, & Shur, 2002; Everington & Fulero, 1999; Fulero & Everington, 1995; O'Connell, Garmoe, & Goldstein, 2005). For example, O'Connell et al. (2005) found that 50% of people with mild mental retardation in their sample could not correctly paraphrase *any* of the five *Miranda* components (see also Everington & Fulero, 1999). In comparison, less than 1% of adults in the general population score similarly low (Grisso, 1996). Moreover, research on the capacity of persons with mental retardation to learn and retain the knowledge and skills necessary to be competent suspects and defendants demonstrates that a significant number cannot meet this threshold, even with education (Anderson & Hewitt, 2002).

Everington and Fulero (1999) also examined the suggestibility of persons with mental retardation. Using the Gudjonsson Suggestibility Scale (GSS; a measure of interrogative suggestibility), they found that people with mental retardation were more likely to yield to leading questions and change their answers in response to mild negative feedback (see also O'Connell et al., 2005). Gudjonsson (1991) examined GSS scores among three groups: alleged false confessors, alleged true confessors, and suspects who resisted confession during questioning. He found the alleged false confessors to have the lowest IQ scores as well as the highest suggestibility scores compared to the other two groups (Gudjonsson & Clare, 1995). Finally, Clare and Gudjonsson (1995) examined perceptions of a videotaped suspect who provides a true and false confession during an interrogation and found that 38% of perceivers with intellectual disabilities, compared to only 5% of those without intellectual disabilities, believed the suspect would be allowed to go home while awaiting trial. Additionally, only 52% believed that the suspect should obtain legal advice if innocent, compared to 90% of others.

Personality and Psychopathology

In terms of susceptibility to false confession, it is important to consider other individual factors of relevance to a person's decision to confess. Gudjonsson (2003) discusses a number of personal risk factors, including enduring personality traits (e.g., suggestibility, compliance) as well as psychopathology and personality disorders—categories within the DSM-IV Axis I and II diagnostic framework that are relevant to false confessions.

A number of large-scale studies of false confessions, carried out in Iceland, show the importance of antisocial personality traits and history of offending both among prison inmates (Sigurdsson & Gudjonsson, 2001) and community samples (Gudjonsson, Sigurdsson, Asgeirsdottir, & Sigfusdottir, 2006, 2007; Gudjonsson, Sigurdsson, Bragason, et al., 2004; Gudjonsson et al., 2004). There have also been cases in which the personality disorder was considered crucial to understanding the false confession (Gudjonsson, 2006; Gudjonsson & Grisso, 2008). One interpretation of this finding is that persons with antisocial personality disorder, or antisocial traits, are more likely to be involved in offending, more often interviewed by police, and prone to lie for short-term instrumental gain, and are less concerned about the consequences of their behavior. This increases their tendency to make false denials as well as false confessions depending on their need at the time.

Psychopathology seems to be linked to false confessions in that persons with mental illness are over-represented in these cases. Psychological disorder is often accompanied by faulty reality monitoring, distorted perception, impaired judgment, anxiety, mood disturbance, poor self-control, and feelings of guilt. Gudjonsson (2003) provided a number of examples of cases where false confessions were directly related to specific disorders. Following the release of the Birmingham Six in 1991, research conducted for the British Royal Commission on Criminal Justice found that about 7% of suspects detained at police stations had a

history of mental illness and that many more were in an abnormal mental state due to anxiety and mood disturbance (Gudjonsson, Clare, Rutter, & Pearse, 1993). Similar findings were found in a recent study among suspects at Icelandic police stations (Sigurdsson, Gudjonsson, Einarsson, & Gudjonsson, 2006). In the U.S., research has consistently shown that rates of serious mental illness in the criminal justice system are at least two to five times higher than rates in the general population (e.g., James & Glaze, 2006; Lamb & Weinberger, 1998). To further compound the problem, the majority (75–80%) of offenders with mental illness have co-occurring substance abuse or dependence disorders (Abram, Teplin, & McClelland, 2003), which is an additional risk factor for false confessions (see Sigurdsson & Gudjonsson, 2001).

There is currently little research available to show how different disorders (e.g., anxiety, depression, and schizophrenia) potentially impair the suspect's capacity to waive legal rights and navigate his or her way through a police interview (Redlich, 2004). However, there is recent evidence from two separate studies to suggest that depressed mood is linked to a susceptibility to provide false confession to police (Gudjonsson et al., 2006; Sigurdsson et al., 2006). Gudjonsson et al. (2007) also recently found that multiple exposures to unpleasant or traumatic life events were significantly associated with self-reported false confessions during interrogation. Rogers et al. (2007a) found that most mentally disordered offenders exhibited insufficient understanding of *Miranda*, particularly when the warnings required increased levels of reading comprehension. Finally, Redlich (2007) found that offenders with mental illness self-reported a 22% lifetime false confession rate—notably higher than the 12% found in samples of prison inmates without mental illness (Sigurdsson & Gudjonsson, 1996).

An important type of psychopathology in relation to false confessions is attention deficit hyperactivity disorder (ADHD), which consists of three primary symptoms: inattention, hyperactivity, and impulsivity (American Psychiatric Association, 1994). This condition is commonly found among offenders (Young, 2007). Moreover, research shows that people with ADHD cope during questioning by answering a disproportionate number of questions with "don't know" replies—which may lead police to be suspicious of their answers (Gudjonsson, Young, & Bramham, 2007). They may also exhibit high levels of compliance. Gudjonsson et al. (2008) found that the rate of self-reported false confessions was significantly higher among prisoners who were currently symptomatic for attention deficit hyperactivity disorder (ADHD) than among the other prisoners (41 and 18%, respectively). These findings highlight the potential vulnerability during questioning of people who are currently symptomatic for ADHD.

Protections for Vulnerable Suspects in England

When the police interview mentally disordered persons and juveniles in England and Wales, there are special legal provisions available to ensure that their statements to police are reliable and properly obtained—for example, in the presence of "appropriate adults." The current legal provisions are detailed in the Codes of Practice (Home Office, 2003). Even when the police adhere to all the legal provisions, a judge may consider it unsafe and unfair to allow the statement to go before the jury. Here the crucial issue may be whether or not the defendant was "mentally fit" when interviewed. The term "fitness for interview" was first introduced formally in the current Codes of Practice, which became effective in 2003.

Fitness for interview is closely linked to the concept of "legal competencies," which refers to an individual's physical, mental, and social vulnerabilities that may adversely affect his or her capacity to cope with the investigative and judicial process (Grisso, 1986). Historically, legal competence constructs relating to confession evidence have focused primarily on the functional deficits of juveniles (Drizin & Colgan, 2004), and adult defendants with mental retardation (Fulero & Everington, 2004) and mental illnesses (Melton, Petrila, Poythress, & Slobogin, 1997). Increasingly, the construct of legal competence in criminal cases is also being applied to defendants with "personality disorder" (Gudjonsson & Grisso, 2008). The introduction of "fitness to be interviewed" within the current Codes of Practice in England and Wales is a significant step toward protecting vulnerable suspect populations (Gudjonsson, 2005). Indeed, a similar framework has been introduced in New Zealand and Australia (Gall & Freckelton, 1999).

Innocence as a Risk Factor

On September 20, 2006, Jeffrey Mark Deskovic was released from a maximum-security prison in New York, where he spent 15 years for a murder he said he committed but did not. Why did he confess? "Believing in the criminal justice system and being fearful for myself, I told them what they wanted to hear," Deskovic said. Certain that DNA testing on the semen would establish his innocence, he added: "I thought it was all going to be okay in the end" (Santos, 2006, p. A1).

On the basis of anecdotal and research evidence, Kassin (2005) suggested the ironic hypothesis that *innocence* itself may put *innocents* at risk. Specifically, it appears that people who stand falsely accused tend to believe that truth and justice will prevail and that their innocence will become transparent to investigators, juries, and others. As a result, they cooperate fully with police, often failing to

realize that they are suspects not witnesses, by waiving their rights to silence and a lawyer and speaking freely to defend themselves. Thus, although mock criminals vary their disclosures according to whether the interrogator seems informed about the evidence, innocents are uniformly forthcoming—regardless of how informed the interrogator seems (Hartwig, Granhag, Strömwall, & Kronkvist, 2006; Hartwig, Granhag, Strömwall, & Vrij, 2005).

Based on observations of live and videotaped interrogations, Leo (1996b) found that four out of five suspects waive their rights and submit to questioning—and that people who have no prior record of crime are the most likely to do so. In light of known recidivism rates, this result suggested that innocent people in particular are at risk to waive their rights. Kassin and Norwick (2004) tested this hypothesis in a controlled laboratory setting in which some subjects but not others committed a mock theft of $100. Upon questioning, subjects who were innocent were more likely to sign a waiver than those who were guilty, 81 to 36%. Afterward, most innocent subjects said that they waived their rights precisely because they were innocent: "I did nothing wrong," "I had nothing to hide." The feeling of reassurance that accompanies innocence may be rooted in a generalized and perhaps motivated belief in a just world in which human beings get what they deserve and deserve what they get (Lerner, 1980). It may also stem from the "illusion of transparency," a tendency for people to overestimate the extent to which their true thoughts, emotions, and other inner states can be seen by others (Gilovich, Savitsky, & Medvec, 1998; Miller & McFarland, 1987). Whatever the mechanism, it is clear that *Miranda* warnings may not adequately protect the citizens who need it most—those accused of crimes they did not commit (Kassin, 2005).

These findings suggest that people have a naïve faith in the power of innocence to set them free. This phenomenology was evident in the classic case of Peter Reilly, an 18-year-old who falsely confessed to the murder of his mother. When asked years later why he did not invoke his *Miranda* rights, Reilly said, "My state of mind was that I hadn't done anything wrong and I felt that only a criminal really needed an attorney, and this was all going to come out in the wash" (Connery, 1996, p. 93). Innocence may lead innocents to forego other important safeguards as well. Consider the case of Kirk Bloodsworth, the first death row inmate to be exonerated by DNA. In 1985, based solely on eyewitness identifications, Bloodsworth was convicted for the rape and murder of a 9-year-old girl. He was exonerated by DNA 8 years later and ultimately vindicated when the true perpetrator was identified. The day of his arrest, Bloodsworth was warned that there would be cameras present and asked if he wanted to cover his head with a blanket. He refused, saying he did nothing wrong and was not going to hide—even though potential witnesses might see him on TV (Junkin, 2004).

THE CONSEQUENCES OF CONFESSION

It is inevitable that some number of innocent people will be targeted for suspicion and subjected to excessively persuasive interrogation tactics, and many of them will naively and in opposition to their own self-interest waive their rights and confess. One might argue that this unfortunate chain of events is tolerable, not tragic, to the extent that the resulting false confessions are detected by authorities at some point and corrected. Essential to this presumed safety net is the belief that police, prosecutors, judges, and juries are capable of distinguishing true and false confessions.

The process begins with the police. Numerous false confession cases reveal that once a suspect confesses, police often close their investigation, deem the case solved, and overlook exculpatory evidence or other possible leads–even if the confession is internally inconsistent, contradicted by external evidence, or the product of coercive interrogation (Drizin & Leo, 2004; Leo & Ofshe, 1998). This trust in confessions may extend to prosecutors as well, many of whom express skepticism about police-induced false confessions, stubbornly refusing to admit to such an occurrence even after DNA evidence has unequivocally established the defendant's innocence (Findley & Scott, 2006; Hirsch, 2005b; Kassin & Gudjonsson, 2004). Upon confession, prosecutors tend to charge suspects with the highest number and types of offenses, set bail higher, and are far less likely to initiate or accept a plea bargain to a reduced charge (Drizin & Leo, 2004; Leo & Ofshe, 1998; but see Redlich, in press).

Part of the problem is that confessions can taint other evidence. In one case, for example, Pennsylvania defendant Barry Laughman confessed to rape and murder, which was later contradicted by blood typing evidence. Clearly influenced by the confession, the state forensic chemist went on to concoct four "theories," none grounded in science, to explain away the mismatch. Sixteen years later, Laughman was set free (http://www.innocenceproject.org). Recent empirical studies have demonstrated the problem as well. In one study, Dror and Charlton (2006) presented five latent fingerprint experts with pairs of prints from a crime scene and suspect in an actual case in which they had previously made a match or exclusion judgment. The prints were accompanied either by no extraneous information, an instruction that the suspect had confessed (suggesting a match), or an instruction that the suspect was in custody at the time (suggesting an exclusion). The misinformation

produced a change in 17% of the original, previously correct judgments. In a second study, Hasel and Kassin (2009) staged a theft and took photographic identification decisions from a large number of eyewitnesses who were present. One week later, individual witnesses were told that the person they had identified denied guilt, or that he confessed, or that a specific other lineup member confessed. Influenced by this information, many witnesses went on to change their identification decisions, selecting the confessor with confidence, when given the opportunity to do so.

Not surprisingly, confessions are particularly potent in the courtroom. When a suspect in the U.S. retracts his or her confession, pleads not guilty, and goes to trial, a sequence of two decisions is set into motion. First, a judge determines whether the confession was voluntary and hence admissible as evidence. Then a jury, hearing the admissible confession, determines whether the defendant is guilty beyond a reasonable doubt. But can people distinguish between true and false confessions? And what effect does this evidence have within the context of a trial?

Research on the impact of confessions throughout the criminal justice system is unequivocal. Mock jury studies have shown that confessions have more impact than other potent forms of evidence (Kassin & Neumann, 1997) and that people do not fully discount confessions—even when they are judged to be coerced (Kassin & Wrightsman, 1980) and even when the confessions are presented secondhand by an informant who is motivated to lie (Neuschatz, Lawson, Swanner, Meissner, & Neuschatz, 2008). For example, Kassin and Sukel (1997) presented mock jurors with one of three versions of a murder trial transcript. In a low-pressure version, the defendant was said to have confessed to police immediately upon questioning. In a high-pressure version, participants read that the suspect was in pain and interrogated aggressively by a detective who waved his gun in a menacing manner. A control version contained no confession in evidence. Presented with the high-pressure confession, participants appeared to respond in the legally prescribed manner. They judged the statement to be involuntary and said it did not influence their decisions. Yet when it came to the all-important verdict measure, this confession significantly increased the conviction rate. This increase occurred even in a condition in which subjects were specifically admonished to disregard confessions they found to be coerced. Similar results have recently been reported in mock jury studies involving defendants who are minors (Redlich, Ghetti, & Quas, 2008; Redlich, Quas, & Ghetti, 2008).

This point concerning the power of confession evidence is bolstered by recent survey evidence indicating that although laypeople understand that certain interrogation tactics are psychologically coercive, they do not believe that these tactics elicit false confessions (Leo & Liu, 2009). Archival analyses of actual cases also reinforce this point. When proven false confessors pleaded not guilty and proceeded to trial, the jury conviction rates ranged from 73% (Leo & Ofshe, 1998) to 81% (Drizin & Leo, 2004). These figures led Drizin and Leo to describe confessions as "inherently prejudicial and highly damaging to a defendant, even if it is the product of coercive interrogation, even if it is supported by no other evidence, and even if it is ultimately proven false beyond any reasonable doubt" (p. 959).

There are at least three reasons why people cannot easily identify as false the confessions of innocent suspects. First, generalized common sense leads people to trust confessions the way they trust other behaviors that counter self-interest. Over the years, and across a wide range of contexts, social psychologists have found that social perceivers fall prey to the fundamental attribution error—that is, they tend to make dispositional attributions for a person's actions, taking behavior at face value, while neglecting the role of situational factors (Jones, 1990; Ross, 1977). Gilbert and Malone (1995) offered several explanations for this bias, the most compelling of which is that people draw quick and relatively automatic dispositional inferences from behavior and then fail to adjust or correct for the presence of situational constraints. Common sense further compels the belief that people present themselves in ways that are self-serving and that confessions must therefore be particularly diagnostic of guilt. Indeed, most people reasonably believe that they would never confess to a crime they did not commit and have only rudimentary understanding of the predispositional and situational factors that would lead someone to do so (Henkel, Coffman, & Dailey, 2008).

A second reason is that people are typically not adept at deception detection. We saw earlier that neither lay people nor professionals distinguish truths from lies at high levels of accuracy. This problem extends to judgments of true and false confessions. To demonstrate, Kassin, Meissner, and Norwick (2005) videotaped male prison inmates providing true confessions to the crimes for which they were incarcerated and concocting false confessions to crimes selected by the experimenter that they did not commit. When college students and police investigators later judged these statements from videotapes or audiotapes, the results showed that neither group was particularly adept, exhibiting accuracy rates that ranged from 42 to 64%—typically not much better than chance performance. These findings suggest people cannot readily distinguish true and false confessions and that law enforcement experience does not improve performance. This latter result is not surprising, as many of the behavioral cues that typically form part of the basis for training (e.g., gaze aversion, postural cues, and

grooming gestures) are not statistically correlated with truth-telling or deception (DePaulo et al., 2003).

On the assumption that "I'd know a false confession if I saw one," there is a third reason for concern: Police-induced false confessions often contain *content* cues presumed to be associated with truthfulness. In many documented false confessions, the statements ultimately presented in court contained not only an admission of guilt but vivid details about the crime, the scene, and the victim that became known to the innocent suspect through leading questions, photographs, visits to the crime scene, and other secondhand sources invisible to the naïve observer. To further complicate matters, many false confessors state not just what they allegedly did, and how they did it, but *why*—as they self-report on revenge, jealousy, provocation, financial desperation, peer pressure, and other prototypical motives for crime. Some of these statements even contain apologies and expressions of remorse. To the naïve spectator, such statements appear to be voluntary, textured with detail, and the product of personal experience. Uninformed, however, this spectator mistakes illusion for reality, not realizing that the taped confession is scripted by the police theory of the case, rehearsed during hours of unrecorded questioning, directed by the questioner, and ultimately enacted on paper, tape, or camera by the suspect (see Kassin, 2006).

RECOMMENDATIONS FOR REFORM

Confession is a potent form of evidence that triggers a chain of events from arrest, prosecution, and conviction, through post-conviction resistance to change in the face of exculpatory information. Recent DNA exonerations have shed light on the problem that innocent people, confident in the power of their innocence to prevail, sometimes confess to crimes they did not commit. Research has identified two sets of risks factors. The first pertains to the circumstances of interrogation, situational factors such as a lengthy custody and isolation, possibly accompanied by a deprivation of sleep and other need states; presentations of false evidence, a form of trickery that is designed to link the suspect to the crime and lead him or her to feel trapped by the evidence; and minimization tactics that lead the suspect and others to infer leniency even in the absence of an explicit promise. The second set of risk factors pertains to dispositional characteristics that render certain suspects highly vulnerable to influence and false confessions—namely, adolescence and immaturity; cognitive and intellectual impairments; and personality characteristics and mental illness.

In light of the wrongful convictions involving false confessions that have recently surfaced, as well as advances in psychological research on interviewing, interrogations, and confessions, there are renewed calls for caution regarding confessions and the reform of interrogation practices not seen since the Wickersham Commission Report (1931) and U.S. Supreme Court opinion in *Miranda* (1966). Professionals from varying perspectives may differ in their perceptions of both the problems and the proposed solutions. Hence, it is our hope that the recommendations to follow will inspire a true collaborative effort among law enforcement professionals, district attorneys, defense lawyers, judges, social scientists, and policy makers to scrutinize the systemic factors that put innocent people at risk and devise effective safeguards.

Electronic Recording of Interrogations

Without equivocation, our most essential recommendation is to lift the veil of secrecy from the interrogation process in favor of the principle of transparency. Specifically, *all custodial interviews and interrogations of felony suspects should be videotaped in their entirety and with a camera angle that focuses equally on the suspect and interrogator.* Stated as a matter of requirement, such a policy evokes strong resistance in some pockets of the law enforcement community. Yet it has also drawn advocates from a wide and diverse range of professional, ideological, and political perspectives (e.g., American Bar Association, 2004; Boetig, Vinson, & Weidel, 2006; Cassell, 1996a; Drizin & Colgan, 2001; Geller, 1994; Gudjonsson, 2003; Leo, 1996c; Slobogin, 2003; Sullivan, 2004; The Justice Project, 2007).

In England, under the Police and Criminal Evidence Act of 1984, the mandatory requirement for tape-recording police interviews was introduced to safeguard the legal rights of suspects and the integrity of the process. At first resisted by police, this requirement has positively transformed the ways in which police interviews are conducted and evaluated. Over the years, the need for taping has pressed for action within the U.S. as well. In *Convicting the Innocent*, a classic study of wrongful convictions, Edwin Borchard (1932) expressed concern that police abuses during interrogations led to involuntary and unreliable confessions. His solution, utilizing the technology of the time, was to make "[phonographic records" [of interrogations] which shall alone be introducible in court" (pp. 370–371).

Throughout the twentieth century, other advocates for recording were less concerned with preventing false confessions and more concerned with increasing the accuracy of the justice system by eliminating the swearing contests between police officers and suspects over what occurred during the interrogation (Kamisar, 1977; Weisberg, 1961). Still others saw that recording interrogations held

tremendous benefits for law enforcement by discouraging note-taking and other practices that could inhibit suspects, helping police officers obtain voluntary confessions, nabbing accomplices, and protecting officers from false allegations of abuse (Geller, 1993; O'Hara, 1956). Despite these calls for recording, by the turn of the twentieth century only two states, by virtue of state Supreme Court decisions—Alaska (*Stephan v. State*, 1985) and Minnesota (*State v. Scales*, 1994)—required law enforcement officers to electronically record suspect interrogations. The pace of reform in this area, however, is picking up and once again a concern about false confessions seems to be the impetus. In the post-DNA age, and particularly in the past 5 years, as the number of wrongful convictions based on false confessions has continued to climb, concerns about the reliability of confession evidence have led to a renewed push for recording requirements (Drizin & Reich, 2004). As a result of statutes and court rulings, seven additional jurisdictions—Illinois, Maine, New Mexico, New Jersey, Wisconsin, North Carolina, and the District of Columbia—have joined Minnesota and Alaska, in requiring recordings of custodial interrogations in some circumstances (Robertson, 2007; Sullivan, 2004). In several other states, supreme courts have stopped short of requiring recording but either have issued strongly worded opinions endorsing recording—e.g., New Hampshire (*State v. Barnett*, 2002) and Iowa (*State v. Hajtic*, 2007)—or, in the case of Massachusetts, held that where law enforcement officers have no excuse for the failure to record interrogation, defendants are entitled to a strongly worded instruction admonishing jurors to treat unrecorded confessions with caution (*Commonwealth v. DiGiambattista*, 2004).

In addition to recent developments in state courts and legislatures, there is a growing movement among law enforcement agencies around the country to record interrogations voluntarily. Over the past 70 years, the idea has been anathema to many in law enforcement—including the FBI, which prohibits electronic recording, and John Reid & Associates, which used to vigorously oppose the practice of recording interrogations (Inbau et al., 2001; but see Buckley & Jayne's [2005] recent publication, *Electronic Recording of Interrogations*; for an historical review, see Drizin & Reich, 2004). Yet there are now signs that police opposition is thawing (e.g., Boetig et al., 2006). Several years ago, a National Institute of Justice study found that one-third of large police and sheriff's departments throughout the U.S. were already videotaping at least some interrogations or confessions and that their experiences with the practice were positive (Geller, 1993). A more recent survey of more than 465 law enforcement agencies in states that do not require electronic recording of interrogations has revealed that the practice is widespread. Without any legislative or judicial compulsion, police departments in many states routinely record interviews and interrogations in major felony investigations. Without exception, they have declared strong support for the practice (Sullivan, 2004; Sullivan, Vail, & Anderson, 2008).

There are numerous advantages to a videotaping policy. To begin, the presence of a camera may deter interrogators from using the most egregious, psychologically coercive tactics—and deter frivolous defense claims of coercion where none existed. Second, a videotaped record provides trial judges (ruling on voluntariness) and juries (determining guilt) an *objective* and *accurate* record of the process by which a statement was taken—a common source of dispute that results from ordinary forgetting and self-serving distortions in memory. In a study that demonstrates the problem, Lamb, Orbach, Sternberg, Hershkowitz, and Horowitz (2000) compared interviewers' verbatim contemporaneous accounts of 20 forensic interviews with alleged child sex abuse victims with tape recordings of these same sessions. Results showed that more than half of the interviewers' utterances and one quarter of the details that the children provided did not appear in their verbatim notes. Even more troubling was that interviewers made frequent and serious source attribution errors—for example, often citing the children, not their own prompting questions, as the source of details. This latter danger was inadvertently realized by D.C. Detective James Trainum (2007) who—in an article entitled "I took a false confession – so don't tell me it doesn't happen!"—recounted a case in which a suspect who had confessed to him was later exonerated: "Years later, during a review of the videotapes, we discovered our mistake. We had fallen into a classic trap. We believed so much in our suspect's guilt that we ignored all evidence to the contrary. To demonstrate the strength of our case, we showed the suspect our evidence, and unintentionally fed her details that she was able to parrot back to us at a later time. It was a classic false confession case and without the video we would never have known" (see also Trainum, 2008). Similarly, Police Commander Neil Nelson, of St. Paul, Minnesota, said that he too once elicited a false confession, which he came to doubt by reviewing the interrogation tape: "You realize maybe you gave too much detail as you tried to encourage him and he just regurgitated it back" (Wills, 2005; quoted online by Neil Nelson & Associates; http://www.neilnelson.com/pressroom.html).

To further complicate matters of recollection, police interrogations are not prototypical social interactions but, rather, extraordinarily stressful events for those who stand accused. In a study that illustrates the risk to accurate retrieval, Morgan et al. (2004) randomly assigned trainees in a military survival school to undergo a realistic high-stress or low-stress mock interrogation. Twenty-four hours later, he found that those in the high-stress condition had

difficulty even identifying their interrogators in a lineup. In real criminal cases, questions constantly arise about whether rights were administered and waived, whether the suspect was cooperative or evasive, whether detectives physically intimidated the suspect, whether promises or threats were made or implied, and whether the details in a confession emanated from the police or suspect, are among the many issues that become resolvable (in Great Britain, as well, taping virtually eliminated the concern that police officers were attributing to suspects admissions that would later be disputed; see Roberts, 2007).

In recent years, Sullivan (2004, 2007) has tirelessly interviewed law enforcement officials from hundreds of police and sheriff's departments that have recorded custodial interrogations and found that they enthusiastically favored the practice. Among the collateral benefits they often cited were that recording permitted detectives to focus on the suspect rather than take copious notes, increased accountability, provided an instant replay of the suspect's statement that sometimes revealed incriminating comments that were initially overlooked, reduced the amount of time detectives spent in court defending their interrogation practices, and increased public trust in law enforcement. Countering the most common apprehensions, the respondents in these interview studies reported that videotaping interrogations did not prove costly or inhibit suspects from talking to police or incriminating themselves. Typical of this uniformly positive reaction, Detective Trainum (2007) notes: "When videotaping was first forced upon us by the D.C. City Council, we fought it tooth and nail. Now, in the words of a top commander, we would not do it any other way."

It is beyond the scope of this article to draft a model rule that would address such specific details as what conditions should activate a recording requirement, how the recordings should be preserved, whether exceptions to the rule should be made (e.g., if the equipment malfunctions, if the suspect refuses to make a recorded statement), and what consequences would follow from the failure to record (e.g., whether the suspect's statement would be excluded or admitted to the jury with a cautionary instruction). As a matter of policy, however, research does suggest that it is important not only that entire sessions be recorded, triggered by custodial detention, but that the camera adopt a neutral "equal focus" perspective that shows both the accused and his or her interrogators. In 20-plus years of research on illusory causation effects in attribution, Lassiter and his colleagues have taped mock interrogations from three different camera angles so that the suspect, the interrogator, or both were visible. Lay participants who saw only the suspect judged the situation as less coercive than those focused on the interrogator. By directing visual attention toward the accused, the camera can thus lead jurors to underestimate the amount of pressure actually exerted by the "hidden" detective (Lassiter & Irvine, 1986; Lassiter, Slaw, Briggs, & Scanlan, 1992). Additional studies have confirmed that people are more attuned to the situational factors that elicit confessions whenever the interrogator is on camera than when the focus is solely on the suspect (Lassiter & Geers, 2004; Lassiter, Geers, Munhall, Handley, & Beers, 2001). Under these more balanced circumstances, juries make more informed attributions of voluntariness and guilt when they see not only the final confession but the conditions under which it was elicited (Lassiter, Geers, Handley, Weiland, & Munhall, 2002). Indeed, even the perceptions of experienced trial judges are influenced by variations in camera perspective (Lassiter, Diamond, Schmidt, & Elek, 2007).

Reform of Interrogation Practices

In light of recent events, the time is ripe for police, district attorneys, defense lawyers, judges, researchers, and policymakers to evaluate current methods of interrogation. All parties would agree that the surgical objective of interrogation is to secure confessions from perpetrators but not from innocent suspects. Hence, the process of interrogation should be structured in theory and in practice to produce outcomes that are accurate, as measured by the observed ratio of true to false confessions. Yet except for physical brutality or deprivation, threats of harm or punishment, promises of leniency or immunity, and flagrant violations of a suspect's constitutional rights, there are no clear criteria by which to regulate the process. Instead, American courts historically have taken a "totality of the circumstances" approach to voluntariness and admissibility. Because *Miranda* does not adequately safeguard the innocent, we believe that the time is right to revisit the factors that comprise those circumstances.

As illustrated by the Reid technique and other similar approaches, the modern American police interrogation is, by definition, a guilt-presumptive and confrontational process—aspects of which put innocent people at risk. There are two ways to approach questions of reform. One is to completely reconceptualize this model at a macro level and propose that the process be converted from "confrontational" to "investigative." Several years ago, after a number of high-profile false confessions, the British moved in this direction, transitioning police from a classic interrogation to a process of "investigative interviewing." The Police and Criminal Evidence (PACE) Act of 1984 sought to reduce the use of psychologically manipulative tactics. In a post-PACE study, Irving and McKenzie (1989) found that the use of psychologically manipulative tactics had significantly declined—without a corresponding drop in the frequency of confessions. The post-PACE confession rate

is also somewhat higher in the UK than in the U.S. (Gudjonsson, 2003). In 1993, the Royal Commission on Criminal Justice further reformed the practice of interrogation by proposing the PEACE model described earlier ("Preparation and Planning," "Engage and Explain," "Account," "Closure," and "Evaluate"), the purpose of which is fact finding rather than confession. Observational research suggests that such investigative interviews enable police to inculpate offenders—and youthful suspects as well (Hershkowitz, Horowitz, Lamb, Orbach, & Sternberg, 2004; Lamb, Orbach, Hershkowitz, Horowitz, & Abbott, 2007)—by obtaining from them useful, evidence-generating information about the crime (for reviews, see Bull & Soukara, 2009; Williamson, 2006).

Similar techniques have been taught and employed in the U.S. as well, where Nelson (2007) reports from experience that it is highly effective. Recent laboratory research has also proved promising in this regard. In one series of experiments, interviewers more effectively exposed deceptive mock criminals when they strategically withheld incriminating evidence than when they confronted the suspects with that evidence (Hartwig et al., 2005, 2006). In an experiment using the Russano et al. (2005) cheating paradigm described earlier, Rigoni and Meissner (2008) independently varied and compared accusatorial and inquisitorial methods and found that the latter produced more diagnostic outcomes—lowering the rate of false confessions without producing a corresponding decrease in the rate of true confessions. Although more systematic research is needed, it is clear that investigative interviewing offers a potentially effective macro alternative to the classic American interrogation. Indeed, New Zealand and Norway have recently adopted the PEACE approach to investigative interviewing as a matter of national policy.

A second approach to the question of reform is to address specific risk factors inherent within a confrontational framework for interrogation. On the basis of converging evidence from actual false confession cases, basic principles of psychology, and forensic research, the existing literature suggests that certain interrogation practices alone and in combination with each other pose a risk to the innocent—whether they are dispositionally vulnerable or not. Focused in this way, but stopping short of making specific recommendations, we propose that the following considerations serve as a starting point for collaborative discussion.

Custody and Interrogation Time

As noted earlier, the human needs for belonging, affiliation, and social support, especially in times of stress, are a fundamental human motive. Prolonged isolation from significant others thus constitutes a form of deprivation that can heighten a suspect's distress and increase his or her incentive to escape the situation. Excessive time in custody may also be accompanied by fatigue and feelings of helplessness and despair as well as the deprivation of sleep, food, and other biological needs. The vast majority of interrogations last from 30 minutes up to 2 hours (Baldwin, 1993; Irving, 1980; Kassin et al., 2007; Leo, 1996b; Wald et al., 1967). Inbau et al. (2001) cautioned against surpassing 4 hours, and Blair (2005) argued that interrogations exceeding 6 hours are "legally coercive." Yet research shows that in proven false confession cases the interrogations had lasted for an average of 16.3 hours (Drizin & Leo, 2004). Following PACE in Great Britain, policy discussions should begin with a proposal for the imposition of time limits, or at least flexible guidelines, when it comes to detention and interrogation, as well as periodic breaks from questioning for rest and meals. At a minimum, police departments should consider placing internal time limits on the process that can be exceeded—initially and at regular intervals thereafter, if needed—only with authorization from a supervisor of detectives.

Presentations of False Evidence

A second problem concerns the tactic of presenting false evidence, which is often depicted as incontrovertible, and which takes the form of outright lying to suspects—for example, about an eyewitness identification that was not actually made; an alibi who did not actually implicate the suspect; fingerprints, hair, or blood that was not actually found; or polygraph tests that they did not actually fail. In *Frazier v. Cupp* (1969), the U.S. Supreme Court reviewed a case in which police falsely told the defendant that his cousin (whom he said he was with), had confessed, which immediately prompted the defendant to confess. The Court sanctioned this type of deception—seeing it as relevant to its inquiry on voluntariness but not a reason to disqualify the resulting confession. Although some state courts have distinguished between mere false assertions, which are permissible, and the fabrication of reports, tapes, and other evidence, which are not, the Supreme Court has not revisited the issue.

From a convergence of three sources, there is strong support for the proposition that outright lies can put innocents at risk to confess by leading them to feel trapped by the inevitability of evidence against them. These three sources are: (1) the aggregation of actual false confession cases, many of which involved use of the false evidence ploy; (2) one hundred-plus years of basic psychology research, which proves without equivocation that misinformation can substantially alter people's visual perceptions, beliefs, motivations, emotions, attitudes, memories, self-assessments, and even certain physiological outcomes, as seen in

studies of the placebo effect; and (3) numerous experiments, from different laboratories, demonstrating that presentations of false evidence increase the rate at which innocent research participants agree to confess to prohibited acts they did not commit. As noted earlier, scientific evidence for the malleability of people's perceptions, decisions, and behavior when confronted with misinformation is broad and pervasive. With regard to a specific variant of the problem, it is also worth noting that the National Research Council Committee to Review the Scientific Evidence on the Polygraph (2003) recently expressed concern over the risk of false confessions produced by telling suspects they had failed the polygraph (see also Lykken, 1998).

Over the years, legal scholars have debated the merits of trickery and deception in the interrogation room (e.g., Magid, 2001; Slobogin, 2007; Thomas, 2007) and some law enforcement professionals have argued that lying is sometimes a necessary evil, effective, and without risk to the innocent (Inbau et al., 2001). To this argument, two important points must be noted. First, direct observations and self-report surveys of American police suggest that the presentation of false evidence is a tactic that is occasionally used (e.g., Feld, 2006a, 2006b; Kassin et al., 2007; Leo, 1996b). Some interrogators no doubt rely on this ploy more than others do. Yet in a position paper on false confessions, the Wisconsin Criminal Justice Study Commission (2007) concluded that "Experienced interrogators appear to agree that false evidence ploys are relatively rare" (p. 6). Second, it is instructive that in Great Britain, where police have long been prohibited from deceiving suspects about the evidence, relying instead on the investigative interviewing tactics described earlier, there has been no evidence of a decline in confession rates (Clarke & Milne, 2001; Gudjonsson, 2003; Williamson, 2006).

In light of the demonstrated risks to the innocent, we believe that the false evidence ploy, which is designed to thrust suspects into a state of inevitability and despair, should be addressed. The strongest response would be an outright ban on the tactic, rendering all resulting confessions per se inadmissible—as they are if elicited by promises, threats, and physical violence (such a ban currently exists in England, Iceland, and Germany; suspects are differently protected in Spain and Italy, where defense counsel must be present for questioning). A second approach, representing a relatively weak response, would involve calling for no direct action, merely a change of attitude in light of scientific research that will lead the courts to weigh the false evidence ploy more heavily when judging voluntariness and reliability according to a "totality of the circumstances."

Representing a compromise between an outright ban and inaction, we urge police, prosecutors, and the courts, in light of past wrongful convictions and empirical research, to heighten their sensitivity to the risks that false evidence poses to the innocent suspect. One way to achieve this compromise would be to curtail some variants of the false evidence ploy but not others—or in the case of some suspects but not others. As noted earlier, some state courts have distinguished between mere false assertions and the fabrication of reports, tapes, photographs, and other evidence, the latter being impermissible. This particular distinction seems arbitrary. False evidence puts innocents at risk to the extent that a suspect is vulnerable (e.g., by virtue of his or her youth, naiveté, intellectual deficiency, or acute emotional state) and to the extent that the alleged evidence it is presented as incontrovertible, sufficient as a basis for prosecution, and impossible to overcome. By this criterion, which the courts would have to apply on a case-by-case basis, a confession produced by telling an adult suspect that his cousin had confessed, the ploy used in *Frazier v. Cupp* (1969), might well be admissible. Yet a confession produced by telling a traumatized 14-year-old boy that his hair was found in his murdered sister's grasp, that her blood was found in his bedroom, and that he failed an infallible lie detector test—the multiple lies presented to false confessor Michael Crowe—would be excluded (White, 2001).

Minimization Tactics

A third area of concern involves the use of minimization techniques (often called "themes," "scenarios," or "inducements") that can communicate promises of leniency indirectly through pragmatic implication. While American federal constitutional law has long prohibited the use of explicit promises of leniency (*Bram v. United States*, 1897; *Leyra v. Denno*, 1954; *Lynumn v. Illinois*, 1963), uses of minimization are less clear. There is some legal support for the proposition that implicit promises of leniency are also prohibited in federal constitutional law (White, 1997), although a majority of states hold that a promise of leniency is only one factor to be considered in determining whether a confession is involuntary (White, 2003).

Multiple sources support the proposition that implicit promises can put innocents at risk to confess by leading them to perceive that the only way to lessen or escape punishment is by complying with the interrogator's demand for confession, especially when minimization is used on suspects who are also led to believe that their continued denial is futile and that prosecution is inevitable. These sources are: (1) the aggregation of actual false confession cases, the vast majority of which involved the use of minimization or explicit promises of leniency (Drizin & Leo, 2004; Leo & Ofshe, 1998; Ofshe & Leo, 1997a, 1997b; White, 2001); (2) basic psychological

research indicating, first, that people are highly responsive to reinforcement and make choices designed to maximize their outcomes (Hastie & Dawes, 2001), and second that people can infer certain consequences in the absence of explicit promises and threats by pragmatic implication (Chan & McDermott, 2006; Harris & Monaco, 1978; Hilton, 1995); and (3) experiments specifically demonstrating that minimization increases the rate at which research participants infer leniency in punishment and confess, even if they are innocent (Kassin & McNall, 1991; Klaver, Lee, & Rose, 2008; Russano et al., 2005).

In light of the demonstrated risks to the innocent, we believe that techniques of minimization, as embodied in the "themes" that interrogators are trained to develop, which communicate promises of leniency via pragmatic implication, should be scrutinized. Some law enforcement professionals have argued that minimization is a necessary interrogation technique (Inbau et al., 2001). As with the false evidence ploy, there are several possible approaches to the regulation of minimization techniques—ranging from the recommendation that no action be taken to an outright ban on minimization. Between these extreme positions one might argue that some uses of minimization but not others should be limited or modified.

Minimization techniques come in essentially three forms: those that minimize the *moral* consequences of confessing, those that minimize the *psychological* consequences of confessing, and those that minimize the *legal* consequences of confessing (Inbau et al., 2001; Ofshe & Leo, 1997a, 1997b). One possible compromise between the two extreme positions noted above would be to permit moral and psychological forms of minimization, but ban legal minimization that communicates promises of leniency via pragmatic implication. With this distinction in mind, interrogators would be permitted, for example, to tell a suspect that he or she will feel better after confession (psychological minimization) or that he or she is still a good person (moral minimization), but not that the legal consequences of his actions will be minimized if he confesses (e.g., as may be implied by self-defense and other themes). More research is thus needed to distinguish among the different tactics that interrogators are trained to use (e.g., the provocation, peer pressure, and accident scenarios), and the pragmatic inferences that these tactics lead suspects to draw concerning the consequences of confession.

Protection of Vulnerable Suspect Populations

There is a strong consensus among psychologists, legal scholars, and practitioners that juveniles and individuals with cognitive impairments or psychological disorders are particularly susceptible to false confession under pressure. Yet little action has been taken to modulate the methods by which these vulnerable groups are questioned when placed into custody as crime suspects. More than 45 years ago, the 1962 President's Panel on Mental Retardation questioned whether confessions from defendants with mental retardation should ever be admissible at trial (see Appelbaum & Appelbaum, 1994). In 1991, Fred Inbau wrote that "special protections must be afforded to juveniles and to all other persons of below-average intelligence, to minimize the risk of untruthful admissions due to their vulnerability to suggestive questioning" (1991, pp. 9–10). More recently, Inbau et al. (2001) advised against use of the false evidence ploy with youthful suspects or those with diminished mental capacity: "These suspects may not have the fortitude or confidence to challenge such evidence and, depending on the nature of the crime, may become confused as to their own possible involvement" (p. 429; also see Buckley, 2006).

It is uniformly clear to all parties that vulnerable suspect populations—namely, juveniles and people who are cognitively impaired or psychologically disordered—need to be protected in the interrogation room. In operational terms, we believe that there are two possible ways to protect these vulnerable populations. The first concerns the mandatory presence of an attorney. A least with regard to juveniles, a parent, guardian, or other interested adult is required in some states to protect young suspects who face interrogation. Yet research suggests that the presence of an interested adult does not increase the rate at which juveniles assert their constitutional rights because these adults, often passive, frequently urge their youths to cooperate with police—a tendency observed both in the U.S. (Grisso & Ring, 1979; Oberlander & Goldstein, 2001) and in the UK, where the law provides for access to an "appropriate adult" (Pearse & Gudjonsson, 1996). For this reason, juveniles—at least those under the age of 16 (at present, the research evidence is less clear when it comes to older adolescents)—should be accompanied and advised by a professional advocate, preferably an attorney, trained to serve in this role (see Gudjonsson, 2003).

As a second possible means of protection, law enforcement personnel who conduct interviews and interrogations should receive special training—not only on the limits of human lie detection, false confessions, and the perils of confirmation biases—but on the added risks to individuals who are young, immature, mentally retarded, psychologically disordered, or in other ways vulnerable to manipulation. In a survey of 332 Baltimore police officers, Meyer and Reppucci (2007) found that while respondents understood in general terms that adolescents lack maturity of judgment and are more malleable than adults, they did not by implication believe that juvenile suspects were at greater risk in the interrogation room. Hence, they reported using roughly the same Reid-like techniques with juveniles

as they do with adults (e.g., confrontation, repetition, refusal to accept denials, false evidence, minimization, and use of alternative questions). Interestingly, one-third of these respondents stated that police could benefit from special training with regard to the interrogation of juvenile suspects. In light of research described earlier, as well as Inbau et al.'s (2001) cautionary notes on the interrogation of minors and their heightened risk for false confession, we agree.

Summary and Conclusion

In 1932, Edwin Borchard published *Convicting the innocent: Sixty-five actual errors of criminal justice*, in which several false confession cases were included. Addressing the question of how these errors were uncovered, he noted how "sheer good luck" played a prominent role and lamented on "how many unfortunate victims of error have no such luck, it is impossible to say, but there are probably many." Today's generation of post-conviction exonerations well illustrate the role that sheer good luck plays (e.g., as when DNA, long ago collected, was preserved; as when the true perpetrator finds a conscience and comes forward). With increased scientific attention to the problem of false confessions, and the reforms recommended in this article, we believe it possible to reduce the serendipitous nature of these discoveries and to increase both the diagnosticity of suspects' statements and the ability of police, prosecutors, judges, and juries to make accurate decisions on the basis of these statements.

Acknowledgment For their helpful comments on earlier drafts of this manuscript, the authors are indebted to Ray Bull, Michael Lamb, Dan Lassiter, Timothy Moore, Edward Mulvey, Richard Petty, Daniel Schacter, Laurence Steinberg, Gary Wells, and two anonymous reviewers. We also want to thank Bill Thompson, AP-LS Chair of the Scientific Review Committee, not only for his useful comments but for his invaluable support and advice throughout the process.

REFERENCES

Abram, K. M., Teplin, L. A., & McClelland, G. M. (2003). Comorbidity of severe psychiatric disorders and substance use disorders among women in jail. *American Journal of Psychiatry, 160*, 1007–1010.

Abramovitch, R., Higgins-Biss, K., & Biss, S. (1993). Young persons' comprehension of waivers in criminal proceedings. *Canadian Journal of Criminology, 35*, 309–322.

Abramovitch, R., Peterson-Badali, M., & Rohan, M. (1995). Young people's understanding and assertion of their rights to silence and legal counsel. *Canadian Journal of Criminology, 37*, 1–18.

American Bar Association. (2004). *Resolution 8A—Videotaping custodial interrogations*. Approved February 9, 2004.

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders-IV*. Washington, DC: American Psychiatric Association.

Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology, 39*, 1037–1049.

Anderson, S. D., & Hewitt, J. (2002). The effect of competency restoration training on defendants with mental retardation found not competent to proceed. *Law and Human Behavior, 26*, 343–351.

Appelbaum, K. L., & Appelbaum, P. S. (1994). Criminal justice-related competencies in defendants with mental retardation. *Journal of Psychiatry and Law, 22*, 483–503.

Arizona v. Fulminante, 499 U.S. 279 (1991).

Aronson, E. (1999). *The social animal*. New York: Worth/Freeman.

Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs, 70*, 416.

Atkins v. Virginia, 536, U.S. 304 (2002).

Ayling, C. J. (1984). Corroborating false confessions: An empirical analysis of legal safeguards against false confessions. *Wisconsin Law Review, 1984*, 1121–1204.

Baker, F., Johnson, M. W., & Bickel, W. K. (2003). Delay discounting differs between current and never-smokers across commodities, sign, and magnitudes. *Journal of Abnormal Psychology, 112*, 382–392.

Baldwin, J. (1993). Police interview techniques: Establishing truth or proof? *British Journal of Criminology, 33*, 325–352.

Baumeister, R. F., & Leary, M. R. (1996). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*, 497–529.

Bedau, H. A., & Radelet, M. L. (1987). Miscarriages of justice in potentially capital cases. *Stanford Law Review, 40*, 21–179.

Bem, D. J. (1966). Inducing belief in false confessions. *Journal of Personality and Social Psychology, 3*, 707–710.

Bickel, W. K., & Marsch, L. A. (2001). Toward a behavioral economic understanding of drug dependence: Delay discounting processes. *Addiction, 96*, 73–86.

Bickel, W. K., Odum, A. L., & Madden, G. L. (1999). Impulsivity and cigarette smoking: Delay discounting in current, never, and ex-smokers. *Psychopharmacology, 146*, 447–454.

Billings, F. J., Taylor, T., Burns, J., Corey, D. L., Garven, S., & Wood, J. M. (2007). Can reinforcement induce children to falsely incriminate themselves? *Law and Human Behavior, 31*, 125–139.

Blagrove, M. (1996). Effects of length of sleep deprivation on interrogative suggestibility. *Journal of Experimental Psychology: Applied, 2*, 48–59.

Blair, J. P. (2005). A test of the unusual false confession perspective using cases of proven false confessions. *Criminal Law Bulletin, 41*, 127–144.

Boetig, B. P., Vinson, D. M., & Weidel, B. R. (2006). Revealing incommunicado. *FBI Law Enforcement Bulletin, 75*(12), 1–8.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality & Social Psychology Review, 10*, 214–234.

Borchard, E. M. (1932). *Convicting the innocent: Errors of criminal justice*. New Haven, CT: Yale University Press.

Bram v. United States, 168 U.S. 532 (1897).

Brewer, W. F. (1977). Memory for pragmatic implications of sentences. *Memory and Cognition, 5*, 673–678.

Brown v. Mississippi, 297 U.S. 278 (1936).

Brown, W. A. (1998). The placebo effect. *Scientific American, 278*, 90–95.

Buckley, D. M., & Jayne, B. C. (2005). *Electronic recording of interrogations*. Eagle River, WI: Hahn Printing, Inc.

Buckley, J. (2006). The Reid technique of interviewing and interrogation. In T. Williamson (Ed.), *Investigative interviewing: Rights, research, regulation* (pp. 190–206). Devon, UK: Willan.

Bull, R., & Soukara, S. (2009). A set of studies of what really happens in police interviews with suspects. In G. D. Lassiter & C. A. Meissner (Eds.), *Interrogations and confessions: Research, practice, and policy*. Washington, DC: American Psychological Association.

Caldwell, J. A., Caldwell, J. L., Brown, D. L., & Smith, J. K. (2004). The effects of 37 hours of continuous wakefulness on the physiological arousal, cognitive performance, self-reported mood, and simulator flight performance of F-117A pilots. *Military Psychology, 16*, 163–181.

Candel, I., Merckelbach, H., Loyen, S., & Reyskens, H. (2005). "I hit the Shift-key and then the computer crashed": Children and false admissions. *Personality and Individual Differences, 38*, 1381–1387.

Cassell, P. G. (1996a). Miranda's social costs: An empirical reassessment. *Northwestern University Law Review, 90*, 387–499.

Cassell, P. G. (1996b). All benefits, no costs: The grand illusion of Miranda's defenders. *Northwestern University Law Review, 90*, 1084–1124.

Cassell, P. G., & Hayman, B. S. (1996). Police interrogation in the 1990s: An empirical study of the effects of Miranda. *UCLA Law Review, 43*, 839–931.

Chan, J. C. K., & McDermott, K. B. (2006). Remembering pragmatic inferences. *Applied Cognitive Psychology, 20*, 633–639.

Cialdini, R. B. (2001). *Influence: Science and practice* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Clare, I., & Gudjonsson, G. H. (1991). Recall and understanding of the caution and rights in police detention among persons of average intellectual ability and persons with a mild mental handicap. *Issues in Criminological and Legal Psychology, 1*, 34–42.

Clare, I., & Gudjonsson, G. H. (1995). The vulnerability of suspects with intellectual disabilities during police interviews: A review and experimental study of decision-making. *Mental Handicap Research, 8*, 110–128.

Clarke, C., & Milne, R. (2001). *National evaluation of the PEACE investigative interviewing course. Police Research Award Scheme*. London: Home Office.

Cloud, M., Shepard, G. B., Barkoff, A. N., & Shur, J. V. (2002). Words without meaning: The Constitution, confessions, and mentally retarded suspects. *University of Chicago Law Review, 69*, 495–624.

Colorado v. Connelly, 479 U.S. 157 (1986).

Colwell, L., Cruise, K., Guy, L., McCoy, W., Fernandez, K., & Ross, H. (2005). The influence of psychosocial maturity on male juvenile offenders' comprehension and understanding of the Miranda warning. *Journal of the American Academy of Psychiatry and Law, 33*, 444–454.

Commonwealth v. DiGiambattista, 813 N.E.2d 516 (Mass. 2004).

Connery, D. S. (Ed.). (1996). *Convicting the innocent*. Cambridge, MA: Brookline.

Cooper, V. G., & Zapf, P. A. (2008). Psychiatric patients' comprehension of *Miranda* rights. *Law and Human Behavior, 32*, 390–405.

Crocker, J., Voelkl, K., Testa, M., & Major, B. (1991). Social stigma: The affective consequences of attributional ambiguity. *Journal of Personality and Social Psychology, 60*, 218–228.

Davis, D., & O'Donahue, W. (2004). The road to perdition: Extreme influence tactics in the interrogation room. In W. O'Donahue (Ed.), *Handbook of forensic psychology* (pp. 897–996). San Diego, CA: Academic Press.

Deluty, M. Z. (1978). Self-control and impulsiveness involving aversive events. *Journal of Experimental Psychology: Animal Behavior Processes, 4*, 250–266.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*, 74–112.

Dickerson v. United States, 530 U.S. 428 (2000).

Donahue, J. (1998). Did *Miranda* diminish police effectiveness. *Stanford Law Review, 50*, 1147–1180.

Doyle, J. (2005). *True witness: Cops, courts, science, and the battle against misidentification*. New York: Palgrave Macmillan.

Drizin, S. A., & Colgan, B. A. (2001). Let the cameras roll: Mandatory videotaping of interrogations is the solution to Illinois' problem of false confessions. *Loyola University Chicago Law Journal, 32*, 337–424.

Drizin, S. A., & Colgan, B. (2004). Tales from the juvenile confession front: A guide to how standard police interrogation tactics can produce coerced and false confessions from juvenile suspects. In G. D. Lassiter (Ed.), *Interrogations, confessions, and entrapment* (pp. 127–162). New York: Kluwer Academic/Plenum.

Drizin, S. A., & Leo, R. A. (2004). The problem of false confessions in the post-DNA world. *North Carolina Law Review, 82*, 891–1007.

Drizin, S. A., & Reich, M. J. (2004). Heeding the lessons of history: The need for mandatory recording of police interrogations to accurately assess the reliability and voluntariness of confessions. *Drake Law Review, 52*, 619–646.

Dror, I. E., & Charlton, D. (2006). Why experts make errors. *Journal of Forensic Identification, 56*, 600–616.

Egan, C. (2006, February 22). A murderer no more. *The Australian Newspaper*, p. 13.

Escobedo v. Illinois, 378 U.S. 478 (1964).

Everington, C., & Fulero, S. (1999). Competence to confess: Measuring understanding and suggestibility of defendants with mental retardation. *Mental Retardation, 37*, 212–220.

Faigman, D. L., Kaye, D. H., Saks, M. J., & Sanders, J. (2002). *Science in the law: Forensic science issues*. St. Paul, MN: West.

Feeney, F. (2000). Police clearances: A poor way to measure the impact of *Miranda* on the police. *Rutgers Law Review, 32*, 1–114.

Feld, B. (1999). *Bad kids: Race and the transformation of the juvenile court*. New York: Oxford University Press.

Feld, B. (2006a). Juveniles' competence to exercise Miranda rights: An empirical study of policy and practice. *Minnesota Law Review, 91*, 26–100.

Feld, B. (2006b). Police interrogations of juveniles: An empirical study of policy and practice. *Journal of Criminal Law and Criminology, 97*, 219–316.

Findley, K. A., & Scott, M. S. (2006). The multiple dimensions of tunnel vision in criminal cases. *Wisconsin Law Review, 2006*, 291–397.

Firstman, R., & Salpeter, J. (2008). *A criminal injustice: A true crime, a false confession, and the fight to free Marty Tankleff*. New York: Ballantine Books.

Fisher, R. P., & Geiselman, R. E. (1992). *Memory enhancing techniques for investigative interviewing: The cognitive interview*. Springfield, IL: Thomas.

Forrest, K. D., Wadkins, T. A., & Miller, R. L. (2002). The role of preexisting stress on false confessions: An empirical study. *Journal of Credibility Assessment and Witness Psychology, 3*, 23–45.

Frank, J., & Frank, B. (1957). *Not guilty*. New York: Doubleday.

Frazier v. Cupp, 394 U.S. 731 (1969).

Fulero, S., & Everington, C. (1995). Assessing competency to waive Miranda rights in defendants with mental retardation. *Law and Human Behavior, 19*, 533–545.

Fulero, S., & Everington, C. (2004). Mental retardation, competency to waive Miranda rights, and false confessions. In G. D. Lassiter

(Ed.), *Interrogations, confessions, and entrapment* (pp. 163–179). New York: Kluwer Academic/Plenum.

Gall, J. A., & Freckelton, I. (1999). Fitness for interview: Current trends, views and an approach to the assessment procedure. *Journal of Clinical Forensic Medicine, 6*, 213–223.

Garner, B. A. (Ed.). (2004). *Black's law dictionary* (8th ed.). Eagan, MN: West.

Garrett, B. (2008). Judging innocence. *Columbia Law Review, 108*, 55–142.

Garven, S., Wood, J. M., & Malpass, R. S. (2000). Allegations of wrongdoing: The effects of reinforcement on children's mundane and fantastic claims. *Journal of Applied Psychology, 85*, 38–49.

Geller, W. A. (1993). *Videotaping interrogations and confessions: National Institute of Justice Research in Brief*. Washington, DC: U.S. Department of Justice.

Geller, W. A. (1994, January). Videotaping interrogations and confessions. *FBI Law Enforcement Bulletin.*

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*, 21–38.

Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The illusion of transparency: Biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology, 75*, 332–346.

Gohara, M. (2006). A lie for a lie: False confessions and the case for reconsidering the legality of deceptive interrogation techniques. *Fordham Urban Law Journal, 33*, 791–842.

Goldstein, N., Condie, L., Kalbeitzer, R., Osman, D., & Geier, J. (2003). Juvenile offenders' Miranda rights comprehension and self-reported likelihood of false confessions. *Assessment, 10*, 359–369.

Gordon, N. J., & Fleisher, W. L. (2006). *Effective interviewing and interrogation techniques* (2nd ed.). San Diego, CA: Academic Press.

Grano, J. D. (1994). *Confessions, truth, and the law*. Ann Arbor, MI: University of Michigan Press.

Grisso, T. (1980). Juveniles' capacities to waive Miranda rights: An empirical analysis. *California Law Review, 68*, 1134–1166.

Grisso, T. (1981). *Juveniles' waiver of rights: Legal and psychological competence*. New York: Plenum.

Grisso, T. (1986). *Evaluating competencies. Forensic assessments and instruments*. New York: Plenum.

Grisso, T. (1996). Society's retributive response to juvenile violence: A developmental perspective. *Law and Human Behavior, 20*, 229–247.

Grisso, T., & Ring, J. (1979). Parents' attitudes toward juveniles' rights in interrogation. *Criminal Justice and Behavior, 6*, 221–226.

Grisso, T., & Schwartz, R. (Eds.). (2000). *Youth on trial: A developmental perspective on juvenile justice*. Chicago: University of Chicago Press.

Grisso, T., Steinberg, L., Woolard, J., Cauffman, E., Scott, E., Graham, S., et al. (2003). Juveniles' competence to stand trial: A comparison of adolescents' and adults' capacities as trial defendants. *Law and Human Behavior, 27*(4), 333–363.

Gross, S. R., Jacoby, K., Matheson, D. J., Montgomery, N., & Patel, S. (2005). Exonerations in the United States, 1989 through 2003. *Journal of Criminal Law & Criminology, 95*, 523–553.

Gudjonsson, G. H. (1991). The effects of intelligence and memory on group differences in suggestibility and compliance. *Personality and Individual Differences, 5*, 503–505.

Gudjonsson, G. H. (1992). *The psychology of interrogations, confessions, and testimony*. London: Wiley.

Gudjonsson, G. H. (2003). *The psychology of interrogations and confessions: A handbook*. Chichester, England: Wiley.

Gudjonsson, G. H. (2005). Fitness to be interviewed. In J. Payne-James, R. W. Byard, T. S. Corey, & C. Henderson (Eds.), *Encyclopedia of forensic and legal medicine* (Vol. 2, pp. 169–174). London: Elsevier.

Gudjonsson, G. H. (2006). Disputed confessions and miscarriages of justice in Britain: Expert psychological and psychiatric evidence in the court of appeal. *The Manitoba Law Journal, 31*, 489–521.

Gudjonsson, G. H., Clare, I., Rutter, S., & Pearse, J. (1993). *Persons at risk during interviews in police custody: The identification of vulnerabilities*. London: HMSO.

Gudjonsson, G. H., & Clare, I. C. H. (1995). The relationship between confabulation and intellectual ability, memory, interrogative suggestibility, and acquiescence. *Personality and Individual Differences, 3*, 333–338.

Gudjonsson, G. H., & Grisso, T. (2008). Legal competencies in relation to confession evidence. In A. R. Felthous & H. Sass (Eds.), *International handbook on psychopathic disorders and the law* (Vol. 2, pp. 177–187). New York: Wiley.

Gudjonsson, G. H., & MacKeith, J. A. C. (1982). False confessions: Psychological effects of interrogation. In A. Trankell (Ed.), *Reconstructing the past: The role of psychologists in criminal trials* (pp. 253–269). Deventer, The Netherlands: Kluwer.

Gudjonsson, G. H., & MacKeith, J. A. C. (1994). Learning disability and the Police and Criminal Evidence Act of 1984. Protection during investigative interviewing: A video-recorded false confession to double murder. *Journal of Forensic Psychiatry, 5*, 35–49.

Gudjonsson, G. H., & Sigurdsson, J. F. (1999). The Gudjonsson Confession Questionnaire-Revised (GCQ-R). Factor structure and its relationship with personality. *Personality and Individual Differences, 27*, 953–968.

Gudjonsson, G. H., Sigurdsson, J. F., Asgeirsdottir, B. B., & Sigfusdottir, I. D. (2006). Custodial interrogation, false confession, and individual differences: A national study among Icelandic youth. *Personality and Individuals Differences, 41*, 49–59.

Gudjonsson, G. H., Sigurdsson, J. F., Asgeirsdottir, B. B., & Sigfusdottir, I. D. (2007a). Custodial interrogation: What are the background factors associated with claimed false confessions? *The Journal of Forensic Psychiatry and Psychology, 18*, 266–275.

Gudjonsson, G. H., Sigurdsson, J. F., Asgeirsdottir, B. B., & Sigfusdottir, I. D. (in press). Interrogation and false confession among adolescents in seven European countries: What background and psychological variables best discriminate between false confessors and non-false confessors? *Psychology, Crime, and Law*.

Gudjonsson, G. H., Sigurdsson, J. F., Bragason, O., Einarsson, E., & Valdimarsdottir, E. B. (2004). Confessions and denials and the relationship with personality. *Legal and Criminological Psychology, 9*, 121–133.

Gudjonsson, G. H., Sigurdsson, J. F., & Einarsson, E. (2004). The role of personality in relation to confessions and denials. *Psychology, Crime and Law, 10*, 125–135.

Gudjonsson, G. H., Sigurdsson, J. F., Einarsson, E., Bragason, O. O., & Newton, A. K. (2008). Interrogative suggestibility, compliance and false confessions among prison inmates and their relationship with attention deficit hyperactivity disorder (ADHD) symptoms. *Psychological Medicine, 38*, 1037–1044.

Gudjonsson, G. H., Young, S., & Bramham, J. (2007b). Interrogative suggestibility in adults diagnosed with attention-deficit hyperactivity disorder (ADHD). A potential vulnerability during police questioning. *Personality and Individual Differences, 43*, 737–745.

Haley v. Ohio, 332 U.S. 596 (1948).

Harris, R. J., & Monaco, G. E. (1978). Psychology of pragmatic implication: Information processing between the lines. *Journal of Experimental Psychology: General, 107*, 1–22.

Harrison, Y., & Horne, J. A. (2000). The impact of sleep deprivation on decision making: A review. *Journal of Experimental Psychology: Applied, 6*, 236–249.

Hartwig, M., Granhag, P. A., Strömwall, L. A., & Kronkvist, O. (2006). Strategic use of evidence during police interviews: When training to detect deception works. *Law and Human Behavior, 30*, 603–619.

Hartwig, M., Granhag, P. A., Strömwall, L., & Vrij, A. (2005). Detecting deception via strategic closure of evidence. *Law and Human Behavior, 29*, 469–484.

Hasel, L. E., & Kassin, S. M. (2009). On the presumption of evidentiary independence: Can confessions corrupt eyewitness identifications? *Psychological Science, 20*, 122–126.

Hastie, R., & Dawes, R. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision-making*. Thousand Oaks, CA: Sage.

Henkel, L. A., Coffman, K. A. J., & Dailey, E. M. (2008). A survey of people's attitudes and beliefs about false confessions. *Behavioral Sciences and the Law, 26*, 555–584.

Herrnstein, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior, 7*, 243–266.

Herrnstein, R. J., Rachlin, H., & Laibson, D. I. (Eds.). (1997). *The matching law: Papers in psychology and economics*. New York: Russell Sage Foundation.

Hershkowitz, I., Horowitz, D., Lamb, M. E., Orbach, Y., & Sternberg, K. J. (2004). Interviewing youthful suspects in alleged sex crimes: A descriptive analysis. *Child Abuse and Neglect, 28*, 423–438.

Hill, C., Memon, A., & McGeorge, P. (2008). The role of confirmation bias in suspect interviews: A systematic evaluation. *Legal and Criminological Psychology, 13*, 357–371.

Hilton, D. J. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin, 118*, 248–271.

Hirsch, A. (2005a). Threats, promises, and false confessions: Lessons of slavery. *Howard Law Journal, 49*, 31–60.

Hirsch, A. (2005b). The tragedy of false confessions and a common sense proposal (book review). *North Dakota Law Review, 81*, 343–350.

Home Office. (1985). *Police and Criminal Evidence Act 1984*. London: HMSO.

Home Office. (2003). *Police and Criminal Evidence Act 1984. Codes of Practice A-E Revised Edition*. London: HMSO.

Hopkins, E. J. (1931). *Our lawless police: A study of the unlawful enforcement of the law*. New York: Viking Press.

Hopt v. Utah, 110 U.S. 574 (1884).

Horselenberg, R., Merckelbach, H., & Josephs, S. (2003). Individual differences and false confessions: A conceptual replication of Kassin and Kiechel (1996). *Psychology, Crime and Law, 9*, 1–18.

Horselenberg, R., Merckelbach, H., Smeets, T., Franssens, D., Ygram Peters, G.-J., & Zeles, G. (2006). False confessions in the lab: Do plausibility and consequences matter? *Psychology, Crime and Law, 12*, 61–75.

Hourihan, P. (1995). Earl Washington's confession: Mental retardation and the law of confessions. *Virginia Law Review, 81*, 1473–1501.

Haynes v. Washington, 373 U.S. 503 (1963).

In re Gault, 387 U.S. 1 (1967).

Inbau, F. E. (1991). Miranda's immunization of low intelligence offenders. *The Prosecutor: Journal of the National District Attorney's Association, 24*(Spring), 9–10.

Inbau, F. E., Reid, J. E., Buckley, J. P., & Jayne, B. C. (2001). *Criminal interrogation and confessions* (4th ed.). Gaithersberg, MD: Aspen.

Inglis, T. (2004). *Truth power and lies: Irish society and the case of the Kerry babies*. Dublin: University College Dublin Press.

Irving, B. (1980). *Police interrogation. A case study of current practice. Research Studies No 2*. London: HMSO.

Irving, B., & McKenzie, I. K. (1989). *Police interrogation: The effects of the Police and Criminal Evidence Act*. London: Police Foundation G.B.

James, D. J., & Glaze, L. E. (2006). *Mental health problems of prison and jail inmates*. Washington, DC: U.S. Dept of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Jones, E. E. (1990). *Interpersonal perception*. New York: Freeman.

Junkin, T. (2004). *Bloodsworth: The true story of the first death row inmate exonerated by DNA*. Chapel Hill, NC: Algonquin Books.

The Justice Project. (2007). *Electronic recording of custodial interrogations: A policy review*. Washington, DC: The Justice Project.

Kahn, J. (2005, September 21). Deep flaws, and little justice, in China's court system. *The New York Times*.

Kahn, R., Zapf, P., & Cooper, V. (2006). Readability of Miranda warnings and waivers: Implications for evaluating Miranda comprehension. *Law & Psychology Review, 30*, 119–142.

Kamisar, Y. (1963). What is an "involuntary" confession? Some comments on Inbau and Reid's criminal interrogation and confessions. *Rutgers Law Review, 17*, 728–732.

Kamisar, Y. (1977). Foreword: Brewer v. Williams—A hard look at a discomfiting record. *Georgetown Law Journal, 66*, 209–243.

Karlsen, C. F. (1989). *The devil in the shape of a woman: Witchcraft in colonial New England*. New York: Vintage.

Kassin, S. M. (1997). The psychology of confession evidence. *American Psychologist, 52*, 221–233.

Kassin, S. M. (2002, November 1). False confessions and the jogger case, *New York Times*, p. A31.

Kassin, S. M. (2005). On the psychology of confessions: Does *innocence* put *innocents* at risk? *American Psychologist, 60*, 215–228.

Kassin, S. M. (2006). A critical appraisal of modern police interrogations. In T. Williamson (Ed.), *Investigative interviewing: Rights, research, regulation* (pp. 207–228). Devon, UK: Willan.

Kassin, S. M. (2008). False confessions: Causes, consequences, and implications for reform. *Current Directions in Psychological Science, 17*(4), 249–253.

Kassin, S. M., Goldstein, C. J., & Savitsky, K. (2003). Behavioral confirmation in the interrogation room: On the dangers of presuming guilt. *Law and Human Behavior, 27*, 187–203.

Kassin, S. M., & Gudjonsson, G. H. (2004). The psychology of confession evidence: A review of the literature and issues. *Psychological Science in the Public Interest, 5*, 35–69.

Kassin, S. M., & Kiechel, K. L. (1996). The social psychology of false confessions: Compliance, internalization, and confabulation. *Psychological Science, 7*, 125–128.

Kassin, S. M., Leo, R. A., Meissner, C. A., Richman, K. D., Colwell, L. H., Leach, A.-M., et al. (2007). Police interviewing and interrogation: A Self-report survey of police practices and beliefs. *Law and Human Behavior, 31*, 381–400.

Kassin, S. M., & McNall, K. (1991). Police interrogations and confessions: Communicating promises and threats by pragmatic implication. *Law and Human Behavior, 15*, 233–251.

Kassin, S. M., Meissner, C. A., & Norwick, R. J. (2005). "I'd know a false confession if I saw one": A comparative study of college students and police investigators. *Law and Human Behavior, 29*, 211–227.

Kassin, S. M., & Neumann, K. (1997). On the power of confession evidence: An experimental test of the "fundamental difference" hypothesis. *Law and Human Behavior, 21*, 469–484.

Kassin, S., & Norwick, R. (2004). Why people waive their Miranda rights: The power of innocence. *Law and Human Behavior, 28*, 211–221.

Kassin, S. M., & Sukel, H. (1997). Coerced confessions and the jury: An experimental test of the "harmless error" rule. *Law and Human Behavior, 21*, 27–46.

Kassin, S. M., & Wrightsman, L. S. (1980). Prior confessions and mock juror verdicts. *Journal of Applied Social Psychology, 10*, 133–146.

Kassin, S. M., & Wrightsman, L. S. (1985). Confession evidence. In S. Kassin & L. Wrightsman (Eds.), *The psychology of evidence and trial procedure* (pp. 67–94). Beverly Hills, CA: Sage.

The King v. Warrickshall (1793), 168 Eng. Rep. 234, 234-35 (K.B. 1783).

Klaver, J., Lee, Z., & Rose, V. G. (2008). Effects of personality, interrogation techniques and plausibility in an experimental false confession paradigm. *Legal and Criminological Psychology, 13*, 71–88.

Kollins, S. H. (2003). Delay discounting is associated with substance use in college students. *Addictive Behaviors, 28*, 1167–1173.

Lamb, H. R., & Weinberger, L. E. (1998). Persons with severe mental illness in jails and prisons: A review. *Psychiatric Services, 49*, 483–492.

Lamb, M. E., Orbach, Y., Hershkowitz, I., Horowitz, D., & Abbott, C. B. (2007). Does the type of prompt affect the accuracy of information provided by alleged victims of abuse in forensic interviews? *Applied Cognitive Psychology, 21*, 1117–1130.

Lamb, M. E., Orbach, Y., Sternberg, K. J., Hershkowitz, I., & Horowitz, D. (2000). Accuracy of investigators' verbatim notes of their forensic interviews with alleged child abuse victims. *Law and Human Behavior, 24*, 699–707.

Lambert, B. (2008, July 1). Freed after 17 years in prison, L.I. man will not face new trial. *The New York Times*, p. A1.

Lassiter, G. D. (Ed.). (2004). *Interrogations, confessions, and entrapment*. New York: Kluwer Academic.

Lassiter, G. D., Diamond, S. S., Schmidt, H. C., & Elek, J. K. (2007). Evaluating videotaped confessions: Expertise provides no defense against the camera-perspective effect. *Psychological Science, 18*, 224–226.

Lassiter, G. D., & Geers, A. L. (2004). Evaluation of confession evidence: Effects of presentation format. In G. D. Lassiter (Ed.), *Interrogations, confessions, and entrapment* (pp. 197–214). New York: Kluwer Academic.

Lassiter, G. D., Geers, A. L., Handley, I. M., Weiland, P. E., & Munhall, P. J. (2002). Videotaped confessions and interrogations: A change in camera perspective alters verdicts in simulated trials. *Journal of Applied Psychology, 87*, 867–874.

Lassiter, G. D., Geers, A. L., Munhall, P. J., Handley, I. M., & Beers, M. J. (2001). Videotaped confessions: Is guilt in the eye of the camera? *Advances in Experimental Social Psychology, 33*, 189–254.

Lassiter, G. D., & Irvine, A. A. (1986). Videotaped confessions: The impact of camera point of view on judgments of coercion. *Journal of Applied Social Psychology, 16*, 268–276.

Lassiter, G. D., Slaw, R. D., Briggs, M. A., & Scanlan, C. R. (1992). The potential for bias in videotaped confessions. *Journal of Applied Social Psychology, 22*, 1838–1851.

Latane, B. (1981). The psychology of social impact. *American Psychologist, 36*, 343–356.

Leo, R. A. (1996a). Miranda's revenge: Police interrogation as a confidence game. *Law and Society Review, 30*, 259–288.

Leo, R. A. (1996b). Inside the interrogation room. *Journal of Criminal Law and Criminology, 86*, 266–303.

Leo, R. A. (1996c). The impact of Miranda revisited. *The Journal of Criminal Law and Criminology, 86*, 621–692.

Leo, R. A. (2004). The third degree and the origins of psychological police interrogation in the United States. In G. D. Lassiter (Ed.), *Interrogations, confessions, and entrapment* (pp. 37–84). New York: Kluwer Academic.

Leo, R. A. (2005). Re-thinking the study of miscarriages of justice: Developing a criminology of wrongful conviction. *Journal of Contemporary Criminal Justice, 21*, 201–223.

Leo, R. A. (2008). *Police interrogation and American justice*. Cambridge, MA: Harvard University Press.

Leo, R. A., Drizin, S., Neufeld, P., Hall, B., & Vatner, A. (2006). Bringing reliability back in: False confessions and legal safeguards in the twenty-first century. *Wisconsin Law Review, 2006*, 479–539.

Leo, R. A., & Liu, B. (2009). What do potential jurors know about police interrogation techniques and false confessions? *Behavioral Sciences and the Law, 27*(3), 381–399.

Leo, R. A., & Ofshe, R. J. (1998). The consequences of false confessions: Deprivations of liberty and miscarriages of justice in the age of psychological interrogation. *Journal of Criminal Law and Criminology, 88*, 429–496.

Leo, R. A., & Ofshe, R. J. (2001). The truth about false confessions and advocacy scholarship. *The Criminal Law Bulletin, 37*, 293–370.

Lerner, M. J. (1980). *The belief in a just world*. New York: Plenum.

Leyra v. Denno, 347 U.S. 556 (1954).

Loftus, E. F. (1997). Creating false memories. *Scientific American, 277*, 70–75.

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory, 12*, 361–366.

Lykken, D. T. (1998). *A tremor in the blood: Uses and abuses of the lie detector*. Reading, MA: Perseus Books.

Lynumn v. Illinois, 372 U.S. 528 (1963).

Lyznicki, J. M., Doege, T. C., Davis, R. M., & Williams, M. A. (1998). Sleepiness, driving, and motor vehicle crashes. *Journal of the American Medical Association, 279*, 1908–1913.

Magid, L. (2001). Deceptive police interrogation practices: How far is too far? *Michigan Law Review, 99*, 1168.

McCann, J. T. (1998). A conceptual framework for identifying various types of confessions. *Behavioral Sciences and the Law, 16*, 441–453.

McCormick, C. T. (1972). *Handbook of the law of evidence* (2nd ed.). St. Paul, MN: West.

Meissner, C. A., & Kassin, S. M. (2002). "He's guilty!": Investigator bias in judgments of truth and deception. *Law and Human Behavior, 26*, 469–480.

Melton, G., Petrila, J., Poythress, N., & Slobogin, C. (1997). *Psychological evaluations for the courts* (2nd ed.). New York: Guilford.

Meyer, J. R., & Reppucci, N. D. (2007). Police practices and perceptions regarding juvenile interrogations and interrogative suggestibility. *Behavioral Sciences and the Law, 25*, 757–780.

Meyer, R. G., & Youngjohn, J. R. (1991). Effects of feedback and validity expectancy on response in a lie detector interview. *Forensic Reports, 4*, 235–244.

Milgram, S. (1974). *Obedience to authority: An experimental view*. New York: Harper & Row.

Miller, D. T., & McFarland, C. (1987). Pluralistic ignorance: When similarity is interpreted as dissimilarity. *Journal of Personality and Social Psychology, 53*, 298–305.

Milne, R., & Bull, R. (1999). *Investigative interviewing: Psychology and practice*. Chichester, England: Wiley.

Miranda v. Arizona, 384 U.S. 436 (1966).

Missouri v. Seibert, 542 U.S. (2004).

Morgan, C. A., Hazlett, G., Doran, A., Garrett, S., Hoyt, G., Thomas, P., et al. (2004). Accuracy of eyewitness memory for persons encountered during exposure to highly intense stress. *International Journal of Law and Psychiatry, 27*, 265–279.

Moston, S., Stephenson, G. M., & Williamson, T. (1992). The effects of case characteristics on suspect behaviour during police questioning. *British Journal of Criminology, 32*, 23–39.

Munsterberg, H. (1908). *On the witness stand*. Garden City, NY: Doubleday.

Nash, R. A., & Wade, K. A. (2009). Innocent but proven guilty: Using false video evidence to elicit false confessions and create false beliefs. *Applied Cognitive Psychology, 23*, 624–637.

National Research Council, Committee to Review the Scientific Evidence on the Polygraph, Division of Behavioral and Social Sciences and Education. (2003). *The polygraph and lie detection*. Washington, DC: National Academies Press.

Navarick, D. J. (1982). Negative reinforcement and choice in humans. *Learning and Motivation, 13*, 361–377.

Nelson, N. P. (2007, March). *Interviewing using the RIP technique*. Paper presented at "Off the Witness Stand, Using Psychology in the Practice of Justice," John Jay College of Criminal Justice, New York City.

Neuschatz, J. S., Lawson, D. S., Swanner, J. K., Meissner, C. A., & Neuschatz, J. S. (2008). The effects of accomplice witnesses and jailhouse informants on jury decision making. *Law and Human Behavior, 32*, 137–149.

Oberlander, L., Goldstein, N., & Goldstein, A. (2003). Competence to confess. In I. Wiener & A. Goldstein (Eds.), *Handbook of psychology: Volume 22, forensic psychology* (pp. 335–357). Hoboken, NJ: Wiley.

Oberlander, L. B., & Goldstein, N. E. (2001). A review and update on the practice of evaluating Miranda comprehension. *Behavioral Sciences and the Law, 19*, 453–471.

O'Connell, M. J., Garmoe, W., & Goldstein, N. E. S. (2005). Miranda comprehension in adults with mental retardation and the effects of feedback style on suggestibility. *Law and Human Behavior, 29*, 359–369.

Ofshe, R. J., & Leo, R. A. (1997a). The social psychology of police interrogation: The theory and classification of true and false confessions. *Studies in Law, Politics, and Society, 16*, 189–251.

Ofshe, R. J., & Leo, R. A. (1997b). The decision to confess falsely: Rational choice and irrational action. *Denver University Law Review, 74*, 979–1122.

O'Hara, C. (1956). *Fundamentals of criminal investigation*. Springfield, IL: Charles C. Thomas.

Onishi, N. (2007, May 11). Pressed by police, even innocent confess in Japan. *The New York Times*.

Opper v. United States, 348 U.S. 84 (1954).

Otto, H. D. (2006). *"Im namen des irrtums!" Fehlurteile in mordprozessen*. Mänchen: F.A. Herbig.

Owen-Kostelnik, J., Reppucci, N. D., & Meyer, J. D. (2006). Testimony and interrogation of minors: Assumptions about maturity and morality. *American Psychologist, 61*, 286–304.

Parke, R. D., Ornstein, P. A., Reiser, J. J., & Zahn-Waxler, C. (1994). *A century of developmental psychology*. Washington, DC: APA.

Pearse, J., & Gudjonsson, G. H. (1996). Police interviewing techniques at two south London police stations. *Psychology, Crime and Law, 3*, 63–74.

Penney, S. (1998). Theories of confession admissibility: A historical view. *American Journal of Criminal Law, 25*, 309–383.

People of the State of New York v. Kharey Wise, Kevin Richardson, Antron McCray, Yusef Salaam, & Raymond Santana: Affirmation in Response to Motion to Vacate Judgment of Conviction (2002). Indictment No. 4762/89, December 5, 2002.

People v. Daoud, 614 N.W.2d 152 (Mich. S. Ct., 2000).

Perske, R. (2004). Understanding persons with intellectual disabilities in the criminal justice system: Indicators of progress? *Mental Retardation, 42*, 484–487.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer.

Pierce v. United States, 160 U.S. 355 (1896).

Pilcher, J. J., & Huffcut, A. (1996). Effects of sleep deprivation on performance: A meta-analysis. *Sleep, 19*, 318–326.

Price, D. D., Finniss, D. G., & Benedetti, F. (2008). A comprehensive review of the placebo effect: Recent advances and current thought. *Annual Review of Psychology, 59*, 565–590.

Rachlin, H. (2000). *The science of self-control*. Cambridge, MA: Harvard University Press.

Radelet, M., Bedau, H., & Putnam, C. (1992). *In spite of innocence: Erroneous convictions in capital cases*. Boston: Northeastern University Press.

Redlich, A. D. (2004). Mental illness, police interrogations, and the potential for false confession. *Psychiatric Services, 55*, 19–21.

Redlich, A. D. (2007). Double jeopardy in the interrogation room: Young age and mental illness. *American Psychologist, 62*, 609–611.

Redlich, A. D. (in press). False confessions and false guilty pleas. In G. D. Lassiter & C. A. Meissner (Eds.), *Interrogations and confessions: Current research, practice and policy*. Washington, DC: APA Books.

Redlich, A. D., & Drizin, S. (2007). Police interrogation of youth. In C. L. Kessler & L. Kraus (Eds.), *The mental health needs of young offenders: Forging paths toward reintegration and rehabilitation* (pp. 61–78). Cambridge, England: Cambridge University Press.

Redlich, A. D., Ghetti, S., & Quas, J. A. (2008a). Perceptions of children during a police interview: A comparison of suspects and alleged victims. *Journal of Applied Social Psychology, 38*, 705–735.

Redlich, A. D., & Goodman, G. S. (2003). Taking responsibility for an act not committed: Influence of age and suggestibility. *Law and Human Behavior, 27*, 141–156.

Redlich, A. D., Quas, J. A., & Ghetti, S. (2008b). Perceptions of children during a police interview: Guilt, confessions, and interview fairness. *Psychology, Crime and Law, 14*, 201–223.

Redlich, A. D., Silverman, M., Chen, J., & Steiner, H. (2004). The police interrogation of children and adolescents. In G. D. Lassiter (Ed.), *Interrogations, confessions, and entrapment* (pp. 107–125). New York: Kluwer Academic.

Redlich, A. D., Silverman, M., & Steiner, H. (2003). Pre-adjudicative and adjudicative competence in juveniles and young adults. *Behavioral Sciences and the Law, 21*, 393–410.

Reynolds, B., Richards, J. B., Horn, K., & Karraker, K. (2004). Delay discounting and probability discounting as related to cigarette smoking status in adults. *Behavioral Processes, 65*, 35–42.

Rigoni, M. E., & Meissner, C. A. (2008). *Is it time for a revolution in the interrogation room? Empirically validating inquisitorial methods*. Paper presented at Meeting of the American Psychology-Law Society, Jacksonville, FL.

Roberts, P. (2007). Law and criminal investigation. In T. Newburn, T. Williamson, & A. Wright (Eds.), *Handbook of criminal investigation* (pp. 92–145). Devon, UK: Willan.

Robertson, G. D. (July 24, 2007 NC 01:44:29). NC lawmakers approve lineup, interrogation recording standards. *AP-Alerts.*

Rofe, Y. (1984). Stress and affiliation: A utility theory. *Psychological Review, 91*, 235–250.

Rogers, R., Harrison, K., Hazelwood, L., & Sewell, K. (2007a). Knowing and intelligent: A study of Miranda warnings in mentally disordered defendants. *Law and Human Behavior, 31*, 401–418.

Rogers, R., Harrison, K., Shuman, D., Sewell, K., & Hazelwood, L. (2007b). An analysis of Miranda warnings and waivers: Comprehension and coverage. *Law and Human Behavior, 31*, 177–192.

Rogers, R., Hazelwood, L., Sewell, K., Harrison, K., & Shuman, D. (2008). The language of Miranda warnings in American jurisdictions: A replication and vocabulary analysis. *Law and Human Behavior, 32*, 124–136.

Roper v. Simmons, 543 U.S. 551 (2005).

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. New York: Holt Rinehart & Winston.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology, 10*, 174–221.

Russano, M. B., Meissner, C. A., Narchet, F. M., & Kassin, S. M. (2005). Investigating true and false confessions within a novel experimental paradigm. *Psychological Science, 16*, 481–486.

Santos, F. (2006, September 21). DNA evidence frees a man imprisoned for half his life. *New York Times*, p. A1.

Santtila, P., Alkiora, P., Ekholm, M., & Niemi, P. (1999). False confessions to robbery: The role of suggestibility, anxiety, memory disturbance and withdrawal symptoms. *The Journal of Forensic Psychiatry, 10*, 399–415.

Saywitz, K., Nathanson, R., & Snyder, L. S. (1993). Credibility of child witnesses: The role of communicative competence. *Topics in Language Disorders, 13*, 59–78.

Schachter, S. (1959). *The psychology of affiliation: Experimental studies of the sources of gregariousness*. Stanford, CA: Stanford University Press.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69*, 379–399.

Schacter, D. L. (2001). *The seven sins of memory: How the mind forgets and remembers*. Boston: Houghton Mifflin.

Scheck, B., Neufeld, P., & Dwyer, J. (2000). *Actual innocence*. Garden City, NY: Doubleday.

Schulhofer, S. (1981). Confessions and the court. *Michigan Law Review, 79*, 865–893.

Schulhofer, S. J. (1996). Miranda's practical effect: Substantial benefits and vanishingly small social costs. *Northwestern University Law Review, 90*, 500–564.

Sherif, M. (1936). *The psychology of social norms*. New York: Harper.

Sherrer, H. (2005). Murdered woman's innocent boyfriend exonerated after bizarre "confession" is exposed as false. *Justice: Denied*, January 2005.

Sigurdsson, J., & Gudjonsson, G. (1997). The criminal history of 'false confessors' and other prison inmates. *Journal of Forensic Psychiatry, 8*, 447–455.

Sigurdsson, J. F., & Gudjonsson, G. H. (1996). The psychological characteristics of false confessors: A study among Icelandic prison inmates and juvenile offenders. *Personality and Individual Differences, 20*, 321–329.

Sigurdsson, J. F., & Gudjonsson, G. H. (2001). False confessions: The relative importance of psychological, criminological and substance abuse variables. *Psychology, Crime and Law, 7*, 275–289.

Sigurdsson, J. F., & Gudjonsson, G. H. (2004). Forensic psychology in Iceland: A survey of members of the Icelandic Psychological Society. *Scandinavian Journal of Psychology, 45*, 325–329.

Sigurdsson, J. F., Gudjonsson, G. H., Einarsson, E., & Gudjonsson, G. (2006). Differences in personality and mental state between suspects and witnesses immediately after being interviewed by the police. *Psychology, Crime and Law, 12*, 619–628.

Simon, D. (1991). *Homicide: A year on the killing streets*. New York: Ivy Books.

Singh, K., & Gudjonsson, G. (1992). Interrogative suggestibility among adolescent boys and its relationship to intelligence, memory, and cognitive set. *Journal of Adolescence, 15*, 155–161.

Skinner, B. F. (1938). *The behavior of organisms*. New York: Appleton-Century-Crofts.

Skolnick, J. H., & Leo, R. A. (1992). The ethics of deceptive interrogation. *Criminal Justice Ethics, 11*, 3–12.

Slobogin, C. (2003). Toward taping. *Ohio State Journal of Criminal Law, 1*, 309–322.

Slobogin, C. (2007). Lying and confessing. *Texas Tech Law Review, 39*, 1275–1292.

Smith v. United States, 348 U.S. 147 (1954).

Snyder, H. (2006). *Juvenile arrests 2004*. Washington, DC: Office of Juvenile Justice and Delinquency Prevention, Office of Justice Programs.

Soukara, S., Bull, R., Vrij, A., Turner, M., & Cherryman, C. (in press). A study of what really happens in police interviews with suspects. *Psychology, Crime and Law*.

Spano v. New York, 360 U.S. 315 (1959).

Sparf v. United States, 156 U.S. 51 (1895).

State v. Barnett, 789 A.2d 629–633 (N.H. 2002).

State v. Cayward, 552 So. 2d 921 (Fla. 1989).

State v. Chirokovskcic, 860 A.2d 986 (N.J.Super.2004).

State v. Hajtic, 724 N.W.2d 449,455 (Ia. 2006).

State v. Mauchley, 67 P.3d 477 (Utah 2003).

State v. Patton, 826 A.2d 783, N.J. Super. A.D. (2003).

State v. Scales, 518 N.W.2d 587 (Minn. 1994).

Stephan v. State, 711 P.2d 1156 (Alaska 1985).

Steinberg, L. (2005). Cognitive and affective development in adolescence. *Trends in Cognitive Sciences, 9*, 69–74.

Steinberg, L. (2007). Risk taking in adolescence: New perspectives from brain and behavioral science. *Current Directions in Psychological Science, 16*, 55–59.

Steinberg, L., & Cauffman, E. (1996). Maturity of judgment in adolescence: Psychosocial factors in adolescent decision making. *Law and Human Behavior, 20*, 249–272.

Steinberg, L., & Morris, A. S. (2001). Adolescent development. *Annual Review of Psychology, 52*, 83–110.

Steingrimsdottir, G., Hreinsdottir, H., Gudjonsson, G. H., Sigurdsson, J. F., & Nielsen, T. (2007). False confessions and the relationship with offending behaviour and personality among Danish adolescents. *Legal and Criminological Psychology, 12*, 287–296.

Suedfeld, P. (Ed.). (1990). *Psychology and torture*. Washington, DC: Hemisphere.

Sullivan, T. (2007, April). Federal law enforcement agencies should record custodial interrogations. *The Champion*, 8–12.

Sullivan, T. P. (2004). *Police experiences with recording custodial interrogations*. Chicago: Northwestern University Law School, Center on Wrongful Convictions.

Sullivan, Y. P., Vail, A. W., & Anderson, H. W. (2008). The case for recording police interrogation. *Litigation, 34*(3), 1–8.

Swanner, J. K., Beike, D. R., & Cole, A. T. (in press). Snitching, lies and computer crashes: An experimental investigation of secondary confessions. *Law and Human Behavior*.

Taffinder, N. J., McManus, I. C., Gul, Y., Russell, R. C., & Darzi, A. (1998). Effect of sleep deprivation on surgeons' dexterity on laparoscopy simulator. *Lancet, 352*, 1191.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behavior. *European Journal of Social Psychology, 1*, 149–178.

Taylor, B. D. (2005). Evidence beyond confession: Abolish Arizona's corpus delicti rule. *Arizona Attorney, 41*, 22–28.

Technical Working Group for Eyewitness Evidence. (1999). *Eyewitness evidence: A guide for law enforcement*. Washington, DC: U.S. Department of Justice, Office of Justice Programs.

Thomas, G. C. (1996). Is Miranda a real-world failure? A plea for more (and better) empirical evidence. *UCLA Law Review, 43*, 821.

Thomas, G. C. (2007). Regulating police deception during interrogation. *Texas Tech Law Review*, *39*.

Thomas, G. C., & Leo, R. A. (2002). The effects of *Miranda v. Arizona*: "Embedded" in our national culture? *Crime and Justice: A Review of Research, 29*(20), 3–271.

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. New York: MacMillan.

Trainum, J. (2007, September 20). I took a false confession—So don't tell me it doesn't happen! *The California Majority Report*. http://www.camajorityreport.com/index.php?module=articles&func=display&ptid=9&aid=2306.

Trainum, J. (2008, October 24). The case for videotaping interrogations: A suspect's false confession to a murder opened an officer's eyes. *The Los Angeles Times.*

Uchino, B. N., Cacioppo, J. T., & Kiecolt-Glaser, J. K. (1996). The relationship between social support and physiological processes: A review with emphasis on underlying mechanisms and implications for health. *Psychological Bulletin, 119*, 488–531.

Valins, S. (1966). Cognitive effects of false heart-rate feedback. *Journal of Personality and Social Psychology, 4*, 400–408.

Veasey, S., Rosen, R., Barzansky, B., Rosen, I., & Owens, J. (2002). Sleep loss and fatigue in residency training: A reappraisal. *Journal of the American Medical Association, 288*, 1116–1124.

Viljoen, J., Klaver, J., & Roesch, R. (2005). Legal decisions of preadolescent and adolescent defendants: Predictors of confessions, pleas, communication with attorneys, and appeals. *Law and Human Behavior, 29*, 253–278.

Viljoen, J., & Roesch, R. (2005). Competence to waive interrogation rights and adjudicative competence in adolescent defendants: Cognitive development, attorney contact, and psychological symptoms. *Law and Human Behavior, 29*, 723–742.

Viljoen, J., Zapf, P., & Roesch, R. (2007). Adjudicative competence and comprehension of Miranda rights in adolescent defendants: A comparison of legal standards. *Behavioral Sciences and the Law, 25*, 1–19.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. Chichester, England: Wiley.

Wagenaar, W. A. (2002). False confessions after repeated interrogation: The Putten murder case. *European Review, 10*, 519–537.

Wald, M., Ayres, R., Hess, D. W., Schantz, M., & Whitebread, C. H. (1967). Interrogations in New Haven: The impact of Miranda. *The Yale Law Journal, 76*, 1519–1648.

Wall, S., & Furlong, J. (1985). Comprehension of Miranda rights by urban adolescents with law-related education. *Psychological Reports, 56*, 359–372.

Warden, R. (2005). *Wilkie Collins's the dead alive: The novel, the case, and wrongful convictions*. Evanston, IL: Northwestern University Press.

Weinger, M. B., & Ancoli-Israel, S. (2002). Sleep deprivation and clinical performance. *Journal of American Medical Association, 287*, 955–957.

Weisberg, B. (1961). Police interrogation of arrested persons: A skeptical view. In C. R. Sowle (Ed.), *Police power and individual freedom* (pp. 153–181). Chicago: Aldine.

Wells, G. L., Malpass, R. S., Lindsay, R. C. L., Fisher, R. P., Turtle, J. W., & Fulero, S. M. (2000). From the lab to the police station: A successful application of eyewitness research. *American Psychologist, 55*, 581–598.

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22*, 1–39.

White, W. S. (1997). False confessions and the constitution: Safeguards against untrustworthy confessions. *Harvard Civil Rights-Civil Liberties Law Review, 32*, 105–157.

White, W. S. (1998). What is an involuntary confession now? *Rutgers Law Review, 200*, 1–2057.

White, W. S. (2001). *Miranda's waning protections: Police interrogation practices after Dickerson*. Ann Arbor, MI: University of Michigan Press.

White, W. S. (2003). Confessions in capital cases. *University of Illinois Law Review, 2003*, 979–1036.

Wickersham Commission Report (1931). National Commission on Law Observance and Law Enforcement. (1931). *Report on lawlessness in law enforcement*. Washington, DC: U.S. Government Printing Office.

Wiggins, E. C., & Wheaton, S. (2004). So what's a concerned psychologist to Do? Translating the research on interrogations, confessions and entrapment into policy. In G. D. Lassiter (Ed.), *Interrogations, confessions, and entrapment* (pp. 265–280). New York: Springer.

Wigmore, J. H. (1970). *Evidence in trials at common law* (3rd ed.). Little, Brown: Boston.

Williamson, T. (Ed.). (2006). *Investigative interviewing: Rights, research, regulation*. Devon, UK: Willan.

Williamson, T. (2007). Psychology and criminal investigations. In T. Newburn, T. Williamson, & A. Wright (Eds.), *Handbook of criminal investigation* (pp. 68–91). Devon, UK: Willan.

Wills, C. (2005, July 17). Taped interrogations can still be false. *Los Angeles Times* (online edition).

Wilson v. United States, 162 U.S. 613 (1896).

Wisconsin Criminal Justice Study Commission. (2007). *Position paper on false confessions.* Madison, WI. www.wcjsc.org/Position_Paper_on_False_Confessions.pdf

Wrightsman, L. S., & Kassin, S. M. (1993). *Confessions in the courtroom*. Newbury Park, CA: Sage.

Young, S. (2007). Forensic aspects of ADHD. In M. Fitzgerald, M. Bellgrove, & M. Gill (Eds.), *Handbook of attention deficit hyperactive disorder*. Chichester, England: Wiley.

Zimbardo, P. G. (1967). The psychology of police confessions. *Psychology Today, 1*(17–20), 25–27.

Zimring, F. E., & Hawkins, G. (1986). *Capital punishment and the American agenda*. Cambridge, England: Cambridge University Press.

Why Confessions Trump Innocence

Saul M. Kassin
John Jay College of Criminal Justice, City University of New York

As illustrated by the story of Amanda Knox and many others wrongfully convicted, false confessions often trump factual innocence. Focusing on consequences, recent research suggests that confessions are powerfully persuasive as a matter of logic and common sense; that many false confessions contain richly detailed narratives and accurate crime facts that appear to betray guilty knowledge; and that confessions in general can corrupt other evidence from lay witnesses and forensic experts—producing an illusion of false support. This latter phenomenon, termed "corroboration inflation," suggests that pretrial corroboration requirements as well as the concept of "harmless error" on appeal are based on an erroneous presumption of independence among items of evidence. In addition to previously suggested reforms to police practices that are designed to curb the risk of false confessions, measures should be taken as well to minimize the rippling consequences of those confessions.

Keywords: confession, innocence, wrongful conviction

On November 2, 2007, British exchange student Meredith Kercher was found raped and murdered in Perugia, Italy. Almost immediately, police suspected 20-year-old Amanda Knox, an American student and one of Kercher's roommates—the only one who stayed in Perugia after the murder. Knox had no history of crime or violence and no motive. But something about her demeanor—such as an apparent lack of affect, an outburst of sobbing, or her girlish and immature behavior—led police to believe she was involved and lying when she claimed she was with Raffaele Sollecito, her new Italian boyfriend, that night.

Armed with a prejudgment of Knox's guilt, several police officials interrogated the girl on and off for four days. Her final interrogation started on November 5 at 10 p.m. and lasted until November 6 at 6 a.m., during which time she was alone, without an attorney, tag-teamed by a dozen police, and did not break for food or sleep. In many ways, Knox was a vulnerable suspect—young, far from home, without family, and forced to speak in a language in which she was not fluent. Knox says she was repeatedly threatened and called a liar. She was told, falsely, that Sollecito, her boyfriend, disavowed her alibi and that physical evidence placed her at the scene. She was encouraged to shut her eyes and imagine how the gruesome crime had occurred, a trauma, she was told, that she had obviously repressed. Eventually she broke down crying, screaming, and hitting herself in the head. Despite a law that mandates the recording of interrogations, police and prosecutors maintain that these sessions were not recorded.

Two "confessions" were produced in this last session, detailing what Knox called a dreamlike "vision." Both were typed by police—one at 1:45 a.m., the second at 5:45 a.m. She retracted the statements in a handwritten letter as soon as she was left alone ("In regards to this 'confession' that I made last night, I want to make it clear that I'm very doubtful of the verity of my statements because they were made under the pressures of stress, shock, and extreme exhaustion."). Notably, nothing in the confessions indicated that she had guilty knowledge. In fact, the statements attributed to Knox were factually incorrect on significant core details (e.g., she named as an accomplice a man whom police had suspected but who later proved to have an ironclad alibi; she failed to name another man, unknown to police at the time, whose DNA was later identified on the victim). Nevertheless, Knox, Sollecito, and the innocent man she implicated were all immediately arrested. In a media-filled room, the chief of police announced: *Caso chiuso* (case closed).

Police had failed to provide Knox with an attorney or record the interrogations, so the confessions attributed to her were ruled inadmissible in court. Still, the damage was done. The confession set into motion a hypothesis-confirming investigation, prosecution, and conviction. The man whose DNA was found on the victim, after specifically stating that Knox was not present, changed his story and implicated her while being prosecuted. Police forensic experts concluded that Knox's DNA on the handle of a knife found in her boyfriend's apartment also contained Kercher's blood on the blade and that the boyfriend's DNA was on the victim's bra clasp. Several eyewitnesses came forward. An elderly woman said she was awakened by a scream followed by the sound of two people running; a homeless drug addict said he saw Knox and Sollecito in the vicinity that night; a convicted drug dealer said he saw all three suspects together; a grocery store owner said he saw Knox the next morning looking for cleaning products; one witness said he saw Knox wielding a knife.

On December 5, 2009, an eight-person jury convicted Amanda Knox and Raffaele Sollecito of murder. The two were sentenced to 26 and 25 years in prison, respectively.

This article was published Online First April 30, 2012.

Correspondence concerning this article should be addressed to Saul M. Kassin, John Jay College of Criminal Justice, City University of New York, 445 West 59th Street, New York, NY 10019. E-mail: skassin@jjay.cuny.edu

Vol. 67, No. 6, 431–445 DOI: 10.1037/a0028212

Saul M. Kassin

Finally, on October 3, 2011, after having been granted a new trial, they were acquitted. Ten weeks later, the Italian appeals court released a strongly worded 143-page opinion in which it criticized the prosecution and concluded that there was no credible evidence, motive, or plausible theory of guilt. For the four years of their imprisonment, this story drew international attention (for comprehensive overviews of the case, see Dempsey, 2010, and Burleigh, 2011).[1]

It is now clear that the proverbial mountain of discredited evidence used to convict Amanda Knox and Raffaele Sollecito was nothing but a house of cards built upon a false confession. The question posed by this case, and so many others like it, is this: Why do confessions so often trump innocence?

The Psychology of Confessions

This article represents my third in this journal on the psychology of confession evidence. In the first article (Kassin, 1997), I overviewed an emerging study of confessions, described and critically evaluated the influential Reid technique of interrogation (Inbau, Reid, Buckley, & Jayne, 2013), and reiterated three classic types of false confessions previously identified (Kassin & Wrightsman, 1985). The purpose was to describe the phenomenon of false confessions and to note relevant psychological theories and research on the suspect characteristics and police interrogation techniques that can lead innocent people to confess.

Inspired by the founding of the Innocence Project (http://www.innocenceproject.org/; see Scheck, Neufeld, & Dwyer, 2000) and the first wave of DNA exonerations in the 1990s, a startling 25% of which contained false confessions in evidence, and further animated by recent debates over the use of torture or "enhanced" methods of interrogation (Greenberg, 2006), interest in this literature has exploded. This burst of activity can be seen in stories about actual cases (e.g., Burns, 2011; Firstman & Salpeter, 2008; Wells & Leo, 2008), a best-selling crime novel (Grisham, 2010), and publications of review articles, book chapters, and whole books focused on the emerging science of false confessions (Gudjonsson, 2003; Gudjonsson & Pearse, 2011; Kassin, 2008; Kassin & Gudjonsson, 2004, 2005; Lassiter, 2004; Lassiter & Meissner, 2010; Leo, 2008).

On the basis of individual and aggregated case studies (Drizin & Leo, 2004; Garrett, 2011; Warden & Drizin, 2009) and self-reports from civilians (Gudjonsson, Sigurdsson, & Sigfusdottir, 2009) as well as police (Kassin et al., 2007), it is now clear that false confessions are not a new or novel phenomenon and that they occur on a regular basis in all parts of the world and in criminal justice, military, and corporate settings. Research continues at a brisk pace—examining, for example, the practices of police interrogation (Leo, 2008); the extent to which *Miranda* rights comprehension and recall are compromised by language (Rogers, Hazelwood, Sewell, Harrison, & Shuman, 2008) as well as interrogation stress and other situational factors (Rogers, Gillard, Wooley, & Fiduccia, 2011; Scherr & Madon, 2011); the links between mental illness and false confession (Redlich, Kulish, & Steadman, 2011; Redlich, Summers, & Hoover, 2010); adolescence as a risk factor (Owen-Kostelnik, Reppucci, & Meyer, 2006); race differences in interrogation room behavior (Kennard & Kassin, 2009; Najdowski, 2011); "secondary confessions" alleged by informants about the suspect (Swanner, Beike, & Cole, 2010); perceptions of torture in the context of interrogation (Nordgren, McDonnell, & Loewenstein, 2011); similarities and differences between suspect and victim statements (Malloy & Lamb, 2010); basic psychological processes underlying a suspect's decision to confess (Davis & Leo, 2012; Madon, Guyll, Scherr, Greathouse, & Wells, 2012); the effects of guilt-presumptive confirmation biases on behavior in the interrogation room (Hill, Memon, & McGeorge, 2008; Kassin, Goldstein, & Savitsky, 2003; Narchet, Meissner, & Russano, 2011); the use of "investigative interviewing" as an alternative approach to questioning suspects (Williamson, 2006); and the development of new laboratory paradigms to devise more diagnostic police methods of deception detection (Vrij, Granhag, & Porter, 2010) and interrogation (Meissner, Russano, & Narchet, 2010). This literature was comprehensively summarized in an official White Paper of the American Psychology-Law Society (Division 41 of the American Psychological Association [APA]; Kassin et al., 2010).

In a second article (Kassin, 2005), I additionally proposed the paradoxical hypothesis that false confessions are facilitated not only by dispositional characteristics of weak and vulnerable suspects (i.e., youth, intellectual impair-

[1] Additional sources to which I had access include the translated police reports of Knox's statements; personal communications with Amanda Knox, Madison Paxton, and Nina Burleigh; and the Perugia Murder File translation of the Jury/Judge Conviction Report.

ment, mental illness, and personality traits that foster compliance and suggestibility) and situational aspects of custody and interrogation (i.e., lengthy sessions, threats, promises, presentations of false evidence, and minimization themes that imply leniency) but by the phenomenology of innocence. Innocence is a mental state that leads innocent people to waive their *Miranda* rights to silence and to counsel (Kassin & Norwick, 2004; Moore & Gagnier, 2008); to behave in ways that are open and forthcoming in their interactions with police (Hartwig, Granhag, & Strömwall, 2007); to offer up alibis freely, without regard for the fact that police may view minor inaccuracies with suspicion (Olson & Charman, 2011); and to exhibit less physiological arousal in response to the stress of interrogation (Guyll et al., 2012) and on critical items of a concealed information test even when preinformed about the crime (Elaad, 2011). Over the years, laboratory experiments have shown that the vast majority of participants who are accused of a transgression they did not commit—in stark contrast to those who are guilty—refuse to accept plea offers, often to their own detriment, indicating their confidence in acquittal (Gregory, Mowen, & Linder, 1978; Tor, Gazal-Ayal, & Garcia, 2010).

The story of Amanda Knox illustrates just how *innocence* can put *innocents* at risk. Immediately after Meredith Kercher was found murdered, her English roommates left Perugia; her Italian roommates obtained lawyers. Yet Knox, naïve to the risk and exhibiting no consciousness of guilt, wanted to stay to help police. Knox's mother described her daughter as "oblivious to the dark side of the world" (Rich, 2011). Even later, in court, on the day of her first verdict, Knox fully expected to be acquitted (Burleigh, 2011).

Theorizing that innocence is a state of mind that leads people to trust the criminal justice system during interrogation, Perillo and Kassin (2011) examined the relatively benign bluff technique by which interrogators pretend to have evidence without asserting outright that this evidence implicates the suspect (e.g., stating that witnesses were present to be interviewed or that biological evidence was collected and sent to a laboratory for testing). The theory underlying the bluff makes sense: Fearing the evidence to be processed, perpetrators will succumb to police pressure and confess; not fearing that alleged evidence, innocents would not succumb and confess. Yet in two experiments, Perillo and Kassin found that innocent participants were substantially more likely to confess to pressing a forbidden key, causing a computer to crash, when told that their keystrokes had been recorded for later review. In a third experiment, innocent participants were more likely to confess to willful cheating when told that a surveillance camera had taped their session. In both sets of studies, these participants noted that the bluff represented a promise of future exoneration despite confession, which paradoxically made it easier to confess.

The Consequences of Confession

In the present article, I shift the focus from the psychological causes of false confessions, as discussed in my previous articles, to their *consequences* for police investigations, prosecutions, jury trials, and appeals—and the implications that follow for law and the administration of justice. In a nutshell, I propose the empirically generated argument that the vital principle of corroboration is based on a misconception concerning proof of guilty knowledge and the independence of different types of evidence and cannot, therefore, be trusted to safeguard innocent confessors against wrongful conviction.

Once a suspect confesses, police often close the investigation, deem the case solved, and overlook exculpatory information—even if the confession is internally inconsistent, contradicted by external evidence, or the product of coercive interrogation (Drizin & Leo, 2004; Leo & Ofshe, 1998). This trust in confessions may extend to prosecutors as well, some of whom express skepticism about false confessions and stubbornly refuse to admit the possibility of their falsity even after DNA testing has unequivocally excluded the confessor (Findley & Scott, 2006). For example, Bruce Godschalk was exonerated of two rape convictions after 15 years in prison when DNA tests indicated that he was not the rapist. Yet the district attorney refused, at first, to accept the results. When questioned about it, this district attorney said, "I have no scientific basis. I know because I trust my detective and my tape-recorded confession. Therefore the results must be flawed until someone proves to me otherwise" (Rimer, 2002, p. A14). This is not an isolated incident. The Center on Wrongful Convictions (2010) reported on several known cases in which a confessor was tried and convicted despite having being excluded by DNA. Some instances are so flagrant that the *New York Times Magazine* recently published an article titled "The Prosecution's Case Against DNA" about prosecutors who generate improbable arguments to reconcile the DNA exclusion of suspects who have given prior confessions (Martin, 2011).

It is important to note that many tragic false confession stories contain two psychology-rich subplots: (a) why it happened, that is, the dispositional and situational factors that caused an innocent person to confess and (b) why judges, juries, and appeals courts all believed the false confession, making the effect difficult to reverse. It is also important to note that much of what is known about false confessions in the real world is based on a specialized subset of cases, often involving rape and murder, in which the confession both resulted in a wrongful conviction and was later identified as such. Mostly hidden from view are cases in which the confessor's innocence was discovered before conviction or not at all.

Perceptions of Confession Evidence

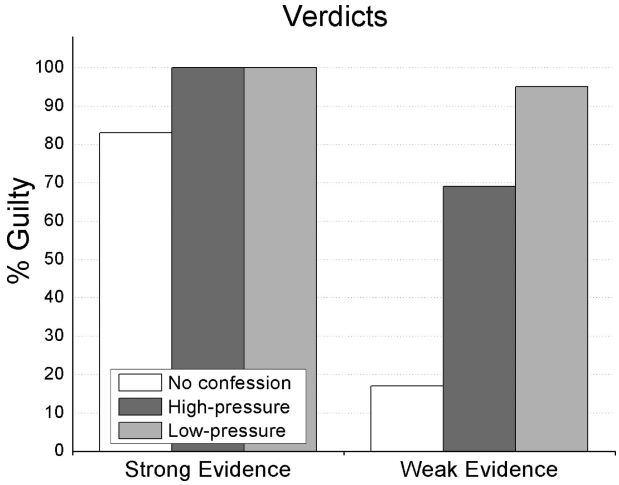
False confession is not a phenomenon that is known to the average layperson as a matter of common sense. Over the years, mock jury studies have shown that confessions have more impact on verdicts than do other potent forms of evidence (Kassin & Neumann, 1997) and that people do not adequately discount confessions—even when they are retracted and judged to be the result of coercion (Kassin & Sukel, 1997; Kassin & Wrightsman, 1980; Redlich, Ghetti,

& Quas, 2008), even when jurors are told that the confessor suffered from psychological illness or interrogation-induced stress (Henkel, 2008), and even when the confessions are provided not by the defendant himself or herself but by an informant who is incentivized to falsely implicate the defendant (Neuschatz, Lawson, Swanner, Meissner, & Neuschatz, 2008). Most people reasonably believe that they would never confess to a crime they did not commit, so they evaluate others accordingly, do not understand the influence of police interrogation practices, and have only a rudimentary understanding of the dispositional and situational factors that would lead someone innocent to confess (Blandón-Gitlin, Sperry, & Leo, 2011; Henkel, Coffman, & Dailey, 2008; Leo & Liu, 2009).

The noncritical acceptance of confessions afflicts judges as well as lay juries. In one study, Wallace and Kassin (2012) presented 132 judges from three states with a murder case summary in which there was strong or weak evidence against the defendant. In a high-pressure confession condition, the defendant was questioned for 15 hours, during which time his interrogators screamed, threatened him with the death penalty, waved a gun, and refused to accept his claims of innocence. In a low-pressure confession condition, the defendant was questioned for only 30 minutes before producing a confession; although he claimed that he was coerced, he described nothing specific and the claim was not borne out by a videotape of the session. In the no-confession condition, participants were told only that the defendant was questioned by police, during which time he denied any involvement.

Reasonably, judges were less likely to see the confession as voluntary, and hence as properly admitted into evidence, when it resulted from a high-pressure than a low-pressure interrogation (29% vs. 84%, respectively). Paralleling past research on mock juries, however, even the high-pressure confession significantly increased guilty verdicts. Figure 1 shows that conviction rates were uniformly high across cells in the strong evidence condition. But in the weak evidence condition, which produced a mere 17% conviction rate absent a confession, a significant increase in conviction rate was produced not only by the low-pressure confession (96%) but by the high-pressure confession as well (69%). As with lay juries, judges see confession as such powerful evidence that they do not discount it when it is legally and logically appropriate to do so.

Figure 1
Percentage of Judges Who Voted Guilty When the Case Was Strong or Weak and When It Contained a High- or Low-Pressure Confession—or None at All



Note. Adapted from "Harmless Error Analysis: How Do Judges Respond to Confession Errors?" by D. B. Wallace & S. M. Kassin, 2012, *Law and Human Behavior, 36,* p. 155. Copyright 2011 by the American Psychological Association.

The Common Sense of Confessions

The tendency to believe confessions begins with the fact that people reflexively accept what is presented to them. In an article titled "How Mental Systems Believe," Gilbert (1991) distinguished between two Western philosophical views on the acquisition of beliefs. René Descartes (1644/1984) opined that people are neutral in their reactions to new assertions—first acquiring and comprehending an idea and then accepting it or not if justified, say, by logic or extrinsic evidence. In contrast, Benedict Spinoza (1677/1982) argued that people automatically and inevitably accept as true every assertion they hear—and must then, later, correct for that belief if it proves not to be credible. Credulity, acceptance, and belief thus precede skepticism, doubt, and disbelief. Describing this latter view, William James (1890) noted, "All propositions, whether attributive or existential, are believed through the very fact of being conceived" (p. 290).

The myth that legal decision makers can be trusted to disbelieve false confessions and serve as a safety net for innocent confessors is debunked by a number of basic tendencies and shortcomings of social perception. To begin with, there is empirical support for Gilbert's (1991) argument "that people are Spinozan systems that, when faced with shortages of time, energy, or conclusive evidence, may fail to unaccept the ideas that they involuntarily accept during comprehension" (p. 115). In one study, for example, research participants read a crime report that contained information indicating that the crime was high or low in severity. That information was explicitly tagged as false upon presentation, yet it led participants to administer harsher or more lenient sentences, respectively, to the defendant (Gilbert, Tafarodi, & Malone, 1993). In a second study, participants who read a short story they knew to be fictional—like a novel, movie, or television show—later incorporated elements of that story into their beliefs about the real world (Gerrig & Prentice, 1991).

This tendency for people to accept what they see and hear at face value manifests itself in two confession-relevant literatures. In one area, researchers have consistently observed that people are notoriously gullible, exhibiting a truth bias that contributes to poor performance at detecting deception (Bond & DePaulo, 2006; Levine, Park, &

McCornack, 1999). It appears that neither laypeople nor professionals distinguish truths from lies at high levels of accuracy, even in high-stakes forensic domains (Hartwig & Bond, 2011; Vrij, 2008; Vrij et al., 2010). This problem can be seen in people's inability to identify false confessions. Kassin, Meissner, and Norwick (2005) videotaped male prison inmates as they gave true confessions for their crimes and concocted false confessions to crimes they did not commit. Neither college students nor police investigators were adept at distinguishing true from false confessions. This finding was later replicated for judgments of juvenile suspects (Honts, Kassin, & Forrest, 2009).

The tendency to accept self-report and other behavior at face value is also evident in the domain of attribution. Over a wide range of contexts, research has shown that social perceivers routinely commit the *fundamental attribution error*, or *correspondence bias*—that is, they tend to make dispositional attributions for other people's actions while underestimating the role of situational factors (Gilbert & Malone, 1995; Jones, 1990; Ross, 1977). Hence, although people recognize the coerciveness of certain interrogation tactics, they do not perceive an accompanying risk of false confessions or the dispositional and situational factors that would increase it (Henkel et al., 2008; Leo & Liu, 2009).

The common sense of confession is particularly problematic for the innocent confessor. In addition to the Spinozan tendency for people to believe what they see and hear from others, people have a strong tendency in attribution—as noted by Heider (1958), Jones and Davis (1965), and other attribution theorists—to especially trust statements against self-interest. This principle of intuitive attributional logic is embedded in the Federal Rules of Evidence (FRE) that prohibit hearsay, a secondhand statement that a witness heard about from someone else and did not perceive directly. In general, hearsay is inadmissible because it cannot be trusted. There are, however, notable exceptions to the hearsay rule. FRE 804-b-3 states that "declarations against interest" (i.e., statements that would expose a declarant to criminal or civil liability) are admissible as an exception to the hearsay rule on the assumption that such statements in particular can be trusted. Illustrating use of this principle of self-interest, research shows that people are far more likely to believe a suspect's admissions of guilt than his or her denials (Levine, Kim, & Blair, 2010).

APA's Amicus Curiae Briefs on Confessions

The impulse to trust confessions, almost regardless of the circumstances under which they are taken or regardless of exculpatory evidence, can be seen in the way U.S. courts often react to defendants convicted by confession. This point is illustrated by two cases in which APA submitted amicus curiae briefs on behalf of convicted confessors—first, on the question of whether they should be eligible, as others are, for DNA testing to establish factual innocence; and second, on the question of whether, if exonerated, they are eligible, along with others who are wrongfully convicted, to receive compensation from the state.

Wright v. Commonwealth of Pennsylvania (see http://www.apa.org/about/offices/ogc/amicus/wright.aspx) concerned the case of Anthony Wright, who was convicted in 1993 of rape and murder on the basis of a confession ruled voluntary and admitted at his trial. Along with many other states, Pennsylvania recently passed a law to ensure a prisoner's right to postconviction DNA testing to establish factual innocence. Wright was denied that right, however, because state courts ruled that if a defendant had confessed, then he or she was later barred from asserting innocence in a request for DNA testing. On November 13, 2008, APA submitted an amicus curiae brief stating that innocent people can be induced to confess through processes of interrogation and that Wright's confession, even if voluntary by law, should not bar his consideration for postconviction DNA testing. In February of 2011, the Supreme Court of Pennsylvania agreed and overruled the lower courts.

In *Warney v. State of New York* (http://www.apa.org/about/offices/ogc/amicus/warney.aspx), Doug Warney—a man with mental retardation and AIDS-related dementia—had been convicted of murder on the basis of a richly detailed false confession produced after hours of interrogation. After serving a nine-year prison term, he was exonerated by DNA testing. When Warney sought reparations, as provided by the state's compensation statute, however, the court ruled that he was ineligible because his conviction resulted from his "own conduct"—which is to say, the false confession. On July 9, 2010, APA filed an amicus brief supporting Warney's petition that false confession should not bar a wrongfully convicted person from recovery under the Unjust Conviction Act. In March 2011, the New York State Court of Appeals unanimously decided in Warney's favor.

In still other briefs, APA has weighed in to note that judges and juries have difficulty assessing confession evidence, that the phenomenon of false confession is counterintuitive, and that psychological experts should be permitted to testify at trial because their testimony would draw from generally accepted research and would assist the trier of fact (*Rivera v. Illinois*, July 12, 2010, http://www.apa.org/about/offices/ogc/amicus/rivera.aspx; *Michigan v. Kowalski*, September 1, 2011, http://www.apa.org/about/offices/ogc/amicus/kowalski.aspx).

Confessions as Hollywood Productions

Analyses of actual cases suggest that police-induced false confessions pose a particular challenge to judges and juries because they often contain not only an admission of guilt but a full narrative replete with content cues commonly associated with truth telling and guilty knowledge. In an examination of 38 false confessions derived from the Innocence Project's DNA exoneration case files, Garrett (2010) found that 36 contained accurate crime details. In fact, most contained nonpublic information that became a centerpiece of their prosecution—information, according to detectives who testified, that only the perpetrator could have known. As these confessors were factually innocent and had no firsthand basis for guilty knowledge, it appears that police had communicated these details, inadvertently

or purposefully—through leading questions and assertions, exposure to photographs, or escorted visits to the crime scene.

To further complicate matters, many false confessions contain vivid details of what the suspect allegedly did, how, why, and with what effects. In a content analysis of 20 false confessions, Appleby, Hasel, and Kassin (2011) found that all the statements contained visual and auditory details about the crime and how it was committed, about the time and location, and about the victim—his or her physical appearance and behavior before, during, and after the crime. Overall, most statements referenced co-perpetrators, witnesses, and other actors; most described a motive (e.g., jealousy, revenge) and a minimization theme that justified, excused, mitigated, or externalized blame (e.g., claiming the crime was spontaneous or accidental; blaming alcohol, peer pressure, or provocation). Still others contained explicit assertions that the confession was voluntary, "illustrators" (e.g., a hand drawn map or a physical reenactment), deliberately inserted errors that were corrected by the confessor, expressions of remorse, and outright apologies. These results appear in Table 1. Not surprisingly, a follow-up mock jury study showed that elaborate narrative confessions in which the defendant recounted how and why he or she committed the crime increased confidence in guilty verdicts.

The case of DNA-exonerated confessor Barry Laughman illustrates the richness of these narratives. Laughman's false confession contained facts about the crime that were verifiable, strikingly accurate, and not in the public domain. Despite Laughman's innocence, his statement revealed where the victim was found and in what position, that a window was open, that she was vaginally raped, that she had suffocated on pills, that she was hit in the head and grabbed by the wrists, and that a handful of cigarette butts had been strewn throughout the house. His confession also contained descriptions of a coverup, statements of motivation for both the rape and the murder, and gratuitous expressions of shame—all of which served to mislead a judge and jury (Garrett, 2010; http://www.innocenceproject.org).

Reflecting the layperson's bias toward making dispositional attributions for behavior, numerous wrongful conviction cases suggest that narrative confessions can be so powerful as to overwhelm contradictory forensic evidence. In the case of Amanda Knox described earlier, the prosecutor theorized in the wake of her coerced confession that Knox was motivated by money or personal envy of her British roommate. Two weeks later, the rapist whose DNA was found in sperm and other biological matter at the crime scene was apprehended. Yet rather than reconsider Knox's confession in light of this contradictory evidence, the prosecutor spun a new and wholly unsupported theory of the crime: that the rapist, Knox, and her boyfriend had come together and killed the victim as part of a satanic sex game (at trial, he redacted the satanic part but still referred to Knox as a "she-devil").

In matters of proof, one would expect that people in general would trust science over self-report. In the courtroom, however, confessions often trump exculpatory DNA evidence. In the infamous Central Park jogger case, for example, five boys were convicted of rape on the basis of their confessions even though all were excluded by the DNA found on the victim. At trial, the prosecutor argued that the results proved only that the defendants failed to ejaculate and that an unknown sixth accomplice was present (Burns, 2011). In a series of studies, Appleby and Kassin (2011) tested the counterintuitive hypothesis that confession trumps DNA. They found that people confronted with a confession and exculpatory DNA evidence seldom voted for conviction, even when the confession conveyed details about the crime. But when the prosecutor offered an explanatory theory to reconcile the contradiction (e.g., the defendant failed to ejaculate and the semen reflected a prior consensual sex act or an unnamed accomplice), the conviction rate increased significantly—from 10% to 33% in a study of college students, and from 14% to 45% in a study of community adults.

Table 1
Content Analysis of 20 False Confessions: Percentages Containing Various Details

Contents	Frequency
Time and place	
Time of day	100%
Location and space	100%
Visual crime detail	100%
Illustrators	45%
The victim	
Victim's behavior	100%
Victim's words and utterances	80%
Victim's physical appearance	75%
Victim's mental state	45%
Self-reflections	
Cognitive/affective inner states	85%
Motives for the crime	80%
Minimization themes	60%
Statement of voluntariness	50%
Expressions of remorse	40%
Explicit apologies	25%

Note. Adapted from "Police-Induced Confessions: An Empirical Analysis of Their Content and Impact' by S. C. Appleby, L. E. Hasel, and S. M. Kassin, 2011, *Psychology, Crime & Law*. Advance online publication, pp. 5–6. Copyright 2011 by Taylor & Francis.

Confessions Corrupt Other Evidence

Precisely because confession evidence is highly trusted as a matter of logic and common sense, it often provides a sufficient basis for jury convictions. However, basic research in social cognition suggests the possibility of a second, more troubling mechanism by which confessions exert influence: by tainting the perceptions of eyewitnesses, forensic experts, and others entrusted to provide independent other evidence to a judge and jury.

Over the years, psychological research across a range of domains has revealed that top-down influences inform human judgment. Classic studies showed that prior exposure to images of a face or a body, an animal or a human, or letters or numbers can bias what people perceive in an ambiguous figure (Bruner & Minturn, 1955; Bugelski & Alampay, 1961; Fisher, 1968; Leeper, 1935). Similarly, people detect more resemblance between an adult and a child when led to believe that the two are parent and offspring (Bressan & Dal Martello, 2002); they perceive more similarity between a suspect and a facial composite when led to believe the suspect is guilty (Charman, Gregory, & Carlucci, 2009); and they hear more incrimination in degraded recordings of speech when led to believe that the interviewee was a criminal suspect (Lange, Thomas, Dana, & Dawes, 2011).

The literature on the primacy of first impressions further illustrates that prior beliefs can bias the interpretation of evidence (Asch, 1946). Depending on one's first impression of a person, the word *proud* can mean self-respecting or conceited; *critical* can mean astute or picky; and *impulsive* can mean spontaneous or reckless (Hamilton & Zanna, 1974; Watkins & Peynircioglu, 1984). Because of the operation of ubiquitous confirmation biases, the presence of objective evidence may even exacerbate the effects of preexisting beliefs (Nickerson, 1998). When subjects were asked to evaluate the academic potential of a schoolgirl from a high or low socioeconomic status background, those who observed her taking a test in which she answered some questions correctly but not others exhibited a greater stereotype effect than those who did not see her test-taking performance. Rather than extinguish the effect of the stereotype, the objective behavioral evidence reinforced it (Darley & Gross, 1983).

Recent research suggests that confessions may trigger the same types of confirmation processes in the high-stakes venue of the criminal justice system. In one study, Elaad, Ginton, and Ben-Shakhar (1994) asked polygraph examiners from the Israeli Police Force to evaluate and interpret charts previously deemed inconclusive. Some examiners, but not others, were told that the suspect had ultimately confessed. Results showed that those in the confession condition rated the charts as significantly more deceptive than those in the control condition (similar results were not obtained on charts that were conclusive). In a second study, Dror and Charlton (2006) presented five latent fingerprint experts with pairs of prints from a crime scene and a suspect in an actual case in which they had previously made a match or exclusion judgment. The prints were accompanied either by no extraneous information; by an instruction that the suspect had confessed, suggesting a match; or by an instruction that the suspect was in custody at the time, suggesting exclusion. The misinformation in the two biasing conditions produced an overall change in 17% of the original, previously correct judgments—a finding that may well extend to visual similarity judgments in other forensic science domains such as ballistics; hair and fiber analysis; bite marks; impression evidence involving shoeprints, bite marks, tire tracks, and handwriting; and bloodstain pattern analysis (Dror & Cole, 2010). Even the interpretation of complex DNA mixtures may require judgment that is subject to bias (Dror & Hampikian, 2011).

Confessions may also influence the testimony of lay witnesses. Hasel and Kassin (2009) staged a theft and asked for photographic identification decisions from witnesses using a lineup that did not contain the culprit. Two days later, individual witnesses were told that the person they had identified denied guilt during a subsequent interrogation, or that he confessed, or that a specific other lineup member confessed. Among those who had made a selection but were told that another lineup member confessed, 61% changed their identifications—and did so with confidence. Among those who had correctly not made an initial identification, 50% went on to select the confessor.

The criminal justice system presumes the independence of different types of evidence. But does this presumption characterize the realities of criminal investigation? The basic and forensic psychology research described above suggests the possibility that confessions have the power to corrupt other evidence, further enhancing its impact on judges and juries. To determine if this phenomenon, amply demonstrated in the laboratory, also occurs in actual cases, Kassin, Bogart, and Kerner (2012) conducted an archival analysis of DNA exonerations from the Innocence Project case files. To test the "corruptive confessions" hypothesis, they compared the number and kind of errors made in wrongful conviction cases containing a false confession with those in which there was no confession. This analysis indicated that additional errors were present in 78% of false confession cases; that false confessions were often accompanied, in order of frequency, by invalid or improper forensic science (63%), by mistaken eyewitness identifications (29%) and by untruthful snitches or informants (19%); and that in 65% of confession cases that contained multiple errors, the confession was obtained first rather than later in the investigation. Of particular interest to psychologists is that the most common problem in DNA-based wrongful convictions is the eyewitness identification error, which was present in 75% of cases in the Innocence Project sample. Using this latter subsample as a point of comparison, Kassin et al. (2012) compared eyewitness and confession cases and found that the latter contained more additional errors overall, more forensic science errors, and more informant errors (see Table 2).

It is interesting that the most common means of corroboration for false confessions comes from bad forensic science, which was present in nearly two thirds of these cases. As a result of improprieties in crime laboratories across the country and the alarming frequency with which errors have surfaced in wrongful convictions, the National Academy of Sciences (2009) recently published a highly critical assessment of a broad range of forensic disciplines such as those involving ballistics, hair and fiber analysis, impression evidence, handwriting, and even fingerprints. The National Academy of Sciences concluded that there are problems with standardization, reliability, accuracy, and error and that there is the potential for contextual bias. In an article on "The Genetics of Innocence," Hampikian,

Table 2
Percentages of "Other Evidence" Errors in DNA Exoneration Cases That Contained Either a False Confession or a Mistaken Eyewitness

Case error	Forensic-science error	Informant error	No other errors
False confessions (N = 42)	67	24	31
Mistaken eyewitnesses (N = 163)	45	6	52

Note. Within each column, the percentages are significantly different at $p < .05$. Adapted from "Confessions That Corrupt: Evidence From the DNA Exoneration Case Files" by S. M. Kassin, D. Bogart, and J. Kerner, 2012, *Psychological Science, 23,* p. 43. Copyright 2012 by Association for Psychological Science.

West, and Akselrod (2011) found that invalid forensic science testimony was found in DNA exonerations in areas as highly regarded as serology (38%), hair comparison (22%), and even fingerprint comparison (2%). Clearly, the presence of a confession and the perception of guilt thus formed constitute the kind of strong contextual bias that can skew expert judgments in these domains.

One out of five false confession cases contained testimony from a jailhouse snitch or other type of incentivized informant claiming to have overheard the defendant confess. Snitching is a commonplace, clandestine, and insufficiently regulated "dirty little secret" in the criminal justice system (Natapoff, 2009). In the first documented wrongful conviction case in U.S. history, in 1819, two brothers in Vermont were convicted and sentenced to death for murder when a cellmate testified that one of the brothers had confessed to him. For his testimony, the snitch was freed—as were the brothers when the alleged victim turned up alive in New Jersey (Warden, 2004). Recent research confirms the fear arising from this practice: In a series of studies, incentives increased the rate at which participants falsely alleged that their lab partner had confessed to causing the experimenter's computer to crash (Swanner & Beike, 2010; Swanner et al., 2010).

The bias set into motion by confession is not a mere laboratory phenomenon—and it can have grave consequences. In the Barry Laughman case described earlier, the defendant confessed to rape and murder during an unrecorded interrogation. The next day, serology tests showed that Laughman had Type B blood; yet the semen recovered from the victim was from a Type A secretor. Aware that Laughman had confessed, the state forensic chemist went on to propose four "novel" theories, none grounded in science, to explain away the mismatch. On the basis of his confession, Laughman was wrongfully convicted and imprisoned for 16 years (see http://www.innocenceproject.org/Content/Barry_Laughman). Another egregious instance occurred in the 2004 trial of Tyler Edmonds in Mississippi. In that case, 13-year-old Edmonds confessed that he had physically assisted his older sister in shooting her husband. Supporting what had become a hotly contested confession, the state pathologist who conducted the autopsy on the victim testified that the gunshot wound suggested a bullet fired by two persons pulling the trigger simultaneously. Highly critical of this expert's unfounded opinion, the Mississippi Supreme Court overturned the conviction (*Tyler Edmonds v. State of Mississippi,* 2007). Edmonds was then retried and acquitted.

The studies described thus far have shown that confessions can spawn other incriminating evidence, creating an illusion of corroboration. It is important to note, however, that this effect may underestimate the problem in two ways. First, just as confessions can taint other evidence, other evidence can taint confessions as well. Indeed, there are numerous studies as well as anecdotal support for the proposition that innocent people can be induced to confess by the true or false presentation of an eyewitness, a failed polygraph, or other incriminating evidence (Gudjonsson & Pearse, 2011; Kassin et al., 2010; Kassin & Kiechel, 1996). Second, there may be instances where false confessions also serve to suppress exculpatory evidence. At present, only anecdotal data are available on this point. In the Laughman case, two witnesses approached police to insist that they had seen the victim alive after the confessed murder was alleged to have occurred. Yet police sent the witnesses home, telling them "You must have seen a ghost." In a second case, DNA exoneree John Kogut named several alibi witnesses he was with on the night of the murder of which he was accused. Research shows that it is not easy for people to generate and validate accurate alibis for a specific time and place (Olson & Charman, 2011). Yet Kogut managed to do so. Initially, his alibi witnesses confirmed his whereabouts. They later withdrew their support, however, once informed that he had confessed. Systematic research is needed to test this phenomenon and the conditions under which exculpatory evidence is suppressed by confession.

Do False Confessions Corrupt the Truth-Seeking Process?

In addition to corrupting the evidence upon which fact finders render judgments of guilt, confessions may also adversely affect the truth-seeking *process* by which justice is administered. Confession evidence is highly and uniquely polarizing when it reaches the courts. On the one hand, confessions have long been considered a gold standard in evidence. In the words of one legal scholar, "The introduction of a confession makes the other aspects of a trial in court superfluous" (McCormick, 1972, p. 316). On the other hand, the confessions presented at trial are those that defendants have invariably

retracted, typically accompanied by the contentious claim that they were coerced by police.

In light of the power of confessions, one wonders if defense lawyers in such cases feel pessimistic if not outright helpless, encouraging their clients to plead guilty and allocating relatively few of their precious resources to discovery and trial. "War stories" from proven false confession cases provide an anecdotal basis for this hypothesis. In addition, one wonders if defense claims of police coercion and contamination, which challenge the centerpiece of the government's case, lead some prosecutors to redouble efforts to procure other forms of incriminating proof, even if questionable in credibility. Once again, a number of proven false confession cases provide an anecdotal basis for this possibility.

To test the hypothesis that confessions corrupt the truth-seeking process, Kassin and Kukucka (2012) conducted an archival analysis of the first 273 DNA exoneration cases from the Innocence Project files, the total number as of September 2011. They compared false confession cases with cases in which there was no confession on instances of "bad lawyering" and "government misconduct" as classified on a per-case basis by the Innocence Project. Consistent with predictions, false confession cases were more likely to involve bad defense lawyering than were nonconfession cases (9.09% vs. 3.38%) and somewhat more likely to involve government misconduct (21.21% vs. 15.46%). Combined, these differences suggest that confession cases skew the adversarial process in ways that are detrimental to the defense. These archival findings are preliminary, not conclusive, and the associations found do not uncover the causal nexus between confessions to police and the subsequent behavior of counsel. The implications, however, are sobering. At this point, more research is needed to retest the hypothesis using surveys, interviews, and experimental methodologies.

Perhaps an even more dramatic effect on process concerns the possibility that false confessions undermine a defendant's opportunity to get his or her proverbial day in court. Using the Innocence Project database, Redlich (2010) found that exonerees who had falsely confessed were four times more likely to plead guilty than were those in the same population who had not confessed. Although based on a small number of guilty pleas, this pattern has continued. Out of 289 DNA exonerations posted by the Innocence Project (E. West, personal communication, March 30, 2012), false confession cases were significantly more likely to be resolved by a guilty plea (25.97%) than were nonconfession cases (3.78%). This difference suggests that many innocents who confess ultimately surrender rather than assert a defense. This is no small matter. Pleading guilty preempts the safeguards inherent in a trial by jury—a process in which a defendant is presumed innocent, the burden is on the state to prove guilt beyond a reasonable doubt, and accusing witnesses can be cross-examined. Pleading guilty also makes it more difficult later for a defendant to gain postconviction scrutiny and assert factual innocence.

Implications for Law, Justice, and Wrongful Convictions

The literature on wrongful convictions, buttressed by research on the dispositional and situational causes of false confessions, has inspired various calls for systemic reform. In particular, research has compelled the conclusion that the video recording of entire interrogations is a necessary safeguard—indeed, it is the primary recommendation in the recent American Psychology-Law Society White Paper (Kassin et al., 2010). Other recommendations have focused on the protection of vulnerable suspect populations (e.g., a requirement that minors be accompanied by a professional advocate, preferably an attorney) and the reform of certain police interrogation practices (e.g., a ban on the false evidence ploy and limits on the use of minimization themes that communicate leniency). As noted earlier, APA has recently weighed in on other matters pertaining to expert testimony, DNA testing, and eligibility for compensation.

Pretrial Corroboration Requirements

The research described in this article has far-reaching implications for criminal law and the safety nets designed to prevent miscarriages of justice. In particular, corroboration requirements are deeply rooted. In a pretrial rule founded in common law in England, many states require that confessions be corroborated as a precondition for admissibility. The rule was designed to prevent false confessions, to incentivize police to continue to investigate a case after obtaining a confession, and to safeguard against the tendency of juries to view confessions as dispositive of guilt regardless of the circumstances under which they were obtained (Ayling, 1984).

According to John Reid and Associates, a Chicago-based firm that has trained over half a million professional interrogators over the past 65 years, there are three means of corroborating a confession (Inbau et al., 2013). The weakest is *rational corroboration,* in which the suspect recounts "a detailed description of how the crime was committed, why it was committed, and perhaps how the suspect felt after committing the crime" (Inbau et al., 2013, pp. 356–357). The second means of support is *dependent corroboration,* which comes from proof of a suspect's guilty knowledge and ability to produce facts that were purposely withheld from all suspects and the media. The third and strongest is *independent corroboration*, which comes from extrinsic evidence consistent with (e.g., an eyewitness) or, better yet, generated by the confession (e.g., the location of the murder weapon or stolen property).

In principle, a corroboration requirement designed to ensure that only trustworthy confessions are used at trial represents an important potential safeguard. But the research cited in this article casts serious doubt as to the diagnosticity of these measures. It now appears that most police-induced false confessions within the database of DNA exonerations contain details about the crime that were allegedly withheld from suspects, thereby suggesting that the confessor had guilty knowledge and providing false dependent corroboration (Garrett, 2010). In these instances,

it is now clear that information was purposefully or unwittingly communicated to innocent suspects through the process of interrogation (Inbau et al., 2013). Most false confessions also contain elements of rational corroboration in the form of crime details on how, why, and with what effect the crime was committed, often including apologies and expressions of remorse (Appleby et al., 2011). Studies also now show that confessions, once taken, can corrupt lay witnesses and forensic experts, thus fostering an illusion of independent corroboration as well (Kassin et al., 2012).

The "Harmless Error" Analysis

Corroboration is also vitally important at the appellate level. In *Arizona v. Fulminante* (1991), the U.S. Supreme Court ruled that an erroneously admitted confession does not, as in the past, automatically entitle a convicted defendant to a new trial. Invoking the principle of "harmless error," the Court ruled that appeals courts reviewing disputed confession cases must determine, first, if a trial error occurred and, second, if that error was prejudicial or harmless (for a history of the harmless error rule, see Bilaisis, 1983). Operationally, the Court stated that even if a confession was coercive and its admission at trial erroneous, the conviction could be maintained if other evidence was so compelling that the jury would still have found the defendant guilty beyond a reasonable doubt.

Over the years, several legal scholars have criticized *Fulminante* on constitutional grounds, out of fear that it will encourage coercive methods of police interrogation, and on the argument that appeals court judges are ill-equipped by intuition, due in part to hindsight biases, to objectively estimate the strength of a prosecutor's case and the cumulative or "harmless" nature of the confession in dispute. Skepticism is justified on the question of whether appeals court judges can perform this retrospective analysis to determine how a jury would have voted in the absence of the known confession. In a study described earlier, Wallace and Kassin (2012) presented judges with a case summary in which the state's evidence was strong or weak and that was accompanied by a high- or low-pressure confession or none at all. The judges, like mock juries, voted to convict the confessor even in the high-pressure condition they deemed coercive. On the harmless error question, however, these same judges reacted in the prescribed manner: They determined both that the admission at trial of the high-pressure confession was erroneous and that the error was prejudicial in its effect on the jury when the totality of other evidence did not form a sufficient basis for conviction.

It appears that judges appreciate how juries are impacted by confessions. However, a serious problem still lurks: The harmless error doctrine—that an erroneously admitted confession can prove harmless when other evidence is sufficient to support a jury's conviction—rests on the core assumption that the alleged other evidence is independent of that confession. Indeed, according to Garrett (2010), appellate courts that conducted postconviction reviews of several confessors who were later exonerated had affirmed the convictions by citing the "overwhelming nature of the evidence against them and describing in detail the nonpublic and 'fully corroborative' facts they each reportedly volunteered" (p. 1107).

In light of studies showing that confessions can taint the judgments of polygraph examiners, latent fingerprint experts, eyewitnesses, and others, and the archival analysis of DNA exonerations indicating that many proven false confessions are accompanied by other subsequently collected evidentiary errors, doubt has been cast over that assumption of independence. It now appears that although a confession can be "subtracted" from the trial record, its influence persists. The courts must therefore consider the proposition that confessions they perceive to have been coerced and erroneously admitted corrupt the very evidence later used to make the confessions appear cumulative and hence harmless. The result: a perception of corroboration that is more illusory than real.

The Supreme Court's *Fulminante* opinion may be flawed on a second front. In reversing the past practice of automatically reversing convictions in which a coerced confession was admitted at trial, the Court asserted that confessions should not be treated differently from other evidence—that such errors do not constitute a "structural defect" in a defendant's ability to get a fair trial (e.g., akin to a lack of competent counsel, government misconduct, or an impartial judge). Although more data are needed to address this claim, recent analyses suggest that such defects are more likely to be found in wrongful convictions in which false confessions were in evidence than in nonconfession cases.

Corroboration Inflation

Taken together, research suggests that judges, juries, and others are doomed to believe the false confessions of innocent people not only because the phenomenon strongly violates common sense but because of *corroboration inflation*—a tendency for confessions to produce an illusion of support from other evidence. This appearance of support can come from the details of the confession statement itself in the form of dependent and rational corroboration, offering "proof" of the confessor's guilty knowledge—and it can also come from extrinsic evidence from lay and expert witnesses whose judgments were tainted by the confession. In both cases, the net effect is to weaken the safeguards designed to protect the accused confessor at the pretrial, trial, and appellate levels.

There are three important points to note about corroboration inflation and its potential to increase the risk of wrongful conviction. First, there is more than one mechanism by which a confession may influence other evidence. One possibility is that subsequent judgments are inadvertently tainted by mere knowledge of the confession and the cognitive confirmation biases resulting from that knowledge (for a review of research on confirmation biases, see Nickerson, 1998; for a review of "tunnel vision" in criminal justice, see Findley & Scott, 2006). A second possibility is that knowledge of the confession and the presumption of guilt it creates increase the *motivation* of lay witnesses and experts to help police and prosecutors implicate the suspect. Just as people tend to see what they

expect to see, recent studies indicate that people also see what they *want* to see (Ask & Granhag, 2007; Balcetis & Dunning, 2006). A third possibility is that the confession effect occurs because of biases by police seeking to procure support for their previously taken and recanted confession. This process is suggested by research showing that nonblind mock investigators often lead witnesses, albeit inadvertently, to falsely identify their suspect (Greathouse & Kovera, 2009). All these mechanisms are plausible. Without making subjective judgments about the mental states of police and witnesses, however, it may not be possible to tease apart these various sources of the effect in actual cases.

A second point about corroboration inflation is that confession is not the only form of evidence persuasive enough to produce false support in these ways. Beginning with the first wave of DNA exonerations, it has been clear that eyewitness mistakes constitute the most common problem in wrongful convictions (Brewer & Wells, 2011; Wells, Memon, & Penrod, 2006;Wells et al., 1998). In fact, many wrongful convictions contain two or more mistaken eyewitnesses who express high levels of certainty in their identifications. These multiple errors can occur independently when the suspect physically resembles the perpetrator—as in the mistaken identification of Ronald Cotton by Jennifer Thompson (Thompson-Cannino, Cotton, & Torneo, 2009). In some instances, however, eyewitnesses may influence one another, as demonstrated in numerous studies (Gabbert, Memon, & Allan, 2003; Shaw, Garven, & Wood, 1997; Skagerberg, 2007). To further complicate matters, eyewitnesses who have been tainted by extrinsic information cannot accurately estimate the extent of the influence, which suggests that self-report cannot be used to diagnose the corruption once it has occurred (Charman & Wells, 2008).

Third, it is important to realize that not all evidence is equally malleable or subject to corroboration inflation. Paralleling classic research indicating that expectations can color judgments of people, objects, and other stimuli that are ambiguous as opposed to those that compel a particular perception, forensic research indicates that ambiguity is a moderating condition. Asked to report on an event or make an identification decision on the basis of a memory trace that cannot be recovered, eyewitnesses are particularly malleable when confronted with evidence of a confession (Hasel & Kassin, 2009). This phenomenon was illustrated in the case against Amanda Knox. When police first interviewed Knox's British roommates, not one reported that there was bad blood between Knox and the victim. After Knox's highly publicized confession, however, the girls brought forth new "memories," telling police that Kercher was uncomfortable with Knox and the boys she would bring home (Burleigh, 2011).

Prior expectations can also bias interpretations of sensory stimuli such as auditory speech—but only when the recordings are degraded in quality and the stimuli are phonologically ambiguous, such as the words *gum* and *gun* (Lange et al., 2011). The same is true of the judgments of polygraph examiners—again, when the physiological test data are ambiguous, not when they contain physiological arousal patterns strongly indicative of truth or deception (Elaad et al., 1994). Within the forensic domains critiqued by the National Academy of Sciences (2009), the potential for confession-induced corroboration inflation is real, more so than previously imagined. In an article titled "The Vision in 'Blind' Justice," Dror and Cole (2010) noted that many forensic judgments involve matching a visual pattern left at a crime scene with a sample taken from a suspect (e.g., shoe prints, tool marks, bite marks, tire marks, handwriting, ballistics). The prototype is fingerprint identification, a forensic "science" long considered nearly perfect (Cole, 2001). Yet no two fingerprint impressions are identical even if lifted from the same source and finger because of variations in skin elasticity, the amount of pressure applied, the material on which the print was left, and other variables. And in real life, most fingerprints are partial and distorted, called *latent* prints. Dror and Charlton (2006) thus found that evidence of a confession led fingerprint experts to alter some judgments they had previously made. As illustrated in the two-trial ordeal of Amanda Knox and Rafaelle Solecito—where court-appointed DNA experts at her second trial flatly contradicted the original results (Povoledo, 2011)—even DNA testing, considered the best of the forensic sciences, is subject to judgment bias when samples are too small or when complex mixtures are analyzed (Dror & Hampikian, 2011).

Conclusion

There are two problems with false confessions. The first is that certain suspect characteristics and police practices can conspire to induce innocent people to confess to crimes they did not commit. The second problem is that false confessions, once taken, arouse a strong inference of guilt, thereby unleashing a chain of confirmation biases that make the consequences difficult to overcome despite innocence.

Supported by 100 plus years of basic psychology and the research reviewed herein, confession-induced corroboration inflation challenges a core premise in law. Both pretrial corroboration requirements and a harmless error analysis on appeal rest on the assumption that the corroborating evidence on record is nonredundant and independent of the confession. It now appears that this assumption is often incorrect, that the other evidence may be tainted by confession, and that the appearances of corroboration at pretrial and the sufficiency of evidence on appeal may be more illusory than real. Going forward, this conclusion has important implications for how criminal investigations are conducted (e.g., use of procedures designed to ensure that lay and expert witnesses are "blind" as to whether the suspect has confessed) and how the evidence, once gathered, is later evaluated in the courts (e.g., probing for the possibility of contamination across items of evidence that are allegedly independent and corroborative).

In recent years, psychologists have been critical of the problems with accuracy, error, subjectivity, and bias in various types of criminal evidence—prominently including eyewitness identification procedures, police interrogation

practices, and the so-called forensic identification sciences, all leading Saks and Koehler (2005) to predict a "coming paradigm shift." With regard to confessions, it now appears that this shift should encompass not only reforms that serve to minimize the risk of false confessions but measures designed to minimize the rippling *consequences* of those confessions—as in the case of Amanda Knox and others who are wrongfully convicted.

REFERENCES

Appleby, S., & Kassin, S. (2011). *When confessions trump DNA: Relative impacts of self-report and DNA evidence on juror decisions*. Paper presented at the meeting of the American Psychology-Law Society, Miami, FL.

Appleby, S. C., Hasel, L. E., & Kassin, S. M. (2011). Police-induced confessions: An empirical analysis of their content and impact. *Psychology, Crime & Law*. Advance online publication. doi:10.1080/1068316X.2011.613389

Arizona v. Fulminante, 499 U.S. 279 (1991).

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41,* 258–290. doi:10.1037/h0055756

Ask, K., & Granhag, P. A. (2007). Motivational bias in criminal investigators' judgments of witness reliability. *Journal of Applied Social Psychology, 37,* 561–591. doi:10.1111/j.1559-1816.2007.00175.x

Ayling, C. J. (1984). Corroborating false confessions: An empirical analysis of legal safeguards against false confessions. *Wisconsin Law Review, 4,* 1121–1204.

Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology, 91,* 612–625. doi:10.1037/0022-3514.91.4.612

Bilaisis, V. (1983). Harmless error: Abettor of courtroom misconduct. *Journal of Criminal Law & Criminology, 74,* 457–475. doi:10.2307/1143084

Blandón-Gitlin, I., Sperry, K., & Leo, R. A. (2011). Jurors believe interrogation tactics are not likely to elicit false confessions: Will expert witness testimony inform them otherwise? *Psychology, Crime & Law, 17,* 239–260. doi:10.1080/10683160903113699

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10,* 214–234. doi:10.1207/s15327957pspr1003_2

Bressan, P., & Dal Martello, M. F. (2002). *Talis Pater, Talis Filius*: Perceived resemblance and the belief in genetic relatedness. *Psychological Science, 13,* 213–218. doi:10.1111/1467-9280.00440

Brewer, N., & Wells, G. L. (2011). Eyewitness identification. *Current Directions in Psychological Science, 20,* 24–27. doi:10.1177/0963721410389169

Bruner, J. S., & Minturn, A. L. (1955). Perceptual identification and perceptual organization. *Journal of General Psychology, 53,* 21–28. doi:10.1080/00221309.1955.9710133

Bugelski, B. R., & Alampay, D. A. (1961). The role of frequency in developing perceptual sets. *Canadian Journal of Psychology, 15,* 205–211. doi:10.1037/h0083443

Burleigh, N. (2011). *The fatal gift of beauty: The trials of Amanda Knox.* New York, NY: Broadway Books.

Burns, S. (2011). *The Central Park five: A chronicle of a city wilding.* New York, NY: Knopf.

Center on Wrongful Convictions. (2010). *Other convictions in the face of exculpatory DNA.* Chicago, IL: Author.

Charman, S. D., Gregory, A. H., & Carlucci, M. (2009). Exploring the diagnostic utility of facial composites: Beliefs of guilt can bias perceived similarity between composite and suspect. *Journal of Experimental Psychology: Applied, 15,* 76–90. doi:10.1037/a0014682

Charman, S. D., & Wells, G. L. (2008). Can eyewitnesses correct for external influences on their lineup identifications? The actual/counterfactual assessment paradigm. *Journal of Experimental Psychology: Applied, 14,* 5–20. doi:10.1037/1076-898X.14.1.5

Cole, S. A. (2001). *Suspect identities.* Cambridge, MA: Harvard University Press.

Darley, J., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology, 44,* 20–33. doi:10.1037/0022-3514.44.1.20

Davis, D., & Leo, R. A. (2012). Interrogation-related regulatory decline: Ego depletion, failures of self-regulation, and the decision to confess. *Psychology, Public Policy, and Law*. Advance online publication. doi:10.1037/a0027367

Dempsey, C. (2010). *Murder in Italy: The shocking slaying of a British student, the accused American girl, and an international scandal.* New York, NY: Berkley Books.

Descartes, R. (1984). Principles of philosophy. In J. Cottingham, R. Stoothoff, & D. Murdoch (Eds. & Trans.), *The philosophical writings of Descartes* (Vol. 1, pp. 193–291). Cambridge, England: Cambridge University Press. (Original work published 1644)

Drizin, S. A., & Leo, R. A. (2004). The problem of false confessions in the post-DNA world. *North Carolina Law Review, 82,* 891–1007.

Dror, I. E., & Charlton, D. (2006). Why experts make errors. *Journal of Forensic Identification, 56,* 600–616.

Dror, I. E., & Cole, S. A. (2010). The vision in "blind" justice: Expert perception, judgment and visual cognition in forensic pattern recognition. *Psychonomic Bulletin & Review, 17,* 161–167. doi:10.3758/PBR.17.2.161

Dror, I. E., & Hampikian, G. (2011). Subjectivity and bias in forensic DNA mixture interpretation. *Science & Justice, 51,* 204–208. doi:10.1016/j.scijus.2011.08.004

Elaad, E. (2011). Effects of incomplete information on the detection of concealed crime details. *Applied Psychophysiology and Biofeedback, 36,* 159–171. doi:10.1007/s10484-011-9153-2

Elaad, E., Ginton, A., & Ben-Shakhar, G. (1994). The effects of prior expectations and outcome knowledge on polygraph examiners' decisions. *Journal of Behavioral Decision Making, 7,* 279–292. doi:10.1002/bdm.3960070405

Findley, K. A., & Scott, M. S. (2006). The multiple dimensions of tunnel vision in criminal cases. *Wisconsin Law Review, 2,* 291–397.

Firstman, R., & Salpeter, J. (2008). *A criminal injustice: A true crime, a false confession, and the fight to free Marty Tankleff.* New York, NY: Ballantine Books.

Fisher, G. (1968). Ambiguity of form: Old and new. *Perception & Psychophysics, 4,* 189–192. doi:10.3758/BF03210466

Gabbert, F., Memon, A., & Allan, K. (2003). Memory conformity: Can eyewitnesses influence each other's memories for an event? *Applied Cognitive Psychology, 17,* 533–543. doi:10.1002/acp.885

Garrett, B. L. (2010). The substance of false confessions. *Stanford Law Review, 62,* 1051–1119.

Garrett, B. L. (2011). *Convicting the innocent: Where criminal prosecutions go wrong.* Cambridge, MA: Harvard University Press.

Gerrig, R. J., & Prentice, D. A. (1991). The representation of fictional information. *Psychological Science, 2,* 336–340. doi:10.1111/j.1467-9280.1991.tb00162.x

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist, 46,* 107–119. doi:10.1037/0003-066X.46.2.107

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117,* 21–38. doi:10.1037/0033-2909.117.1.21

Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology, 65,* 221–233. doi:10.1037/0022-3514.65.2.221

Greathouse, S. M., & Kovera, M. B. (2009). Instruction bias and lineup presentation moderate the effects of administrator knowledge on eyewitness identification. *Law and Human Behavior, 33,* 70–82. doi:10.1007/s10979-008-9136-x

Greenberg, K. J. (Ed.). (2006). *The torture debate in America.* New York, NY: Cambridge University Press.

Gregory, W. L., Mowen, J. C., & Linder, D. E. (1978). Social psychology and plea bargaining: Applications, methodology, and theory. *Journal of Personality and Social Psychology, 36,* 1521–1530. doi:10.1037/0022-3514.36.12.1521

Grisham, J. (2010). *The confession.* New York, NY: Dell Books.

Gudjonsson, G. H. (2003). *The psychology of interrogations and confessions: A handbook.* Chichester, England: Wiley.

Gudjonsson, G. H., & Pearse, J. (2011). Suspect interviews and false confessions. *Current Directions in Psychological Science, 20,* 33–37. doi:10.1177/0963721410396824

Gudjonsson, G. H., Sigurdsson, J. F., & Sigfusdottir, I. D. (2009). Interrogation and false confessions among adolescents in seven European countries. What background and psychological variables best discriminate between false confessors and non-false confessors? *Psychology, Crime & Law, 15,* 711–728. doi:10.1080/10683160802516257

Guyll, M., Madon, S., Yang, Y., Scherr, K. C., Lannin, D., Smalarz, L., Wells, G. L., & Greathouse, S. (2012, March). *Physiologic reactions to interrogation stress: Differences between the innocent and the guilty.* Paper presented at the Annual Conference of the American Psychology-Law Society, San Juan, Puerto Rico.

Hamilton, D. L., & Zanna, M. P. (1974). Context effects in impression formation: Changes in connotative meaning. *Journal of Personality and Social Psychology, 29,* 649–654. doi:10.1037/h0036633

Hampikian, G., West, E., & Akselrod, O. (2011). The genetics of innocence: Analysis of 194 U.S. DNA exonerations. *Annual Review of Genomics and Human Genetics, 12,* 97–120. doi:10.1146/annurev-genom-082509-141715

Hartwig, M., & Bond, C. F., Jr. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin, 137,* 643–659. doi:10.1037/a0023589

Hartwig, M., Granhag, P. A., & Strömwall, L. A. (2007). Guilty and innocent suspects' strategies during police interrogations. *Psychology, Crime & Law, 13,* 213–227. doi:10.1080/10683160600750264

Hasel, L. E., & Kassin, S. M. (2009). On the presumption of evidentiary independence: Can confessions corrupt eyewitness identifications? *Psychological Science, 20,* 122–126. doi:10.1111/j.1467-9280.2008.02262.x

Heider, F. (1958). *The psychology of interpersonal relations.* New York, NY: Wiley. doi:10.1037/10628-000

Henkel, L. A. (2008). Jurors' reactions to recanted confessions: Do the defendant's personal and dispositional characteristics play a role? *Psychology, Crime & Law, 14,* 565–578. doi:10.1080/10683160801995247

Henkel, L. A., Coffman, K. A. J., & Dailey, E. M. (2008). A survey of people's attitudes and beliefs about false confessions. *Behavioral Sciences & the Law, 26,* 555–584. doi:10.1002/bsl.826

Hill, C., Memon, A., & McGeorge, P. (2008). The role of confirmation bias in suspect interviews: A systematic evaluation. *Legal and Criminological Psychology, 13,* 357–371. doi:10.1348/135532507X238682

Honts, C., Kassin, S., & Forrest, K. (2009). *Polygraph examiners unable to discriminate true and false juvenile confessions: Reid training detrimental.* Paper presented at the meeting of the American Psychology-Law Society, San Antonio, TX.

Inbau, F. E., Reid, J. E., Buckley, J. P., & Jayne, B. C. (2013). *Criminal interrogation and confessions* (5th ed.). Burlington, MA: Jones & Bartlett Learning.

James, W. (1890). *The principles of psychology* (Vol. 2). New York, NY: Holt. doi:10.1037/11059-000

Jones, E. E. (1990). *Interpersonal perception.* New York, NY: Freeman.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. *Advances in Experimental Social Psychology, 2,* 219–266. doi:10.1016/S0065-2601(08)60107-0

Kassin, S. M. (1997). The psychology of confession evidence. *American Psychologist, 52,* 221–233. doi:10.1037/0003-066X.52.3.221

Kassin, S. M. (2005). On the psychology of confessions: Does *innocence* put *innocents* at risk? *American Psychologist, 60,* 215–228. doi:10.1037/0003-066X.60.3.215

Kassin, S. M. (2008). The psychology of confessions. *Annual Review of Law and Social Science, 4,* 193–217. doi:10.1146/annurev.lawsocsci.4.110707.172410

Kassin, S. M., Bogart, D., & Kerner, J. (2012). Confessions that corrupt: Evidence from the DNA exoneration case files. *Psychological Science, 23,* 41–45. doi:10.1177/0956797611422918

Kassin, S. M., Drizin, S. A., Grisso, T., Gudjonsson, G. H., Leo, R. A., & Redlich, A. D. (2010). Police-induced confessions: Risk factors and recommendations. *Law and Human Behavior, 34,* 3–38. doi:10.1007/s10979-009-9188-6

Kassin, S. M., Goldstein, C. J., & Savitsky, K. (2003). Behavioral confirmation in the interrogation room: On the dangers of presuming guilt. *Law and Human Behavior, 27,* 187–203. doi:10.1023/A:1022599230598

Kassin, S. M., & Gudjonsson, G. H. (2004). The psychology of confessions: A review of the literature and issues. *Psychological Science in the Public Interest, 5,* 33–67. doi:10.1111/j.1529-1006.2004.00016.x

Kassin, S. M., & Gudjonsson, G. H. (2005). True crimes, false confessions: Why do innocent people confess to crimes they did not commit? *Scientific American Mind, 16,* 24–31. doi:10.1038/scientificamericanmind0605-24

Kassin, S. M., & Kiechel, K. L. (1996). The social psychology of false confessions: Compliance, internalization, and confabulation. *Psychological Science, 7,* 125–128. doi:10.1111/j.1467-9280.1996.tb00344.x

Kassin, S. M., & Kukucka, J. (2012). *Confession errors as "structural defects."* Poster presented at the annual meeting of the American Psychology-Law Society, San Juan, Puerto Rico.

Kassin, S. M., Leo, R. A., Meissner, C. A., Richman, K. D., Colwell, L. H., Leach, A.-M., & La Fon, D. (2007). Police interviewing and interrogation: A self-report survey of police practices and beliefs, *Law and Human Behavior, 31,* 381–400. doi:10.1007/s10979-006-9073-5

Kassin, S. M., Meissner, C. A., & Norwick, R. J. (2005). "I'd know a false confession if I saw one": A comparative study of college students and police investigators. *Law and Human Behavior, 29,* 211–227. doi:10.1007/s10979-005-2416-9

Kassin, S. M., & Neumann, K. (1997). On the power of confession evidence: An experimental test of the "fundamental difference" hypothesis. *Law and Human Behavior, 21,* 469–484. doi:10.1023/A:1024871622490

Kassin, S. M., & Norwick, R. (2004). Why people waive their Miranda rights: The power of innocence. *Law and Human Behavior, 28,* 211–221. doi:10.1023/B:LAHU.0000022323.74584.f5

Kassin, S. M., & Sukel, H. (1997). Coerced confessions and the jury: An experimental test of the harmless error rule. *Law and Human Behavior, 21,* 27–46. doi:10.1023/A:1024814009769

Kassin, S. M., & Wrightsman, L. S. (1980). Prior confessions and mock juror verdicts. *Journal of Applied Social Psychology, 10,* 133–146. doi:10.1111/j.1559-1816.1980.tb00698.x

Kassin, S. M., & Wrightsman, L. S. (1985). Confession evidence. In S. Kassin & L. Wrightsman (Eds.), *The psychology of evidence and trial procedure* (pp. 67–94). Beverly Hills, CA: Sage.

Kennard, J. B., & Kassin, S. M. (2009). *Racial differences in Miranda waiver*. Poster presented at the annual meeting of the American Psychology-Law Society, San Antonio, TX.

Lange, N. D., Thomas, R. P., Dana, J., & Dawes, R. M. (2011). Contextual biases in the interpretation of auditory evidence. *Law and Human Behavior, 35,* 178–187. doi:10.1007/s10979-010-9226-4

Lassiter, G. D. (Ed.). (2004). *Interrogations, confessions, and entrapment.* New York, NY: Kluwer Academic.

Lassiter, G. D., & Meissner, C. A. (Eds.). (2010). *Police interrogations and false confessions: Current research, practice, and policy recommendations.* Washington, DC: American Psychological Association. doi:10.1037/12085-000

Leeper, R. (1935). A study of a neglected portion of the field of learning—the development of sensory organization. *Journal of Genetic Psychology, 46,* 41–75.

Leo, R. A. (2008). *Police interrogation and American justice.* Cambridge, MA: Harvard University Press.

Leo, R. A., & Liu, B. (2009). What do potential jurors know about police interrogation techniques and false confessions? *Behavioral Sciences & the Law, 27,* 381–399. doi:10.1002/bsl.872

Leo, R. A., & Ofshe, R. J. (1998). The consequences of false confessions: Deprivations of liberty and miscarriages of justice in the age of psychological interrogation. *Journal of Criminal Law and Criminology, 88,* 429–496. doi:10.2307/1144288

Levine, T. R., Kim, R. K., & Blair, J. P. (2010). (In)accuracy at detecting true and false confessions and denials: An initial test of a projected motive model of veracity judgments. *Human Communication Research, 36,* 82–102. doi:10.1111/j.1468-2958.2009.01369.x

Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect." *Communication Monographs, 66,* 125–144. doi:10.1080/03637759909376468

Madon, S., Guyll, M., Scherr, K. C., Greathouse, S., & Wells, G. L. (2012). Temporal discounting: The differential effect of proximal and distal consequences on confession decisions. *Law and Human Behavior, 36,* 13–20. doi:10.1007/s10979-011-9267-3

Malloy, L. C., & Lamb, M. E. (2010). Biases in judging victims and suspects whose statements are inconsistent. *Law and Human Behavior, 34,* 46–48. doi:10.1007/s10979-009-9211-y

Martin, A. (2011, November 27). The prosecution's case against DNA. *The New York Times Magazine,* p. MM44.

McCormick, C. T. (1972). *Handbook of the law of evidence* (2nd ed.). St. Paul, MN: West.

Meissner, C. A., Russano, M. B., & Narchet, F. M. (2010). The importance of a laboratory science for improving the diagnostic value of confession evidence. In G. D. Lassiter & C. A. Meissner (Eds.), *Police interrogations and false confessions: Current research, practice, and policy recommendations* (pp. 111–126). Washington, DC: American Psychological Association. doi:10.1037/12085-007

Moore, T. E., & Gagnier, K. (2008). "You can talk if you want to": Is the police caution on the 'right to silence' understandable? *Criminal Reports, 51,* 233–249.

Najdowski, C. J. (2011). Stereotype threat in criminal interrogations: Why innocent Black suspects are at risk for confessing falsely. *Psychology, Public Policy, and Law, 17,* 562–591. doi:10.1037/a0023741

Narchet, F. M., Meissner, C. A., & Russano, M. B. (2011). Modeling the influence of investigator bias on the elicitation of true and false confessions. *Law and Human Behavior, 35,* 452–465. doi:10.1007/s10979-010-9257-x

Natapoff, A. (2009*). Snitching: Criminal informants and the erosion of American justice.* New York, NY: New York University Press.

National Academy of Sciences. (2009). *Strengthening forensic science in the United States: A path forward.* Washington, DC: National Academies Press.

Neuschatz, J. S., Lawson, D. S., Swanner, J. K., Meissner, C. A., & Neuschatz, J. S. (2008). The effects of accomplice witnesses and jailhouse informants on jury decision making. *Law and Human Behavior, 32,* 137–149. doi:10.1007/s10979-007-9100-1

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2,* 175–220. doi:10.1037/1089-2680.2.2.175

Nordgren, L. F., McDonnell, M. M., & Loewenstein, G. (2011). What constitutes torture? Psychological impediments to an objective evaluation of enhanced interrogation tactics. *Psychological Science, 22,* 689–694. doi:10.1177/0956797611405679

Olson, E. A., & Charman, S. D. (2011). "But can you prove it?" - Examining the quality of innocent suspects' alibis. *Psychology, Crime & Law.* Advance online publication. doi:10.1080/1068316X.2010.505567

Owen-Kostelnik, J., Reppucci, N., & Meyer, J. (2006). Testimony and interrogation of minors: Assumptions about maturity and morality. *American Psychologist, 61,* 286–304. doi:10.1037/0003-066X.61.4.286

Perillo, J. T., & Kassin, S. M. (2011). Inside interrogation: The lie, the bluff, and false confessions. *Law and Human Behavior, 35,* 327–337. doi:10.1007/s10979-010-9244-2

Povoledo, E. (2011, June 29). Italian experts question evidence in Knox case. *The New York Times.* Retrieved from http://www.nytimes.com/2011/06/30/world/europe/30knox.html

Redlich, A. D. (2010). False confessions and false guilty pleas: Similarities and differences. In G. D. Lassiter & C. Meissner (Eds.), *Interrogations and confessions: Current research, practice, and policy* (pp. 49–66). Washington, DC: APA Books. doi:10.1037/12085-003

Redlich, A. D., Ghetti, S., & Quas, J. A. (2008). Perceptions of children during a police interview: A comparison of suspects and alleged victims. *Journal of Applied Social Psychology, 38,* 705–735. doi:10.1111/j.1559-1816.2007.00323.x

Redlich, A. D., Kulish, R., & Steadman, H. J. (2011). Comparing true and false confessions among persons with serious mental illness. *Psychology, Public Policy, and Law, 17,* 394–418. doi:10.1037/a0022918

Redlich, A. D., Summers, A., & Hoover, S. (2010). Self-reported false confessions and false guilty pleas among offenders with mental illness. *Law and Human Behavior, 34,* 79–90. doi:10.1007/s10979-009-9194-8

Rich, N. (2011, June 27). The neverending nightmare of Amanda Knox. *Rolling Stone*. Retrieved from http://www.rollingstone.com/culture/news/the-neverending-nightmare-of-amanda-knox-20110627

Rimer, S. (2002, February 6). Convict's DNA sways labs, not a determined prosecutor. *The New York Times,* p. A14.

Rogers, R., Gillard, N. D., Wooley, C. N., & Fiduccia, C. E. (2011). Decrements in *Miranda* abilities: An investigation of situational effects via a mock-crime paradigm. *Law and Human Behavior, 35,* 392–401. doi:10.1007/s10979-010-9248-y

Rogers, R., Hazelwood, L., Sewell, K., Harrison, K., & Shuman, D. (2008). The language of Miranda warnings in American jurisdictions: A replication and vocabulary analysis. *Law and Human Behavior, 32,* 124–136. doi:10.1007/s10979-007-9091-y

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology, 10,* 173–220. doi:10.1016/S0065-2601(08)60357-3

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science, 309,* 892–895. doi:10.1126/science.1111565

Scheck, B., Neufeld, P., & Dwyer, J. (2000). *Actual innocence: Five days to execution and other dispatches from the wrongly convicted.* New York, NY: Doubleday.

Scherr, K. C., & Madon, S. (2011). You have the right to understand: The deleterious effect of stress on suspects' ability to comprehend *Miranda. Law and Human Behavior.* Advance online publication. doi:10.1007/s10979-011-9283-3

Shaw, J. S., Garven, S., & Wood, J. M. (1997). Co-witness information can have immediate effects on eyewitness memory reports. *Law and Human Behavior, 21,* 503–523. doi:10.1023/A:1024875723399

Skagerberg, E. M. (2007). Co-witness feedback in line-ups. *Applied Cognitive Psychology, 21,* 489–497. doi:10.1002/acp.1285

Spinoza, B. (1982). *The Ethics and selected letters* (S. Feldman, Trans.). Indianapolis, IN: Hackett. (Original work published 1677)

Swanner, J. K., & Beike, D. R. (2010). Incentives increase the rate of false but not true secondary confessions from informants with an allegiance to a suspect. *Law and Human Behavior, 34,* 418–428. doi:10.1007/s10979-009-9212-x

Swanner, J. K., Beike, D. R., & Cole, A. T. (2010). Snitching, lies and computer crashes: An experimental investigation of secondary confessions. *Law and Human Behavior, 34,* 53–65. doi:10.1007/s10979-008-9173-5

Thompson-Cannino, J., Cotton, R., & Torneo, E. (2009). *Picking Cotton: Our memoir of injustice and redemption.* New York, NY: St. Martin's Press.

Tor, A., Gazal-Ayal, O., & Garcia, S. M. (2010). Fairness and the willingness to accept plea bargain offers. *Journal of Empirical Legal Studies, 7,* 97–116. doi:10.1111/j.1740-1461.2009.01171.x

Tyler Edmonds v. State of Mississippi, No. 2004-CT-02081-SCT, May 10, 2007.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities.* Chichester, England: Wiley.

Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest, 11,* 89–121. doi:10.1177/1529100610390861

Wallace, D. B., & Kassin, S. M. (2012). Harmless error analysis: How do judges respond to confession errors? *Law and Human Behavior, 36,* 151–157. doi:10.1007/s10979-010-9262-0

Warden, R. (2004). *The snitch system: How incentivized witnesses put 38 innocent Americans on death row.* Evanston, IL: Northwestern University School of Law, Center on Wrongful Convictions.

Warden, R., & Drizin, S. A. (Eds.). (2009). *True stories of false confessions.* Evanston, IL: Northwestern University Press.

Watkins, M. J., & Peynircioglu, Z. F. (1984). Determining perceived meaning during impression formation: Another look at the meaning change hypothesis. *Journal of Personality and Social Psychology, 46,* 1005–1016. doi:10.1037/0022-3514.46.5.1005

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest, 7,* 45–75.

Wells, G. L., Small, M., Penrod, S., Malpass, R., Fulero, S., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22,* 603–647. doi:10.1023/A:1025750605807

Wells, T., & Leo, R. A. (2008). *The wrong guys: Murder, false confessions, and the Norfolk Four.* New York, NY: The New Press.

Williamson, T. (Ed.). (2006). *Investigative interviewing: Rights, research, regulation.* Devon, England: Willan.

Corrections and Updates to Kassin (2012)

In the article "Why Confessions Trump Innocence," by Saul M. Kassin (*American Psychologist,* Vol. 67, No. 6, pp. 431–445, this issue; published Online First April 30, 2012), minor corrections should be made in the description of the Amanda Knox case. The author thanks those whose feedback prompted these changes: (1) On p. 431, paragraph 1, line 5, "the only one" should be deleted; (2) on p. 431, paragraph 2, lines 10–12, "She was told, falsely, that Sollecito, her boyfriend, disavowed her alibi and that physical evidence placed her at the scene" should read "She was told that Sollecito, her boyfriend, disavowed her alibi, which he later retracted, and that physical evidence placed her at the scene, which was not true"; (3) on p. 433, paragraph 2, lines 3–4, "her English roommates left Perugia;" should be deleted; (4) on p. 436, paragraph 3, lines 8–10, "Two weeks later, the rapist whose DNA was found in sperm and other biological matter at the crime scene was apprehended" should read "Two weeks later, Rudy Guede, who was convicted of murdering Meredith Kercher and whose DNA was found inside her body and throughout the crime scene, was apprehended"; (5) on p. 436, paragraph 3, line 13, "the rapist" should be replaced with "Guede"; (6) on p. 441, paragraph 3, line 13, "Knox's British roommates" should be replaced with "Kercher's friends."

For the most recent and "official" opinion on this case, see the Hellmann-Zanetti Report on the Acquittal of Amanda Knox and Raffaele Sollecito–December 16, 2011, Translated into English (http://hellmannreport.wordpress.com/contents/). In addition, since the article was published Online First, the prosecution's appeal of Knox's acquittal has been scheduled to be heard by Italy's highest appeals court on March 25, 2013.

DOI: 10.1037/a0029885

Journal of Applied Research in Memory and Cognition 2 (2013) 42–52

Contents lists available at SciVerse ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Target article

The forensic confirmation bias: Problems, perspectives, and proposed solutions

Saul M. Kassin[a,*], Itiel E. Dror[b], Jeff Kukucka[a]

[a] John Jay College of Criminal Justice, United States
[b] University College London (UCL), United Kingdom

ARTICLE INFO

Article history:
Received 28 October 2012
Received in revised form
29 December 2012
Accepted 3 January 2013

Keywords:
Context effects
Expectancy effects
Confirmation bias

ABSTRACT

As illustrated by the mistaken, high-profile fingerprint identification of Brandon Mayfield in the Madrid Bomber case, and consistent with a recent critique by the National Academy of Sciences (2009), it is clear that the forensic sciences are subject to contextual bias and fraught with error. In this article, we describe classic psychological research on primacy, expectancy effects, and observer effects, all of which indicate that context can taint people's perceptions, judgments, and behaviors. Then we describe recent studies indicating that confessions and other types of information can set into motion *forensic confirmation biases* that corrupt lay witness perceptions and memories as well as the judgments of experts in various domains of forensic science. Finally, we propose best practices that would reduce bias in the forensic laboratory as well as its influence in the courts.

1. The problem

On March 11, 2004, a coordinated series of bombs exploded in four commuter trains in Madrid. The explosions killed 191 people, wounded 1800 others, and set into motion a full-scale international investigation. On the basis of a latent fingerprint lifted from a bag containing detonating devices, the U.S. Federal Bureau of Investigation (FBI) positively identified Brandon Mayfield, an American Muslim from the state of Oregon. Subsequent to 9–11, Mayfield had been on an FBI watch list. Following standard protocol, a number of FBI fingerprint examiners independently concluded that the fingerprint was definitely that of Mayfield. After being arrested and appearing in court, Mayfield requested to have a fingerprint examiner on the defense team examine the prints. That fingerprint examiner concurred with the judgment that the print was Mayfield's. Soon thereafter, however, the Spanish authorities matched the prints to the real Madrid bomber, an Algerian national by the name of Ouhnane Daoud. Following an internal investigation at the FBI and a report by the Office of the Inspector General (OIG, 2006), "confirmation bias" was listed as a contributing factor to the erroneous identification. At that point, the U.S. government issued a formal apology, and paid two million dollars in compensation.

The FBI has rigorous standards of training and practice and highly competent forensic examiners. It is considered one of the best, if not *the* best forensic laboratories in the U.S., if not in the entire world. Thus, it was not easy to dismiss the error and claim it to be the product of mere "bad apples." The Mayfield case (preceded by a decade in which the U.S. Supreme Court had sought to curb the introduction at trial of experts in junk science—see *Daubert v. Merrell Dow Pharmaceuticals, 1993*; *Kumho Tire Co. v. Carmichael*, 1999), along with improprieties discovered in various state laboratories, have come together to draw attention to forensic science and to the fact that is not infallible. Forensic science errors have also surfaced with alarming frequency in DNA exoneration cases and other wrongful convictions (Garrett, 2011; http://www.innocenceproject.org/fix/Crime-Lab-Oversight.php). In "The genetics of innocence," Hampikian, West, and Akselrod (2011) found that several types of forensic science testimony had been used to wrongfully convict innocent individuals. In cases where trial transcripts or reliable forensic science data were available for review, 38% contained incorrect serology testimony, which is highly regarded. In addition, 22% involved hair comparisons; 3% involved bite mark comparisons; and 2% involved fingerprint comparisons.

The National Academy of Sciences (NAS, 2009) published a scathing assessment of a broad range of forensic disciplines. Included in this critique were toolmarks and firearms; hair and fiber analysis; impression evidence; blood spatter; fibers; handwriting; and even fingerprints—until recently considered infallible. NAS concluded that there are problems with standardization, reliability, accuracy and error, and the potential for contextual bias. Specifically, the NAS report went on to advise that: "These disciplines need to develop rigorous protocols to guide these subjective interpretations and pursue equally rigorous research and evaluation programs. The development of such research programs can benefit significantly from other areas, notably from the large body

* Corresponding author.
E-mail address: Skassin@jjay.cuny.edu (S.M. Kassin).

2211-3681/$ – see front matter © 2013 Society for Applied Research in Memory and Cognition. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jarmac.2013.01.001

of research on the evaluation of observer performance in diagnostic medicine and from the findings of cognitive psychology on the potential for bias and error in human observers" (p. 8).

The criticisms of the forensic sciences are twofold. First is the realization that too often the stimulus does not compel a perceptual judgment that is objective and, hence, there is a concern both for inter-rater reliability across experts and for intra-test reliability over time within experts. In many forensic disciplines, the human examiner is the main instrument of analysis. It is the forensic expert who compares visual patterns and determines if they are "sufficiently similar" to conclude that they originate from the same source (e.g., whether two fingerprints were made by the same finger, whether two bullets were fired from the same gun, or whether two signatures were made by the same person). However, determinations of "sufficiently similar" have no criteria and quantification instruments; these judgments are subjective. Indeed, a recent study has shown that when the same fingerprint evidence is given to the same examiners, they reach different conclusions approximately 10% of the time (Ulery, Hicklin, Buscaglia, & Roberts, 2012). Dror et al. (2011) have shown not only that the decisions are inconsistent but that even the initial perception of the stimulus, prior to comparison, lack inter- and intra-expert consistency.

Following from this realization about the lack of reliability is a corollary concern that forensic experts' judgments are "biasable"—that is, they are significantly influenced by psychological factors (Dror & Cole, 2010; Dror & Rosenthal, 2008). The biasability of forensic science is a particular concern because forensic experts work within a variety of contextual influences: Knowing the nature and details of the crime, being pressured by detectives; working within—and as part of—the police; the use of computer-generated lists that feature some suspects ahead of others; appearing in court within an adversarial criminal justice system. Describing the various sources of bias, Saks, Risinger, Rosenthal, and Thompson (2003) note that examiners often receive direct communications from police (e.g., in transmittal letters that accompany submitted evidence, in person, and by phone), that there is often cross-communication among different examiners involved in a case (e.g., via informal channels or as mandated in "peer review" processes designed to ensure the reasonableness of conclusions), and that police and prosecutors sometimes respond to non-supportive test results by requesting a re-examination. In short, the contextual influences that impinge on forensic examiners are numerous and they come in many forms, some of which are subtle. The erroneous identification in the Madrid bomber case illustrated a number of psychological factors at work (e.g., the latent fingerprint was examined against a pre-existing "target," without first being properly analyzed in isolation; the examiners were prearmed with contextual information, leading them to be suspicious of their target; and the case was high in profile and time-urgent, increasing the need for closure).

In this article, we overview prior critiques of the forensic sciences and specific cases in which experts have rendered judgments that were fraught with bias and error. Then we consider classic psychological research on primacy, expectancy effects, and observer effects, and the various confirmation biases that can taint people's perceptions, judgments, and behaviors. Then we examine recent empirical work on confirmation biases in various domains of forensic science. Finally we use psychology to propose best practices that would minimize such effects—both in the crime laboratory and in the courtroom.

2. The forensic sciences: accuracy and error

For over 100 years forensic science disciplines have produced evidence used both to prosecute and convict criminals as well as to exonerate and release those who are innocent. The domains of forensic science are varied and include judgments of fingerprints, firearms examinations, toolmarks, bite marks, tire and shoe impressions, bloodstain pattern analysis, handwriting, hair, coatings such as paint and chemicals—including drugs and such materials as fibers, fluids, fire and explosive analysis, digital evidence, and serological analysis.

Since the 1990s, advances in DNA technology have proved particularly useful in these regards. Many previously unsolved crimes have been solved because of DNA samples left in hair, semen, blood, skin, and saliva. Often, however, these DNA cases have revealed that faulty forensic sciences have contributed to the wrongful convictions of innocent people. As exposed by more than 300 DNA exonerations identified by the Innocence Project, two sets of problems have come to light: (1) Forensic science judgments are often derived from inadequate testing and analysis, if not outright fabrication; and (2) Experts often give imprecise or exaggerated testimony, drawing conclusions not supported by the data—in some cases drawing charges of misconduct. Indeed, some form of invalid or improper forensic science was a contributing factor in the original convictions of more than half of all DNA exonerees (Garrett, 2011; http://www.innocenceproject.org/understand/Unreliable-Limited-Science.php).

In cases that are not subject to bias, certain forensic sciences—such as latent fingerprint identifications—offer a potentially powerful tool in administering justice (e.g., Tangen, Thompson, & McCarthy, 2011; Ulery, Hicklin, Buscaglia, & Roberts, 2011). In most domains, however, there are no quantitatively precise objective measures and no instruments of measurement—just partial samples from a crime scene to be compared against a particular suspect. No two patterns are identical, so an examiner invariably must determine whether they are "sufficiently similar" (a term that has yet to be defined or quantified) to conclude that they originate from the same source. The absence of objective standards is reflected in the lack of consistency not only between examiners but within examiners over time. Hence, not only do inter-variations exist, but intra-variations show that the same examiner inspecting the same data on multiple occasions may reach different conclusions (Ulery et al., 2012). The lack of reliability indicates that the identification process can be subjective and that judgments are susceptible to bias from other sources. This is especially problematic in cases that contain complex forms of forensic evidence, as is often the case in evidence gathered in crime scene.

Popular TV programs, such as *CSI*, communicate a false belief in the powers of forensic science, a problem that can be exacerbated when forensic experts overstate the strength of the evidence. Such occurrences are common when you consider the following: (1) Across many domains, experts are often overconfident in their abilities (e.g., Baumann, Deber, & Thompson, 1991); (2) the courts, for the most part, have blindly accepted forensic science evidence without much scrutiny (Mnookin et al., 2011); (3) errors are often not apparent in the forensic sciences because ground truth is often not known as a matter of certainty; (4) many forensic examiners work for police and appear in court as advocates for the prosecution; and (5) many forensic examiners consider themselves objective and immune to bias. As stated by the Chair of the Fingerprint Society: "Any fingerprint examiner who comes to a decision on identification and is swayed either way in that decision making process under the influence of stories and gory images is either totally incapable of performing the noble tasks expected of him/her or is so immature he/she should seek employment at Disneyland" (Leadbetter, 2007).

3. Classic confirmation biases: a psychological perspective

Over the years, research has identified a number of confirmation biases by which people tend to seek, perceive, interpret, and create new evidence in ways that verify their preexisting beliefs. Confirmation biases are a pervasive psychological phenomenon. Classic studies showed that prior exposure to images of a face or a body, an animal or a human, or letters or numbers, can bias what people see in an ambiguous figure. More recent research shows that our impressions of other people can similarly be tainted.

Recognition of confirmation bias as a human phenomenon is not new. Julius Caesar is cited to have said that "Men freely believe that which they desire" (e.g., Hochschild, 2008). References can also be found in the writings of William Shakespeare and Francis Bacon (Risinger, Saks, Thompson, & Rosenthal, 2002). Indeed, Nickerson (1998) notes that confirmation biases may be implicated in "a significant fraction of the disputes, altercations, and misunderstandings that occur among individuals, groups, and nations"—including, among others, the witch trials of Western Europe and New England, the continuation of ineffective medical treatments, inaccurate medical diagnoses, and adherence to erroneous scientific theories (p. 175).

3.1. Perceptual and cognitive effects

Contemporary work on confirmation biases began with classic research suggesting that the perception of a stimulus is not solely a function of the stimulus itself (i.e., "bottom-up" processing), but is also shaped by the qualities of the observer (i.e., "top-down" processing). For example, Bruner and Goodman (1947) asked children to estimate the size of coins from memory and found that children of low-SES overestimated the size of the coins to a greater degree than did children of high SES. Bruner and Potter (1964) demonstrated that one's expectations can also interfere with visual recognition. Participants were shown photographs of common objects (e.g., a dog, a fire hydrant, etc.) that had been blurred to various degrees, and then watched as the pictures were gradually brought into focus. The blurrier the photographs were at the start, the less able participants were to correctly recognize the objects later. Bruner and Potter explained these results by noting that participants readily generated hypotheses about the blurry images and then maintained these beliefs even as the pictures came into focus. Using simple ambiguous ("reversible") figures, other research as well showed that expectations shape perception (Boring, 1930; Leeper, 1935; for a compendium of such figures, see Fisher, 1968).

Recent studies have demonstrated similar effects using more complex stimuli. For example, Bressan and Dal Martello (2002) showed participants photographs of adult-child pairs and asked them to rate their facial resemblance. When led to believe that the adult and child were genetically related (e.g., parent and offspring), participants rated their facial similarity as higher – even when the two were not truly related. Other studies have similarly shown that people perceive more similarity between a suspect and a facial composite when led to believe the suspect is guilty (Charman, Gregory, & Carlucci, 2009); and people hear more incrimination in degraded speech recordings when the interviewee was thought to be a crime suspect (Lange, Thomas, Dana, & Dawes, 2011).

To sum up: A wealth of evidence indicates that an observer's expectations can impact visual and auditory perception. Although similar effects can be driven by motivation (Balcetis & Dunning, 2006, 2010; Radel & Clement-Guillotin, 2012), confirmation biases are a natural and automatic feature of human cognition that can occur in the absence of self-interest (Nickerson, 1998) and operate without conscious awareness (Findley & Scott, 2006; Kunda, 1990).

3.2. Social perception effects

Strong expectancy effects can also contaminate the processes of social perception. This research literature can be traced to Asch's (1946) initial finding of primacy effects in impression formation by which information about a person presented early in a sequence is weighed more heavily than information presented later which is ignored, discounted, or assimilated into the early-formed impression. Illustrating the process of assimilation, or "change of meaning" hypothesis, later research revealed that depending on one's first impression of a person, the word "proud" can mean self-respecting or conceited; "critical" can mean astute or picky; and "impulsive" can mean spontaneous or reckless (Hamilton & Zanna, 1974; Watkins & Peynircioglu, 1984). As a result of these processes, additional research has shown that beliefs, once they take root, can persist even after the evidence on which they were based has been discredited (Anderson, Lepper, & Ross, 1980). In fact, the presence of objective evidence that can be selectively interpreted may *exacerbate* the biasing effects of pre-existing beliefs (Darley & Gross, 1983).

Research on confirmatory hypothesis testing also explains the power and resistance to change of first impressions. In a classic experiment, Wason (1960) gave participants a three-number sequence, challenged them to discern the rule used to generate the set, and found that very few discovered the correct rule because once they seized upon a hypothesis they would search only for confirming evidence (see also Klayman & Ha, 1997). In a social-interactional context, Snyder and Swann (1978) brought together pairs of participants for a getting-acquainted interview. In each pair, interviewers were led to believe that their partner was either introverted or extroverted. Expecting a certain kind of person, participants unwittingly sought evidence that would confirm their expectations: Those in the introverted condition chose to ask mostly introvert-oriented questions ("Have you ever felt left out of some social group?"); those in the extroverted condition asked extrovert-oriented questions ("How do you liven up a party?"). In doing so, interviewers procured support for their beliefs, causing neutral observers who later listened to the tapes to perceive the interviewees as introverted or extroverted on the basis of their randomly assigned condition.

The fact that people can be jaded by existing beliefs is a phenomenon of potential consequence in forensic settings. In one study, participants reviewed a mock police file of a crime investigation that contained weak circumstantial evidence pointing to a possible suspect. Some participants but not others were asked to form and state an initial hypothesis as to the likely offender. Those who did so proceeded to search for additional evidence and interpret that evidence in ways that confirmed their hypothesis. Hence, a weak suspect became the *prime* suspect (O'Brien, 2009). In another study, Kassin, Goldstein, and Savitsky (2003) had some participants but not others commit a mock crime, after which all were questioned by interrogators who by random assignment were led to presume guilt or innocence. Interrogators who presumed guilt asked more incriminating questions, conducted more coercive interrogations, and tried harder to get the suspect to confess. In turn, this more aggressive style made the suspects sound defensive and led observers who later listened to the tapes to judge them as guilty, even when they were innocent. Follow-up research has confirmed variants of this latter chain of events in the context of suspect interviews (Hill, Memon, & McGeorge, 2008; Narchet, Meissner, & Russano, 2011).

An individual's prior beliefs can produce dramatic behavioral consequences as well, often setting into motion a three-step behavioral confirmation process by which a perceiver forms an impression of a target person, interacts in a manner that is consistent with that impression, and causes the target person unwittingly

to adjust his or her behavior. The net result: a process that transforms expectations into reality (Darley & Fazio, 1980; Rosenthal & Jacobson, 1966; Snyder & Swann, 1978).

In an early demonstration of this phenomenon, Rosenthal and Fode (1963) reported on an experimenter expectancy effect, whereby an experimenter who is aware of the hypothesis of a study and the condition to which a participant is assigned can unwittingly produce results consistent with the expected outcome. Thus, when students were led to believe that the rats they would be training at maze learning were bright or dull, those rats believed to be bright learned more quickly (for an overview of this research, see Rosenthal, 2002). In subsequent research on *teacher* expectancy effects, Rosenthal and Jacobson (1966) extended these findings to human participants and found that when elementary school teachers were led to believe that certain of their students, randomly assigned, were on the verge of an intellectual growth spurt, those selected students exhibited greater improvement in academic tests eight months later. Whether training rats or teaching students, it appears that people unwittingly act upon their beliefs in ways that produced the expected outcomes. Although the interpretation of the teacher expectancy effect is a source of some controversy (Jussim, 2012), self-fulfilling prophecies have amply been demonstrated not only in the laboratory but in schools and other types of organizations as well (for reviews, see Kierein & Gold, 2000; McNatt, 2000).

3.3. Cognitive and motivational sources of bias

It is clear that belief-confirming thought processes are an inherent feature of human cognition. In their classic studies, Tversky and Kahneman (1974) demonstrated that people naturally rely on various cognitive *heuristics* – and that heuristic thinking, while generally beneficial, can also produce systematic errors in judgment, especially where strong prior expectations exist. Over time, and across a range of domains, basic psychological research has shown that strong expectations provide a sufficient and unwitting trigger of our tendency to seek, perceive, interpret, and create new evidence in ways that verify preexisting beliefs.

At times, confirmation biases can be fueled by *motivational goals*. Kunda (1990) argued that motivation influences reasoning indirectly as a result of two types of goals: *accuracy goals*, where individuals strive to form an accurate belief or judgment, and *directional goals*, where individuals seek a particular desired conclusion. In the latter case, people maintain an "illusion of objectivity" that prevents them from recognizing that their cognition has been tainted by preference or desire (Kunda, 1990, p. 483). Motivated reasoning is pervasive. Hence, people exhibit a ubiquitous self-serving positivity bias in the attributions they make for their own successes and failures (Mezulis, Abramson, Hyde, & Hankin, 2004). Likewise, people's attributions for external events are influenced by their political ideologies (Skitka, Mullen, Griffin, Hutchinson, & Chamberlin, 2002).

Recent empirical research supports the notion that directional goals can unconsciously guide perception. In a series of studies, Balcetis and Dunning (2006) showed participants an ambiguous figure that could be readily perceived as either of two different stimuli (e.g., the letter "B" or the number "13"). Depending on which stimulus they perceived, participants were assigned either to drink orange juice or a foul-smelling beverage. For those told that a letter would assign them to the orange juice condition, 72% saw the letter B. For those told that a number would assign them to the orange juice, 61% saw the number 13. Using an array of methods, follow-up studies showed that these results were not due to selective reporting but rather that motivation had a genuine unconscious effect on perception. In additional research on "wishful seeing," Balcetis and Dunning (2010) found that people judged objects that they want as physically closer than more neutral objects (e.g., participants who were thirsty compared to those who were quenched estimated that a bottle of water across a table was closer to them).

Perceptions of form and distance are not limitlessly malleable, even among people who are highly motivated. As Kunda (1990) noted, "people do not seem to be at liberty to conclude whatever they want to conclude merely because they want to" (p. 482). To some extent, reality constrains perception. Evidence in favor of one's biased judgment must be sufficient to allow for the construction of that judgment; a desired outcome cannot be rationalized in the face of irrefutable evidence to the contrary. This is precisely why ambiguous stimuli prove particularly susceptible to confirmation biases. It is also why many forensic judgments are subject to bias.

4. The forensic confirmation bias

Nearly 40 years ago, Tversky and Kahneman (1974) reasoned that confirmation bias effects could extend to the legal system insofar as "beliefs concerning the likelihood of... the guilt of a defendant" could impact judicial decision-making (p. 1124). They further speculated that the operation of such biases would affect not only the layperson but also experienced professionals. These statements proved quite prescient. Empirical and anecdotal evidence now suggests that pre-judgment expectations can indeed influence interrogators (Hill et al., 2008; Kassin, Goldstein, & Savitsky, 2003; Narchet et al., 2011), jurors (Charman et al., 2009; Lange et al., 2011), judges (Halverson, Hallahan, Hart, & Rosenthal, 1997), eyewitnesses (Hasel & Kassin, 2009), and experts in a range of forensic domains (e.g., see Dror & Cole, 2010; Dror & Hampikian, 2011).

Thus, we use the term *forensic confirmation bias* to summarize the class of effects through which an individual's preexisting beliefs, expectations, motives, and situational context influence the collection, perception, and interpretation of evidence during the course of a criminal case. As Findley and Scott (2006) have noted, the pernicious result produces a form of "tunnel vision"—a rigid focus on one suspect that leads investigators to seek out and favor inculpatory evidence, while overlooking or discounting any exculpatory evidence that might exist. A growing body of literature has begun to identify the ways in which such biases can pervade the investigative and judicial processes.

4.1. Context effects on forensic judgments

In an 1894 treatise on distinguishing genuine from forged signatures, William Hagan wrote: "There must be no hypothesis at the commencement, and the examiner must depend wholly on what is seen, leaving out of consideration all suggestions or hints from interested parties... Where the expert has no knowledge of the moral evidence or aspects of the case... there is nothing to mislead him" (p. 82). With this statement, Hagan was among the first scholars to acknowledge the potential biasing effect of expectation and context on perceptual judgments made by forensic examiners. It was not until recently, however, that empirical data emerged to support Hagan's admonition.

A growing body of work now suggests that confessions, a highly potent form of incrimination (Kassin, 1997; Kassin et al., 2010)—and other strong contextual cues—may bias forensic judgments in the criminal justice system, producing an effect that Kassin (2012) has called "corroboration inflation." Saks et al. (2003) note that the resulting non-independence among items of evidence can create an "investigative echo chamber" in which certain items reverberate and seem stronger and more numerous than they really are. Simon (2011) notes that coherence-based reasoning promotes

false corroboration among different witnesses, resulting in trials that are limited in their diagnostic value. Dror (2012) notes that the overall effect on judgments can increase as a result, creating a "bias snowball effect."

To our knowledge, the first study to examine this effect was by Miller (1984), who explored the impact of contextual information on the judgments of 12 college students trained to identify forged signatures. Miller found that participants who were exposed to additional inculpatory evidence formed a belief in the suspect's guilt, which skewed their perceptions. More recent work builds upon this finding. Kukucka and Kassin (2012) found that knowledge of a recanted confession can taint evaluations of handwriting evidence. In this study, lay participants read a bank robbery case in which the perpetrator gave a handwritten note to a bank teller. Soon afterward, they were told that a suspect was apprehended and interrogated, at which point he gave a handwritten *Miranda* waiver. Participants were asked to compare the handwriting samples taken from the perpetrator (bank note) and the defendant (Miranda waiver). When told that the defendant had confessed—even though he later retracted his confession, claiming it was coerced—participants perceived the handwriting samples as more similar and were more likely to conclude, erroneously, that they were authored by the same individual.

Other research indicates that interpretations of polygraph tests may also be shaped by preexisting beliefs. Elaad, Ginton, and Ben-Shakhar (1994) noted two ways in which expectations can impact the outcome of a polygraph test: By influencing the way examiners conduct their interviews and the questions they ask, and by influencing the conclusions they draw from the test results. To test the latter hypothesis, these investigators asked ten polygraph examiners from the Israeli Police to analyze 14 records from polygraph examinations of criminal suspects, all of whom had been judged inconclusive by independent raters. Each chart was accompanied by biasing information—for half of the charts, examiners were told that the interviewee had later confessed; for the remaining half, they were told that someone else had later confessed. Although most charts were judged inconclusive in the absence of biasing information, the charts were more likely to be scored as deceptive in the suspect-confession condition and as truthful in the other-confession condition. This effect was obtained with both experienced and inexperienced examiners—but not when the charts were conclusive. Thus, the conclusions drawn from ambiguous polygraph results were influenced by prior expectations.

Additional studies suggest that even fingerprint judgments may be subject to bias. In one study, Dror, Charlton, and Peron (2006) asked five experienced fingerprint experts to assess pairs of fingerprints that, unbeknownst to them, they had examined years earlier and declared to be a match. Before the stimuli were re-presented, these examiners were told that the fingerprints were taken from a high-profile case of erroneous identification, implying that they were not a match. Given this biasing information, only one of the five experts judged the fingerprints to be a match, indicating that context undermined reliability. This study is particularly troubling because the change as a function of context was obtained among experienced examiners, in a highly trusted forensic science, and in a within-subject experimental design.

In a followup study, Dror and Charlton (2006) presented six latent fingerprint experts with eight pairs of prints from a crime scene and suspect in an actual case in which they had previously made a match or exclusion judgment. The participants did not know they were taking part in a study, believing instead that they were conducting routine casework. The prints were accompanied either by no extraneous information, information that the suspect had confessed, suggesting a match; or information that the suspect was in custody at the time, suggesting exclusion. The results showed that contextual information in the custody condition produced an overall change in 17% of the originally correct match decisions.

Based on a meta-analysis of these two studies, Dror and Rosenthal (2008) estimated that the reliability of fingerprint experts' judgments over time likely falls in the range of 0.33–0.80, implying a considerable degree of subjectivity. Similarly, effect size estimates of biasability were 0.45 and 0.41, respectively, for the two studies. These findings are likely to extend to other forensic science domains that are based on visual similarity judgments, such as firearms; microscopic hair and fiber analysis; bite marks; impression evidence involving shoeprints, bite marks, tire tracks, and handwriting; and bloodstain pattern analysis (Dror & Cole, 2010).

Additional research suggests that confessions can also influence the testimony of lay witnesses. Looking at the possible effects of confession on eyewitnesses themselves, Hasel and Kassin (2009) staged a theft and took photographic identification decisions from eyewitnesses who viewed a culprit-absent lineup. Two days later, individual witnesses were told that the person they had identified denied guilt during a subsequent interrogation, or that he confessed, or that a specific other lineup member confessed. Among those who had made a selection but were told that another lineup member confessed, 61% changed their identifications—and did so with confidence. Among those who had correctly not made an initial identification, 50% went on to select the confessor.

The biasing effect of confessions can have grave consequences. The criminal justice system presupposes that suspects, eyewitnesses, forensic experts, and others offer information that is independent—not subject to taint from outside influences. But does this presupposition describe the reality of criminal investigation? Both basic psychology and forensic psychology research suggest otherwise—and, in particular, suggest the possibility that confessions can corrupt other evidence. To determine if this phenomenon might occur in actual cases, Kassin, Bogart, and Kerner (2012) conducted an archival analysis of DNA exonerations from the Innocence Project case files. Testing the hypothesis that confessions may prompt additional evidentiary errors, they examined whether other contributing factors were present in DNA exoneration cases containing a false confession. They found that additional errors were present in 78% of these cases. In order of frequency, false confessions were accompanied by invalid or improper forensic science (63%), mistaken eyewitness identifications (29%) and snitches or informants (19%). Consistent with the causal hypothesis that the false confessions had influenced the subsequent errors, the confession was obtained first rather than later in the investigation in 65% of these cases.

As a result of improprieties in U.S. laboratories, the frequency with which forensic science errors have surfaced in wrongful convictions, and the scathing critique from the National Academy of Sciences (2009)—which concluded that there are problems with standardization, reliability, accuracy and error, and the potential for contextual bias—it is not surprising that the most common means of corroboration for false confessions comes from bad forensic science (http://www.innocenceproject.org/). When coupled with recent laboratory studies, this presence of numerous forensic errors in Innocence Project confession cases suggests that confession evidence constitutes the kind of contextual bias that can skew expert judgments in many domains.

Confession is not the only form of evidence that can bias people's judgments. Mistaken eyewitness identifications constitute the most common contributing factor in DNA exoneration cases (Brewer & Wells, 2011; Wells et al., 1998). In fact, many Innocence Project cases contained two or more mistaken eyewitnesses who expressed high levels of certainty in their identifications. In some instances, these multiple errors can occur independently—as in the highly publicized mistaken identification of Ronald Cotton by

Jennifer Thompson, where Cotton physically resembled the perpetrator (Thompson-Cannino, Cotton, & Torneo, 2009). In other instances, however, the eyewitnesses may have influenced one another, a phenomenon demonstrated in numerous co-witness experiments (Gabbert, Memon, & Allan, 2003; Skagerberg, 2007). To further complicate matters, eyewitnesses who have been tainted by extrinsic information cannot accurately estimate the extent of the influence, suggesting that self-reports cannot be used to diagnose the corruption once it occurs (Charman & Wells, 2008).

4.2. Elasticity of forensic evidence

It is not surprising that expectations can taint questioned document examination (QDE), the discipline pertaining to documents, the authenticity or source of which are in dispute. QDE has been criticized for being a subjective domain of forensic science (Miller, 1984; Risinger et al., 2002; Risinger & Saks, 1996; *U.S. v. Hines, 1999*). In accordance with the research described earlier, examiners are more likely to exhibit bias when evaluating evidence that is ambiguous. This is consistent with Ask, Rebelius, and Granhag's (2008) assertion that some types of evidence are more "elastic"—i.e., more vulnerable to extraneous influence—than others.

Not all evidence is equally malleable or subject to confirmation bias. Paralleling classic research indicating that expectations can color judgments of stimuli that are ambiguous but not those that compel a particular perception, forensic research indicates that ambiguity is a moderating condition. Asked to make an identification decision on the basis of a memory trace that cannot be recovered for a side-by-side comparison to a stimulus face, eyewitnesses are particularly malleable when informed of a confession (Hasel & Kassin, 2009). Prior expectations can also bias interpretations of sensory stimuli such as auditory speech—but only when the recordings are degraded in quality and the stimuli are phonologically ambiguous, such as the words *gum* and *gun* or *ripped* and *raped* (Lange et al., 2011). The same is true of the judgments of polygraph examiners—again, when the physiological test data are ambiguous but not when the charts are strongly indicative of truth or deception (Elaad, Ginton, & Ben-Shakhar, 1994).

Still, within the forensic domains critiqued by the National Academy of Sciences (2009), the potential for bias is greater than previously imagined. In "The vision in 'blind' justice," Dror and Cole (2010) noted that many forensic judgments involve matching a visual pattern left at a crime scene with a sample taken from a suspect (e.g., shoe prints, tool marks, bite marks, tire marks, handwriting). The prototype is fingerprint identification, a forensic science long considered near-perfect (Cole, 2001). No two fingerprint impressions are totally identical because of variations in skin elasticity, the amount of pressure applied, the material on which the print was left, how the prints were recovered and other variables. And in criminal cases, where prints are lifted from crime scenes, many such latent fingerprints are partial and distorted. Hence, an impressive body of research now indicates that the judgments made by latent fingerprint experts are sensitive to biasing contextual information (Charlton, Fraser-Mackenzie, & Dror, 2010; Dror & Charlton, 2006; Dror, Charlton, & Peron, 2006; Dror, Peron, Hind, & Charlton, 2005).

Even when it comes to DNA testing—commonly considered the "gold standard" of forensic evidence (Lieberman, Carrell, Miethe, & Krauss, 2008; Lynch, 2003; Saks & Koehler, 2005)—the interpretation of certain complex DNA mixtures requires judgment that is subject to bias. To illustrate the risk, Dror and Hampikian (2011) described an actual gang rape case in which one of the assailants had accepted a plea bargain in exchange for testimony against other suspects. In order for the testimony of the cooperating assailant to be admissible, evidence was needed to corroborate his identifications. Aware of the situation, expert DNA analysts were asked to analyze the complex DNA mixture, and they concluded that the forensic evidence implicated those identified in the plea bargain. However, one of the alleged assailants repeatedly denied any involvement in the rape. To test the potential for contextual bias, Dror and Hampikian later took the same sample from this case and presented it, devoid of the biasing contextual information, to 17 neutral DNA analysts. Only one agreed with the original analysts; four deemed the sample inconclusive; 12 concluded that the DNA excluded the suspect in question. Despite the claim that DNA evidence is "inelastic" (e.g., Ask et al., 2008), it thus appears that confirmation biases may influence even the work of DNA analysts.

4.3. Bias and self-insight

Although confirmation bias typically operates outside of conscious awareness, forensic examiners may have some insight into the cognitive, motivational, and emotional factors that guide their job performance. Charlton et al. (2010) conducted semi-structured interviews of 13 experienced fingerprint examiners and identified a number of recurrent themes in their experiences. While describing their methodology in an objective manner, examiners expressed a personal interest in catching criminals and solving crimes, which some reported as more pronounced in serious and high-profile cases. They also expressed a strong need for closure, indicating a desire to provide definitive conclusions as a result of their work, and the feeling of joy that accompanies the discovery of a fingerprint match. At the same time, these experts consistently expressed a fear of making erroneous judgments, and in particular, a fear of committing a false-positive error that would implicate an innocent person. Thus, perhaps some experts deliberately endeavor to be conservative in their judgments to avoid such errors.

Furthermore, mere awareness of the type of crime being investigated may not be sufficient to bias fingerprint expert judgments. Utilizing an experimental paradigm, Hall and Player (2008) asked experienced examiners to judge pairs of fingerprints either in the context of a forgery case or a murder case—but no emotionally-arousing crime scene photos were included. Results indicated that examiners in the murder condition were more likely to self-report feeling influenced by this context, but the type of case had no overall effect on their conclusions. Perhaps for context to influence judgments, participants must really believe it. In short, there may be a dissociation between forensic examiners' insight into their own biases and the actual manifestation of bias in their actual judgments; they can be biased and unaware, or they can be relatively objective despite the self-perception of bias.

4.4. Null effects from the Netherlands

It is important to note that two additional studies from a single lab failed to replicate confirmation bias effects on forensic experts. First, Kerstholt, Paashuis, and Sjerps (2007) recruited twelve Dutch officers trained in forensic shoe print examinations and asked them to evaluate eight pairs of shoes and prints. Each pair was presented in the context of a fictional criminal investigation, which either did or did not contain biasing information to suggest that the shoe had created the print. This manipulation had no effect on evaluations. Similarly, Kerstholt et al. (2010) had six Dutch firearms examiners judge six pairs of bullets that were presented twice, several months apart. Each pair of bullets was presented twice—once with, and once without, a biasing case description—to be categorized as a match, a non-match, or inconclusive. Overall, 10 out of 36 judgments of the same pair of bullets changed from one presentation to the next, indicating a problem with intra-examiner reliability and subjectivity. However, the bias manipulation did not have a significant effect on judgments.

To account for these replication failures, Kerstholt et al. (2010) note that each of these forensic sciences utilizes a highly-standardized procedure in the Netherlands (e.g., shoe print examiners follow a protocol whereby they assign numerical values to various features of the shoe print and then sum the values to obtain a total score). Perhaps examiners in these studies were not biased by expectations precisely because they are trained in evaluation procedures that are well-defined. Alternatively, perhaps the biasing information used in these studies failed to create a strong expectation of guilt. As far as we can tell, the expectation manipulations had not been pilot tested nor was their effectiveness confirmed through manipulation checks. In the study of shoe print examinations, one of the fictional cases described the burglary of an electronics store. To raise the expectation that the print matched the suspect's shoe, examiners were told that the suspect had been found selling electronics on the street and that he owned a van (which would presumably be needed to transport large electronics). Each of these facts arguably constitutes a necessary but not sufficient condition to imply guilt, and thus may not have cultivated an a priori belief that the shoe would match the print. As in Hall and Player (2008), these studies also involved examiners who knew they were taking part in a study—not involved in actual criminal casework (Dror, 2009b).

4.5. Forensic confirmation bias in actual cases

The biases set into motion by confessions and other guilt-presumptive sources of information are not without consequence. A growing number of real-world wrongful convictions, as seen in the opening story about the Madrid bombing case in which Brandon Mayfield was misidentified and as reported in many cases from the Innocence Project, provide ample real world instantiation of this hypothesis.

In one case, in Pennsylvania, suspect Barry Laughman was induced to confess during an unrecorded interrogation to the rape and murder of his elderly neighbor. The next day, serology tests showed that Laughman had Type B blood; yet the semen recovered from the victim was from a Type A secretor. Aware that Laughman had confessed, the state forensic chemist went on to propose four "novel" theories, none grounded in science, to dismiss the mismatch. On the basis of his confession, Laughman was convicted. Sixteen years later, he was exonerated by DNA and set free (http://www.innocenceproject.org/Content/Barry_Laughman).

Another example can be found in the 2004 trial of Mississippi v. Tyler Edmonds. In that case, 13 year-old Edmonds was induced to confess that he had physically assisted his older half-sister in the shooting and killing of her husband. Supporting what had become disputed confession, the state's medical pathologist who conducted the autopsy of the victim's body and submitted his report after the confession was taken testified without any basis in science that the gunshot wound suggested a bullet fired by two persons pulling the trigger simultaneously. Edmonds was convicted at trial and sentenced to life in prison. Highly critical of this expert's "speculative" and "scientifically unfounded" opinion, the state Supreme Court overturned the conviction (*Tyler Edmonds v. State of Mississippi, 2007*). The following year, Edmonds was retried and acquitted. After an investigation by the state's medical board, the pathologist in question was removed from the state's designated list of pathologists.

4.6. Implications for accuracy and error

The fact that confessions and other strong bases for a presumption of guilt can bias the search, collection, perception, and interpretation of subsequently obtained evidence undermines a silent but basic tenet of the judicial system—namely, that the items of evidence presented at trial are independent of one another. When one witness influences another, then a strong bias is created, creating what Kassin (2012) described as "corroboration inflation" and a gathering momentum for more and more bias, or what Dror (2012) referred to as a "bias snowball effect." The influence of one witness or item of evidence on another witness or item of evidence constitutes a biasing process of confirmation, one that can increase the likelihood of error. In the Texas arson-murder case against Cameron Todd Willingham, for example, eyewitnesses changed their account once told about forensic evidence suggesting that the fire was not accidental. Although this forensic conclusion was later found to be erroneous, Willingham was found guilty and executed (Grann, 2009).

Just as forensic science is subject to bias, so too are suspects pressed for confession and eyewitnesses pressed for identification. Many of the studies described above focused on how confessions can spawn other incriminating evidence. This influence can be bidirectional; just as confessions can taint other evidence, other evidence can taint confessions as well. Indeed, numerous studies and case anecdotes support the fact that innocent people can be induced to confess by the true or false presentation of an eyewitness, physical evidence, failed polygraph, or other incriminating evidence (e.g., Gudjonsson, 2003; Kassin, 1997; Kassin & Gudjonsson, 2004; Kassin & Kiechel, 1996; Nash & Wade, 2009; Perillo & Kassin, 2011). In one case, for example, Dwayne Jackson confessed to a crime he did not commit after he was erroneously identified in DNA testing by Las Vegas forensic examiners (Mower & McMurdo, 2011).

Forensic examiners are aware of and trained to avoid physical contamination in an effort to protect the integrity of the evidence. However, "psychological contamination" has not received similar attention and is prevalent throughout the criminal justice system. The sources of psychological contamination are numerous (e.g., knowing the context of the crime, police pressure to influence a forensic evaluation, information about a prior confession or eyewitness identification). Biasing context can take on other subtle forms as well. For example, forensic examiners work with a variety of technologies—including computerized systems that suggest a list of candidates for the human examiner to consider. In a recent study, Dror, Wertheim, Fraser-Mackenzie, and Walajtys (2012) independently varied the order of the candidates on the list and found that examiners spent less time on the same candidate when it was placed further down the list. Examining 55,200 forensic decisions, these investigators also found that examiners are more likely to make false positive errors on candidates on the top of the list and false negative errors on those near the bottom. This result illustrates how meta-data provided by computerized systems can also bias forensic examiners.

5. How to reduce bias: proposed reforms

As detailed earlier, forensic confirmation biases may be particularly problematic in the forensic sciences—where stimulus ambiguity, context-driven expectations, and motivations conspire to create fertile conditions for psychological contamination and bias to operate. There are two levels at which it is necessary to reduce this bias and its consequences: The first level is in the forensic laboratory, and even at the crime scene, where evidence is collected and sometimes analyzed; the second level is at the trial and appellate courts, where that evidence is evaluated. Hence, we offer a number of suggestions.

5.1. Reducing bias in the crime laboratory

In a study of four crime laboratories, Peterson, Mihajlovic, and Gilliland (1984) discovered that very few reports excluded the

known suspect from the crime scene or from a connection to the victim. It is not clear whether this result indicates that police manage to identify actual perpetrators for suspicion at high levels of accuracy or that forensic examiners have strong and biasing base-rate expectations that lab results will prove incriminating. As a result of numerous DNA exonerations since that time, however, it is clear that the forensic sciences have contributed to wrongful convictions (Hampikian et al., 2011)—especially in cases that featured other flawed evidence, most notably mistaken eyewitness identifications and false confessions that chronologically preceded the forensic errors (Kassin et al., 2012). To minimize the problem, we suggest the following:

- Examiners should work "linear" rather than "circular," thus initially examining the evidence from the crime scene and documenting their findings before making comparisons against a target. This will eliminate the potential influence of the target on how information is processed and the weight assigned to it (Dror, 2009a).

  It is conceivable that forensic examiners sometimes "re-assess" the evidence to fit the target. If the initial assessment is done in isolation of the target, then such potential influences are eliminated. Indeed, the FBI recently revised its Standard Operating Procedures (SOPs) to "include some steps to avoid bias: examiners must complete and document analysis of the latent fingerprint before looking at any known fingerprint" and to "instruct examiners conducting analysis of a latent fingerprint to analyze it for evidence of distortion, determine whether it is 'of value,' and document the data used during analysis" (OIG, 2011, p. 27).

  Initial analysis in isolation lacks the direction guided by the comparison to a target. For example, when examining a latent print from a crime scene, it may be hard to know where to look for minutia—the important characteristics in a print. Having a suspect's print can guide the examiner as to where such characteristics may be found on the latent print. It is therefore suggested that examiners be allowed to revisit the analysis stage but document their inquiry, justify it, and limit it (for example, for features that were inconclusive during the initial analysis). Although this revisit may open the door to some bias, we believe it is important to use reasonable procedures that both balance the need to avoid bias but facilitate examiners in doing their work. The Office of the Inspector General (OIG, 2011) supports this cognitively informed approach in its report: "a solution to bias may be requiring initial analysis of the latent fingerprint in isolation from the known fingerprints, but also permitting, with clear and detailed documentation, some 're-analysis' of the latent print after comparison" (p. 28). A recent Expert Group set up by the National Institute of Standards and Technology made a similar recommendation (NIST, 2012, Recommendation 3.2).
- The simplest way to protect against the biasing effects of contextual variables is to conduct blind testing. Too often, examiners are exposed to extraneous information from various sources that may taint their conclusions. It is important to shield them from this information. There is no reason why examiners should receive information that is *not* relevant to their work and that they do not need. Thus, we recommend, as much as possible, that forensic examiners be isolated from undue influences such as direct contact with the investigating officer, the victims and their families, and other irrelevant information—such as whether the suspect had confessed.
- Blind testing can shield the forensic examiner from a confession, eyewitness identification, and other information about an investigation that is irrelevant to their forensic work. But it does not protect against the simple base-rate assumption that any individual identified as a suspect is the likely perpetrator. In current forensic practice, examiners often compare a sample of material to that of a target, presumably belonging to the suspect, in an effort to determine if the two samples derive from the same individual. This protocol is structurally identical to the eyewitness "showup" in which a witness is asked to make a memory-based identification decision via exposure to a single individual. Research shows that showups result in more false positive errors when the suspect and comparison are generally similar to one another (Steblay, Dysart, Fulero, & Lindsay, 2003). Modeled after the extensive scientific literature on best way to collect eyewitness identifications (Wells et al., 1998), which forms the basis for a set of best practice guidelines adopted by U.S. Department of Justice (Technical Working Group for Eyewitness Evidence, 1999), we agree with Saks et al. (2003) in proposing, when possible, the use of an *evidence lineup.*

  Modeled after the practice of administering a photograph eyewitness lineup, often called a "six pack," we would recommend that a target-blind examiner be presented with six samples—one belonging to the suspect and five plausible fillers (for the importance of having lineup identifications conducted by a blind administrator, see Canter, Hammond, & Youngs, 2012). From that array, he or she would then seek to determine which, if any, constitutes a match to the evidence found at the crime scene or on the victim. In the only test of the effects of an evidence lineup, Miller (1987) presented students trained in human hair identification with hair samples recovered from a crime scene, which they compared against either a singular innocent suspect sample or a "target-absent lineup" of five innocent samples. Results indicated that the use of a lineup produced a significantly lower error rate than the traditional method (3.8% vs. 30.4%, respectively). Given that none of the samples presented was a true match, all of the errors committed were false positives.
- The *verification* of forensic decisions should be a more controlled process in which blind and double-blind procedures are used whenever possible. Such procedures would require that the verifier is not informed of the initial conclusion; if possible, that the verifier does not know who the examiner was; and that the examiner does not select the verifier (a common practice in many laboratories). Cross-laboratory verifications are also advisable to provide an independent means of checking on the propriety of the initial forensic work (Koppl, Kurzban, & Kobilinsky, 2008).
- Technology plays an increasing and effective tool in solving crimes, enabling the speedy examination of large databases. As noted earlier, however, such technology that examines millions of potential suspects can also lead to error because the likelihood of finding incidental close non-matches is increased (Dror & Mnookin, 2010). This technology can also unwittingly provide meta-data, such as a ranking of potential candidates, which can bias expectations and cause examiners to miss matches or make incorrect identifications (Dror et al., 2012).

  To minimize this problem, careful consideration should be given to deploying these technologies. When a list of potential candidates is provided, that list should be reasonable in length and the order of entries should be randomized as a way to keep examiners from developing a strategy that considers candidates according to their ordinal position on the list. This simple safeguard will enable human examiners to evaluate each candidate fully, equally, and without bias.
- Finally, we believe that it would be useful for forensic science education and certification to include training in basic psychology that is relevant to forensic

work—for example, psychology coursework that addresses experimental methods as well as aspects of perception, judgment and decision making, and social influence, all illustrated through the use of forensic case materials.

5.2. Reducing bias in the courts

The forensic confirmation bias spawns three problems. The first is that it can corrupt the conclusions and testimony of forensic examiners. The second problem is that these conclusions, once corrupted, can have grave consequences—influencing other lines of evidence, be it other forensic examiners, eyewitnesses, and even inducing false confessions among the suspects themselves. The third problem is that these biased sources of information are presented to judges, juries, and appeals courts, which heavily rely on forensic science evidence in their decision-making.

To address these problems, we believe it is important that legal decision makers be educated with regard to the procedures by which forensic examiners reached their conclusions and the information that was available to them at that time. In particular, both trial and appellate courts should be trained to ask "What did the examiner know and when did he or she know it?" and probe routinely for the possibility of contamination across items of evidence that are allegedly independent and corroborative. In cases in which a forensic examiner was unduly exposed and possibly biased by extraneous information, such forensic evidence should be subject to a pretrial reliability hearing aimed at determining if the judgment was tainted and should be excluded rather than admitted into evidence.

At the trial level, judges and juries need to know that forensic science conclusions that appear to corroborate a confession or eyewitness identification may, in fact, have been influenced by these previously collected forms of evidence. This problem has relevance at the appellate level as well. In the U.S., appeals courts may determine that flawed evidence (e.g., a coerced confession or suggestive eyewitness identification) was erroneously admitted at trial but that this trial error was "harmless" (the implication of which is to affirm a defendant's conviction) based on an assessment whether that error had contributed to the jury's verdict in light of all of the evidence presented (for a history of the harmless error rule, see Bilaisis, 1983; as applied to confessions, see *Arizona v. Fulminante, 1991*; Kassin, 2012).

This harmless error doctrine—that an erroneously admitted confession can prove harmless if other evidence is sufficient to support conviction—rests on the tacit and often incorrect assumption that the alleged other evidence was independent of the erroneously admitted item, say, a coerced confession. Indeed, according to Garrett (2011), appellate courts that conducted post-conviction reviews of several confessors who were later exonerated had affirmed the convictions by citing the "overwhelming nature of the evidence against them" (p. 1107). In light of classic psychology research on perceptual and cognitive confirmation biases and the more recent studies of psychological contamination of forensic evidence, we now believe that the courts must consider the proposition on a case-by-case basis that the erroneous evidence presented at trial had corrupted the very forensic examinations that were used to make the error appear harmless.

Going forward, therefore, we believe that the research reviewed in this article has far reaching implications not only for how forensic examinations are conducted but for how the evidence, once gathered, is later presented and evaluated in the courts. It is clear that forensic science evidence often involves subjective judgments that may be biased in a variety of ways. Such influences are psychological in nature, and therefore an area ripe for further empirical research. This research will not only enhance forensic work and the administration of justice but also provide insights and a testing ground for psychological theory.

Acknowledgements

This work was supported in part by a research grant by the National Institute of Standards and Technology (NIST), Federal Bureau of Investigation (FBI), and Department of Defense (DoD/CTTSO/TSWG), #N41756-10-C-3382, awarded to Itiel Dror; and in part by funds provided by the Research Foundation of the City University of New York to Saul Kassin.

References

Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology*, *39*, 1037–1049.

Arizona v. Fulminante, 499 U.S. 279 (1991).

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, *41*, 258–290.

Ask, K., Rebelius, A., & Granhag, P. A. (2008). The elasticity of criminal evidence: A moderator of investigator bias. *Applied Cognitive Psychology*, *22*, 1245–1259. http://dx.doi.org/10.1002/acp.1432

Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology*, *91*, 612–625. http://dx.doi.org/10.1037/0022-3514.91.4.612

Balcetis, E., & Dunning, D. (2010). Wishful seeing: More desired objects are seen as close. *Psychological Science*, *21*, 147–152.

Baumann, A. O., Deber, R. B., & Thompson, G. G. (1991). Overconfidence among physicians and nurses: The micro-certainty, macro-uncertainty phenomenon. *Social Science and Medicine*, *32*, 167–174.

Bilaisis, V. (1983). Harmless error: Abettor of courtroom misconduct. *Journal of Criminal Law & Criminology*, *74*, 457–475.

Boring, E. G. (1930). A new ambiguous figure. *The American Journal of Psychology*, *42*, 444–445. http://dx.doi.org/10.2307/1415447

Bressan, P., & Dal Martello, M. F. (2002). 'Talis pater, talis filius': Perceived resemblance and the belief in genetic relatedness. *Psychological Science*, *13*, 213–218. http://dx.doi.org/10.1111/1467-9280.00440

Brewer, N., & Wells, G. L. (2011). Eyewitness identification. *Current Directions in Psychological Science*, *20*, 24–27.

Bruner, J. S., & Goodman, C. C. (1947). Value and need as organizing factors in perception. *The Journal of Abnormal and Social Psychology*, *42*, 33–44. http://dx.doi.org/10.1037/h0058484

Bruner, J. S., & Potter, M. C. (1964). Interference in visual recognition. *Science*, *144*, 424–425. http://dx.doi.org/10.1126/science.144.3617.424

Canter, D., Hammond, L., & Youngs, D. (2012). Cognitive bias in line-up identifications: The impact of administrator knowledge. *Science and Justice*, http://dx.doi.org/10.1016/j.scijus.2012.12.001

Charlton, D., Fraser-Mackenzie, P. A. F., & Dror, I. E. (2010). Emotional experiences and motivating factors associated with fingerprint analysis. *Journal of Forensic Sciences*, *55*, 385–393. http://dx.doi.org/10.1111/j. 1556-4029.2009.01295.x

Charman, S. D., Gregory, A. H., & Carlucci, M. (2009). Exploring the diagnostic utility of facial composites: Beliefs of guilt can bias perceived similarity between composite and suspect. *Journal of Experimental Psychology: Applied*, *15*, 76–90. http://dx.doi.org/10.1037/a0014682

Charman, S. D., & Wells, G. L. (2008). Can eyewitnesses correct for external influences on their lineup identifications? The actual/counterfactual assessment paradigm. *Journal of Experimental Psychology: Applied*, *14*, 5–20.

Cole, S. A. (2001). *Suspect identities: A history of fingerprinting and criminal identification*. Cambridge, MA: Harvard University Press.

Darley, J. M., & Fazio, R. H. (1980). Expectancy confirmation processes arising in the social interaction sequence. *American Psychologist*, *35*, 867–881. http://dx.doi.org/10.1037/0003-066X.35.10.867

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*, 20–33. http://dx.doi.org/10.1037/0022-3514.44.1.20

Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579. (1993).

Dror, I. E. (2009a). How can Francis Bacon help forensic science? The four idols of human biases. *Jurimetrics: The Journal of Law, Science, and Technology*, *50*, 93–110.

Dror, I. E. (2009b). On proper research and understanding of the interplay between bias and decision outcomes. *Forensic Science International*, *191*, 17–18.

Dror, I. E. (2012). Cognitive bias in forensic science. In *The 2012 yearbook of science & technology*. New York: McGraw-Hill., pp. 43–45.

Dror, I. E., Champod, C., Langenburg, G., Charlton, D., Hunt, H., & Rosenthal, R. (2011). Cognitive issues in fingerprint analysis: Inter-and intra-expert consistency and the effect of a 'target' comparison. *Forensic Science International*, *208*, 10–17. http://dx.doi.org/10.1016/j.forsciint.2010.10.013

Dror, I. E., & Charlton, D. (2006). Why experts make errors. *Journal of Forensic Identification*, *56*, 600–616.

Dror, I. E., Charlton, D., & Peron, A. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International*, *156*, 174–178. http://dx.doi.org/10.1016/j.forsciint.2005.10.017

Dror, I. E., & Cole, S. A. (2010). The vision in blind justice: Expert perception, judgment, and visual cognition in forensic pattern recognition. *Psychonomic Bulletin & Review*, *17*, 161–167. http://dx.doi.org/10.3758/PBR.17.2.161

Dror, I. E., & Hampikian, G. (2011). Subjectivity and bias in forensic DNA mixture interpretation. *Science & Justice*, *51*, 204–208. http://dx.doi.org/10.1016/j.scijus.2011.08.004

Dror, I. E., & Mnookin, J. (2010). The use of technology in human expert domains: Challenges and risks arising from the use of automated fingerprint identification systems in forensics. *Law, Probability and Risk*, *9*(1), 47–67.

Dror, I. E., Peron, A. E., Hind, S.-L., & Charlton, D. (2005). When emotions get the better of us: The effect of contextual top-down processing on matching fingerprints. *Applied Cognitive Psychology*, *19*, 799–809. http://dx.doi.org/10.1002/acp.1130

Dror, I. E., & Rosenthal, R. (2008). Meta-analytically quantifying the reliability and biasability of forensic experts. *Journal of Forensic Sciences*, *53*, 900–903. http://dx.doi.org/10.1111/j. 1556-4029.2008.00762.x

Dror, I. E., Wertheim, K., Fraser-Mackenzie, P., & Walajtys, J. (2012). The impact of human-technology cooperation and distributed cognition in forensic science: Biasing effects of AFIS contextual information on human experts. *Journal of Forensic Sciences*, *57*(2), 343–352.

Edmonds v. Mississippi, 955 So.2d 787 (2007).

Elaad, E., Ginton, A., & Ben-Shakhar, G. (1994). The effects of prior expectations and outcome knowledge on polygraph examiners' decisions. *Journal of Behavioral Decision Making*, *7*, 279–292. http://dx.doi.org/10.1002/bdm.3960070405

Findley, K. A., & Scott, M. S. (2006). The multiple dimensions of tunnel vision in criminal cases. *Wisconsin Law Review*, *2*, 291–397.

Fisher, G. H. (1968). Ambiguity of form: Old and new. *Attention, Perception, & Psychophysics*, *4*, 189–192. http://dx.doi.org/10.3758/BF03210466

Gabbert, F., Memon, A., & Allan, K. (2003). Memory conformity: Can eyewitnesses influence each other's memories for an event? *Applied Cognitive Psychology*, *88*, 341–347.

Garrett, B. L. (2011). *Convicting the innocent: Where criminal prosecutions go wrong*. Cambridge, MA: Harvard University Press.

Grann, D. (2009). Trial by Fire: Did Texas execute an innocent man? *The New Yorker*, (September).

Gudjonsson, G. H. (2003). *The psychology of interrogations and confessions: A handbook*. Chichester: Wiley.

Hagan, W. E. (1894). *A treatise on disputed handwriting and the determination of genuine from forged signatures*. New York, NY: Banks & Brothers.

Hall, L. J., & Player, E. (2008). Will the introduction of an emotional context affect fingerprint analysis and decision-making? *Forensic Science International*, *181*, 36–39. http://dx.doi.org/10.1016/j.forsciint.2008.08.008

Halverson, A. M., Hallahan, M., Hart, A. J., & Rosenthal, R. (1997). Reducing the biasing effects of judges' nonverbal behavior with simplified jury instruction. *Journal of Applied Psychology*, *82*, 590–598. http://dx.doi.org/10.1037/0021-9010.82.4.590

Hamilton, D. L., & Zanna, M. P. (1974). Context effects in impression formation: Changes in connotative meaning. *Journal of Personality and Social Psychology*, *29*, 649–654.

Hampikian, G., West, E., & Akselrod, O. (2011). The genetics of innocence: Analysis of 194 U.S. DNA exonerations. *Annual Review of Genomics and Human Genetics*, *12*, 97–120.

Hasel, L. E., & Kassin, S. M. (2009). On the presumption of evidentiary independence: Can confessions corrupt eyewitness identifications? *Psychological Science*, *20*, 122–126. http://dx.doi.org/10.1111/j. 1467-9280.2008.02262.x

Hill, C., Memon, A., & McGeorge, P. (2008). The role of confirmation bias in suspect interviews: A systematic evaluation. *Legal and Criminological Psychology*, *13*, 357–371. http://dx.doi.org/10.1348/135532507X238682

Hochschild, J. L. (2008). How ideas affect actions. In R. E. Goodin, & C. Tilly (Eds.), *The Oxford Handbook of Contextual Political Analysis* (p. 284). Oxford University Press.

Jussim, L. (2012). *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy*. New York: Oxford University Press.

Kassin, S. M. (1997). The psychology of confession evidence. *American Psychologist*, *52*, 221–233.

Kassin, S. M. (2012). Why confessions trump innocence. *American Psychologist*, *67*, 431–445.

Kassin, S. M., Bogart, D., & Kerner, J. (2012). Confessions that corrupt: Evidence from the DNA exoneration case files. *Psychological Science*, *23*, 41–45.

Kassin, S. M., Drizin, S. A., Grisso, T., Gudjonsson, G. H., Leo, R. A., & Redlich, A. D. (2010). Police-induced confessions: Risk factors and recommendations. *Law and Human Behavior*, *34*, 3–38. http://dx.doi.org/10.1007/s10979-009-9188-6

Kassin, S. M., Goldstein, C. C., & Savitsky, K. (2003). Behavioral confirmation in the interrogation room: On the dangers of presuming guilt. *Law and Human Behavior*, *27*, 187–203. http://dx.doi.org/10.1023/A:1022599230598

Kassin, S. M., & Gudjonsson, G. H. (2004). The psychology of confessions: A review of the literature and issues. *Psychological Science in the Public Interest*, *5*, 33–67. http://dx.doi.org/10.1111/j.1529-1006.2004.00016.x

Kassin, S. M., & Kiechel, K. L. (1996). The social psychology of false confessions: Compliance, internalization, and confabulation. *Psychological Science*, *7*, 125–128.

Kerstholt, J., Eikelboom, A., Dijkman, T., Stoel, R., Hermsen, R., & van Leuven, B. (2010). Does suggestive information cause a confirmation bias in bullet comparisons? *Forensic Science International*, *198*, 138–142. http://dx.doi.org/10.1016/j.forsciint.2010.02.007

Kerstholt, J., Paashuis, R., & Sjerps, M. (2007). Shoe print examinations: Effects of expectation, complexity and experience. *Forensic Science International*, *165*, 30–34. http://dx.doi.org/10.1016/j.forsciint.2006.02.039

Kierein, N. M., & Gold, M. A. (2000). Pygmalion in work organizations: A meta-analysis. *Journal of Organizational Behavior*, *21*, 913–928.

Klayman, J., & Ha, Y.-W. (1997). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Bulletin*, *94*, 211–228. http://dx.doi.org/10.1037/0033-295X.94.2.211

Koppl, R., Kurzban, R., & Kobilinsky, L. (2008). Epistemics for forensics. *Episteme*, *5*(2), 141–159.

Kukucka, J., & Kassin, S. M. (2012, April). Do confessions taint juror perceptions of handwriting evidence? *Paper presented at the Annual Meeting of the American Psychology-Law Society* San Juan, Puerto Rico, March 14–17.

Kumho Tire Co. v. Carmichael, 526 U.S. 137. (1999).

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498. http://dx.doi.org/10.1037/0033-2909.108.3.480

Lange, N. D., Thomas, R. P., Dana, J., & Dawes, R. M. (2011). Contextual biases in the interpretation of auditory evidence. *Law and Human Behavior*, *35*, 178–187. http://dx.doi.org/10.1007/s10979-010-9226-4

Leadbetter, M. (2007). Letter to the Editor. *Fingerprint World*, *33*, 231.

Leeper, R. (1935). A study of a neglected portion of the field of learning: The development of sensory organization. *The Pedagogical Seminary and Journal of Genetic Psychology*, *46*, 41–75.

Lieberman, J. D., Carrell, C. A., Miethe, T. D., & Krauss, D. A. (2008). Gold versus platinum: Do jurors recognize the superiority and limitations of DNA evidence compared to other types of forensic evidence? *Psychology, Public Policy, & Law*, *14*, 27–62. http://dx.doi.org/10.1037/1076-8971.14.1.27

Lynch, M. (2003). God's signature: DNA profiling, the new gold standard in forensic evidence. *Endeavor*, *27*, 93–97. http://dx.doi.org/10.1016/S0160-9327(03)00068-1

McNatt, D. B. (2000). Ancient Pygmalion joins contemporary management: A meta-analysis of the result. *Journal of Applied Psychology*, *85*, 314–322.

Mezulis, A., Abramson, L., Hyde, J. S., & Hankin, B. L. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin*, *130*, 711–746.

Miller, L. S. (1984). Bias among forensic document examiners: A need for procedural changes. *Journal of Police Science and Administration*, *12*, 407–411.

Miller, L. S. (1987). Procedural bias in forensic science examinations of human hair. *Law and Human Behavior*, *11*, 157–163. http://dx.doi.org/10.1007/BF01040448

Mnookin, J. L., Cole, S. A., Dror, I. E., Fisher, B. A., Houck, M., Inman, K., et al. (2011). The need for a research culture in the forensic sciences. *UCLA Law Review*, *58*(3), 725–779.

Mower, L., & McMurdo, D. (2011, July). Las Vegas police reveal DNA error put wrong man in prison. *Las Vegas Review Journal*.

Narchet, F. M., Meissner, C. A., & Russano, M. B. (2011). Modeling the influence of investigator bias on the elicitation of true and false confessions. *Law and Human Behavior*, *35*, 452–465. http://dx.doi.org/10.1007/s10979-010-9257-x

Nash, R. A., & Wade, K. A. (2009). Innocent but proven guilty: Using false video evidence to elicit false confessions and create false beliefs. *Applied Cognitive Psychology*, *23*, 624–637.

National Academy of Sciences. (2009). *Strengthening Forensic Science in the United States: A Path Forward*. Washington, DC: National Academies Press.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220. http://dx.doi.org/10.1037/1089-2680.2.2.175

NIST. (2012). Expert Working Group on human factors in latent print analysis. *Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach*. U.S. Department of Commerce, National Institute of Standards and Technology. http://www.nist.gov/customcf/get_pdf.cfm?pub_id=910745

O'Brien, B. (2009). Prime suspect: An examination of factors that aggravate and counteract confirmation bias in criminal investigations. *Psychology, Public Policy and Law*, *15*, 315–334.

OIG. (2006). A review of the FBI's handling of the Brandon Mayfield case. Office of the Inspector General, Oversight & Review Division, US Department of Justice.

OIG. (2011). *A review of the FBI's progress in responding to the recommendations in the office of the inspector general report on the fingerprint misidentification in the Brandon Mayfield case*. <http://www.justice.gov/oig/special/s1105.pdf>.

Perillo, J. T., & Kassin, S. M. (2011). Inside interrogation: The lie, the bluff, and false confessions. *Law and Human Behavior*, *35*, 327–337. http://dx.doi.org/10.1007/s10979-010-9244-2

Peterson, J. L., Mihajlovic, S., & Gilliland, M. (1984). *Forensic evidence and the police*. Washington, DC: National Institute of Justice.

Radel, R., & Clement-Guillotin, C. (2012). Evidence of motivational influences in early visual perception: Hunger modulates conscious access. *Psychological Science*, *23*, 232–234.

Risinger, D. M., & Saks, M. J. (1996). Science and nonscience in the courts: Daubert meets handwriting identification expertise. *Iowa Law Review*, *82*, 21–74.

Risinger, D. M., Saks, M. J., Thompson, W. C., & Rosenthal, R. (2002). The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. *California Law Review*, *90*, 1–56.

Rosenthal, R. (2002). Covert communication in classrooms, clinics, courtrooms, and cubicles. *American Psychologist*, *57*, 839–849. http://dx.doi.org/10.1037/0003-066X.57.11.839

Rosenthal, R., & Fode, K. (1963). The effect of experimenter bias on performance of the albino rat. *Behavioral Science*, *8*, 183–189.

Rosenthal, R., & Jacobson, L. (1966). Teachers' expectancies: Determinants of pupils' IQ gains. *Psychological Reports*, *19*, 115–118. http://dx.doi.org/10.2466/pr0.1966.19.1.115

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science*, *309*, 892–895. http://dx.doi.org/10.1126/science.1111565

Saks, M. J., Risinger, D. M., Rosenthal, R., & Thompson, W. C. (2003). Context effects in forensic science: A review and application of the science of science to crime laboratory practice in the United States. *Science & Justice*, *43*, 77–90.

Simon, D. (2011). The limited diagnosticity of criminal trials. *Vanderbilt Law Review*, *64*, 143–223.

Skagerberg, E. M. (2007). Co-witness feedback in line-ups. *Applied Cognitive Psychology*, *21*, 489–497.

Skitka, L. J., Mullen, E., Griffin, T., Hutchinson, S., & Chamberlin, B. (2002). Dispositions, scripts, or motivated correction? Understanding ideological differences in explanations for social problems. *Journal of Personality and Social Psychology*, *83*, 470–487.

Snyder, M., & Swann, W. B., Jr. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology*, *36*, 1202–1212. http://dx.doi.org/10.1037/0022-3514.36.11.1202

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2003). Eyewitness accuracy rates in police showup and lineup presentations: A meta-analytic comparison. *Law and Human Behavior*, *27*, 523–540. http://dx.doi.org/10.1023/A:1025438223608

Tangen, J. M., Thompson, M. B., & McCarthy, D. J. (2011). Identifying fingerprint expertise. *Psychological Science*, *22*, 995–997. http://dx.doi.org/10.1177/0956797611414729

Technical Working Group for Eyewitness Evidence. (1999). *Eyewitness evidence: A guide for law enforcement [booklet]*. Washington, DC: United States Department of Justice, Office of Justice Programs. Document Number NCJ-178240.

Thompson-Cannino, J., Cotton, R., & Torneo, E. (2009). *Picking Cotton: Our memoir of injustice and redemption*. New York: St. Martin's Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131. http://dx.doi.org/10.1126/science.185.4157.1124

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Science of the United States of America*, *108*(19), 7733–7738.

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2012). Repeatability and reproducibility of decisions made by latent fingerprint examiners. *PLoS ONE*, *7*, 1–12. http://dx.doi.org/10.1371/journal.pone.0032800

U.S. v. Hines, 55 F. Supp. 2d 62. (1999).

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129–140. http://dx.doi.org/10.1080/17470216008416717

Watkins, M. J., & Peynircioglu, Z. F. (1984). Determining perceived meaning during impression formation: Another look at the meaning change hypothesis. *Journal of Personality and Social Psychology*, *46*, 1005–1016.

Wells, G. L., Small, M., Penrod, S., Malpass, R., Fulero, S., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law & Human Behavior*, *22*, 603–647.

DANIEL MURRIE

Daniel Murrie completed his PhD in clinical psychology at the University of Virginia. Dr. Murrie serves as Director of Psychology at the *Institute of Law, Psychiatry and Public Policy* (ILPPP), an Associate Professor in the Department of Psychiatry and Neurobehavioral Sciences at the *University of Virginia School of Medicine,* and an instructor in the UVA *School of Law*.

Dr. Murrie's duties at the ILPPP involve training clinicians to perform forensic evaluations and supervising the ILPPP's psychiatry and psychology trainees. He oversees a state-sponsored training program for psychologists and psychiatrists learning to perform court-ordered forensic evaluations.

As a forensic psychologist, Dr. Murrie performs a variety of criminal and civil forensic evaluations of juveniles and adults, both within the ILPPP's forensic clinic and in his private practice. These include forensic evaluations addressing adjudicative competence, legal sanity, death penalty sentencing, sexual offender risk assessment, and violence risk assessment.

Cognitive bias among forensic psychologists and psychiatrists

Daniel Murrie

Institute of Law, Psychiatry, & Public Policy,
University of Virginia

Some research funded by the National Science Foundation,
Law & Social Science Program,
Award No: SES 0961082

National Academy of Sciences Report (2009)

- Warned:
  - -Field reliability of techniques is unknown
  - -Techniques subject to bias because labs "lack independence" from agencies that use them
- Called for:
  - "…research on human observer bias and sources of human error in forensic examinations … to determine the effects of contextual bias in forensic practice."
- Did not specifically addressing forensic psychology or psychiatry. But similar principles apply….

"Crisis of Confidence" in Psychological Science

- Widely publicized misconduct and fraud
  - e.g., Stapels case
- Researchers won't share data for re-analysis
  - see Wicherts et al, 2011
- Widespread "questionable research practices;" Easy manipulations to obtain desired results
  - Simmons et al, 2011; John et al, 2012
- "Allegiance effects" in research on therapy, etc
  - Since Luborsky et al, 1975

Adversarial Allegiance in Forensic Mental Health Evaluation:

- The tendency for forensic evaluators to interpret data and form opinions in a manner that better supports the party that retains them

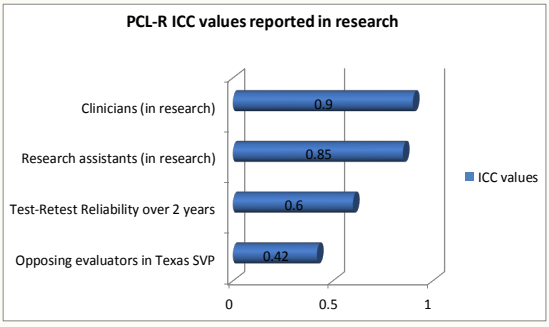
How might we measure allegiance?

- *Forensic Assessment Instruments have well-documented reliability values, at least in formal research studies.*
- *We know what reliability values we should expect from certain instruments*

In the field…

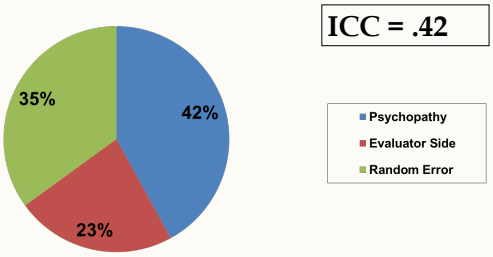
- Does reliability remain as strong?
- If not, do scores differ *systematically*, depending on the side that requested them?

One attempt to measure allegiance effects using Hare's Psychopathy Checklist-revised (PCL-R)



What determines a PCL-R score in Texas SVP cases?

ICC = .42



Combining analyses from:
Boccaccini et al, 2008
Murrie et al., 2008; 2009

Risk Measure Agreement among Opposing Evaluators: Texas SVP cases

	ICC (A, 1)	Mean score: *Prosecution*	Mean score: *Defense*	Effect size (d) for difference
PCL-R	.42	24.3	18.5	.78
MnSOST-R	.44	8.9	5.4	.85
Static-99	.62	4.8	4.3	.34

Murrie et al., 2009

Other perspectives on allegiance in the field

- The Static-99R shows least allegiance effects, perhaps because scoring is so structured
- But there is more room for subjective judgment in selecting the "norms" or comparison group for score reporting
- Do evaluators who work for different sides report different score reporting practices?

– (Chevalier, Boccaccini, & Murrie, in press)

National Survey of Static-99R score report practices

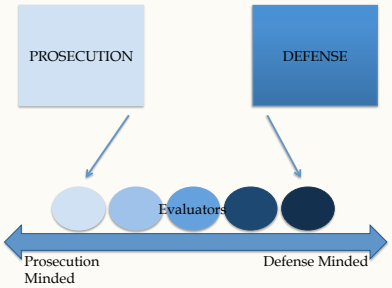
Comparisons of the Static-99R Reporting Practices of Petitioner, State Agency, and Defense Evaluators

	Percentage of evaluators[a]			Odds Ratio		
Survey question/response	Prosecution	State agency	Defense	Pros vs. State	Pros vs. Defense	State vs. Defense
Norms reported[b]						
High risk/need	94.4	64.3	33.3	9.43*	34.00***	3.60*
Non-routine	27.8	28.6	11.1	0.96	3.08	3.19
Preselected treatment	11.1	26.2	16.7	0.35	0.63	1.77
Routine sample	27.8	42.9	88.9	0.51	0.05***	0.09***
Norms most important for SVP evals?[c]						
High-risk/need	77.8	52.4	16.7	3.18	17.54***	5.49*
Routine sample	5.6	23.8	72.2	0.19	0.02***	0.12***
SVP evaluators should usually report high risk/need rates	83.3	66.7	11.1	2.50	40.00***	20.78***
Reports recidivism rate confidence interval	44.4	40.5	77.8	1.18	0.23*	0.19*
Reports classification accuracy statistics	5.6	9.5	38.9	0.56	0.09*	0.17**
Some difficulty choosing norms	27.8	59.5	33.3	0.26*	0.77	2.94

"Allegiance effects"?

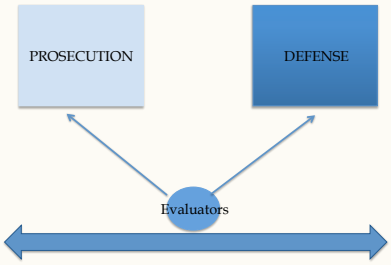
Or just selection effects?

Selection Effects



Shrewd attorneys retain evaluators who are *already* oriented towards their side.

Allegiance Effects



After retention, evaluators *become* more allied as they interpret case data in a way that supports the side that retained them.

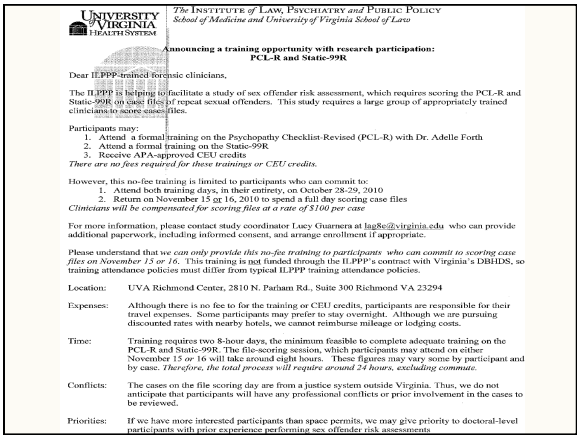
To *really* explore adversarial allegiance:

- Exclude attorney selection effects
- Exclude evaluator selection effects

- Ideally...a true experiment
  - Random assignment to opposing sides
  - Review identical case materials
  - Offer well-quantified opinions (e.g., test scores)

Exploring adversarial allegiance:

A TRUE EXPERIMENT



UNIVERSITY of VIRGINIA HEALTH SYSTEM

The INSTITUTE *of* LAW, PSYCHIATRY *and* PUBLIC POLICY
School of Medicine and University of Virginia School of Law

Announcing a training opportunity with research participation:
PCL-R and Static-99R

Dear ILPPP-trained forensic clinicians,

The ILPPP is helping to facilitate a study of sex offender risk assessment, which requires scoring the PCL-R and Static-99R on case files of repeat sexual offenders. This study requires a large group of appropriately trained clinicians to score cases files.

Participants may:
1. Attend a formal training on the Psychopathy Checklist-Revised (PCL-R) with Dr. Adelle Forth
2. Attend a formal training on the Static-99R
3. Receive APA-approved CEU credits

There are no fees required for these trainings or CEU credits.

However, this no-fee training is limited to participants who can commit to:
1. Attend both training days, in their entirety, on October 28-29, 2010
2. Return on November 15 or 16, 2010 to spend a full day scoring case files

Clinicians will be compensated for scoring files at a rate of $100 per case

For more information, please contact study coordinator Lucy Guarnera at lag8e@virginia.edu who can provide additional paperwork, including informed consent, and arrange enrollment if appropriate.

Please understand that *we can only provide this no-fee training to participants who can commit to scoring case files on November 15 or 16.* This training is not funded through the ILPPP's contract with Virginia's DBHDS, so training attendance policies must differ from typical ILPPP training attendance policies.

Location: UVA Richmond Center, 2810 N. Parham Rd., Suite 300 Richmond VA 23294

Expenses: Although there is no fee to for the training or CEU credits, participants are responsible for their travel expenses. Some participants may prefer to stay overnight. Although we are pursuing discounted rates with nearby hotels, we cannot reimburse mileage or lodging costs.

Time: Training requires two 8-hour days, the minimum feasible to complete adequate training on the PCL-R and Static-99R. The file-scoring session, which participants may attend on either November 15 or 16 will take around eight hours. These figures may vary some by participant and by case. *Therefore, the total process will require around 24 hours, excluding commute.*

Conflicts: The cases on the file scoring day are from a justice system outside Virginia. Thus, we do not anticipate that participants will have any professional conflicts or prior involvement in the cases to be reviewed.

Priorities: If we have more interested participants than space permits, we may give priority to doctoral-level participants with prior experience performing sex offender risk assessments

Response

- >100 applications, from 15 states
- Doctoral-level forensic clinicians
- Most with sex offender evaluation experience
- Two sessions
  - Fall (mostly Virginia and adjacent states)
  - Spring (broader range of states)

Training

- PCL-R
  - Adelle Forth
  - 1.5 days + evening homework
  - Practice scoring
  - Briefer than standard (less research background)
- Static-99
  - Eric Madsen (VaDOC, extensive experience)
  - .5 days training
  - Practice Cases

Why training?

- Recruitment Tool
- Ensures some uniformity in participant background and training
- Ensures more uniformity than most field settings

At the Conclusion of Training…

- Participants committed to return for paid file scoring
- Participants completed:
  - Background questionnaire
  - Experience questionnaire
  - "Typical PCL-R score" questionnaire
  - "Foreshadowing" questions

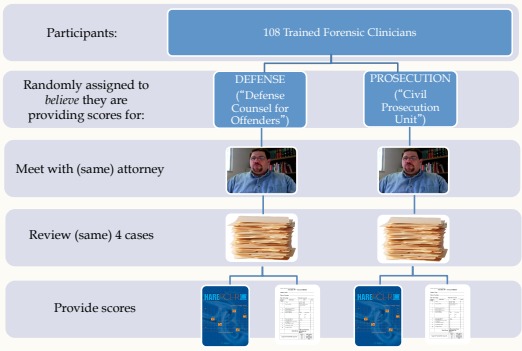
Dropping hints on training day….

21) If a state agency would like to know more details about the scores you assign during the file-scoring portion of this study, would you be open to discussing these? (Assume time would be compensated at reasonably hourly rate, and there would be no request for testimony or travel.)	a) No b) Yes

*Please note: Follow-up correspondence is **not** a part of study procedures and is **not** required to receive CEU credits, no-fee training, or payment for file scoring. UVA and ILPPP have no oversight, involvement, or responsibility for any follow-up correspondence between study participants and external agencies.*

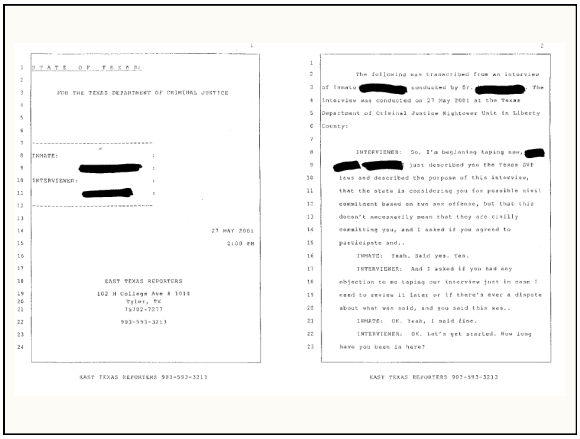
Experiment

- Participants returned 2-3 weeks after training
- Informed that large-scale consultation was arranged by a Texas agency to review pending SVP cases, due to concerns about original screening evaluators



Materials

- Actual SVP files (sanitized)
- Files included
  - Law enforcement records
  - Correctional records
  - Treatment Program Clinical interview
  - *Fabricated* PCL-R interview transcript (designed to correspond to case file)



STATE OF TEXAS:

VVH THE TEXAS DEPARTMENT OF CRIMINAL JUSTICE

INMATE: [redacted]

INTERVIEWER: [redacted]

23 MAY 2001
2:00 PM

EAST TEXAS REPORTERS
102 N College Ave # 1014
Tyler, TX
75702-7277
903-593-3213

EAST TEXAS REPORTERS 903-593-3213

The following was transcribed from an interview of inmate [redacted] conducted by [redacted]. The interview was conducted on 23 May 2001 at the Texas Department of Criminal Justice Hightower Unit in Liberty County:

INTERVIEWER: So, I'm beginning taping now [redacted] just described you the Texas DVP laws and described the purpose of this interview, that the state is considering you for possible civil commitment based on two sex offense, but that this doesn't necessarily mean that they are civilly committing you, and I asked if you agreed to participate and..

INMATE: Yeah. Said yes. Yes.

INTERVIEWER: And I asked if you had any objection to me taping our interview just in case I need to review it later or if there's ever a dispute about what was said, and you said this was..

INMATE: OK. Yeah, I said fine.

INTERVIEWER: OK. Let's get started. How long have you been in here?

EAST TEXAS REPORTERS 903-593-3213

Cases

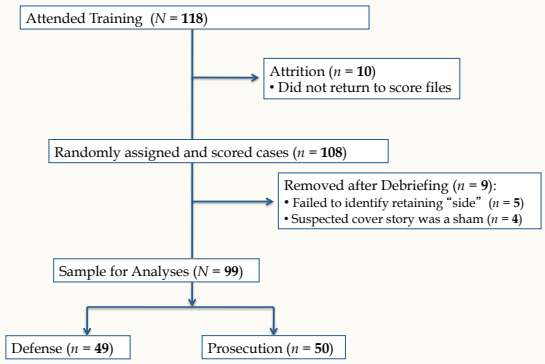
			Victims	
Randomized Order	1	PK	Teenage males	Mid-range PCL-R
	2	TR	Adult females	Higher PCL-R
	3	KL	Child + teen males	Higher PCL-R
Always Last	4	EJ	Children, female	Very low PCL-R

Measures

- When returning each file, participants provided:
  - PCL-R score
  - Static-99 score
- After completing *all* files, participants completed:
  - Attitudes measure (1 = disagree, 5 = agree)
  - e.g., "We need special policies and procedures for sex offenders to help protect the public."

Debriefing

- Manipulation check
  - Did they understand the assignment?
  - Suspicions or doubts?
- Explanation of **true** study purpose
  - Comments
- Still receive payment and CEUs
- Invitation for follow-up survey



Did scores differ depending on the side that requested them?

PCL-R Results

Case 1: Allegiance Effects

	Prosecution	Defense	
PCL-R Score	*M* (*SD*)	*M* (*SD*)	*d*
Total	16.6 (3.5)	13.4 (4.1)	.85**
Factor 1	11.2 (2.6)	8.9 (3.2)	.78**
Factor 2	3.9 (1.7)	3.1 (1.6)	.45*

*$p < .05$. **$p < .01$

Case 1: Score Ranges

- But… evaluators on the *SAME SIDE* often assigned very different PCL-R Total scores

Side	*Min*	*Max*	*Range*
Prosecution	11.0	29.0	18.0
Defense	5.0	22.0	17.0

How Likely are "Large" Differences?

- This variability means that:
  - Although state scores are, *on average*, higher than defense scores…
  - …defense scores are, at times, higher than state scores.
- If we randomly select one state and one defense evaluator,
  - How often do they differ by > 6.0 points (2 SEM)?
  - These (tedious) analyses are more relevant to the field

Difference Scores Example

- Calculate difference between each state and each defense evaluator

Side	*Prosecution*	*Defense*	*dif*
Example 1	20	10	+10
Example 2	20	29	-9

Case 1 Difference > 6.0

Difference	%
Prosecution > Defense by 6.0+	29%
Defense > Prosecution by 6.0+	4%

Case 1 Difference > 3.0

Difference	%
Prosecution > Defense by 3.0+	51%
Defense > Prosecution by 3.0+	11%

Results: mean PCL-R scores

Case:	*Prosecution Expert*	*Defense Expert*	*Effect size*
1	**16.6**	**13.4**	**.85*** **
2	**26.5**	**23.2**	**.76*** **
3	**26.4**	**24.0**	**.55** **
4	**7.8**	**7.8**	**-.01**

Effect size expressed as Cohen's d.
**p < .05. **p < .01. ***p < .001.*

Results: What percentage of opposing evaluator pairs would differ by twice the SEM?

Case:	*Prosecution > Defense*	*Defense > Prosecution*
1	**29%**	**4%**
2	**33%**	**7%**
3	**28%**	**9%**
4	**13%**	**12%**

Results reflect randomly selecting every possible combination of defense/prosecution pairs for each case (~2,400), and calculating the percentage of score differences greater than 2SEM (or 6 points) on PCL-R.
In research contexts, score differences of >2SEM occur in <2% of cases

Quick Summary

- When we control for selection effects…
  - We find adversarial allegiance effect in 3 of 4 cases
  - Prosecution scores about 3 points higher than defense, *on average*
  - Most "Big" (> 3.0 or > 6.0 points) differences are in the direction of adversarial allegiance

But, does it depend on…

- Does allegiance effect depend on ….?
  - Prior PCL-R experience
  - SVP attitudes
  - "Typical" PCL-R scores
- No, no, and no
  - No moderating effects

Static-99R Results

Static-99R

Cases	Prosecution *M* (*SD*)	Defense *M* (*SD*)	*d*
Case 1	4.5 (.85)	4.1 (1.0)	.42*
Case 2	5.6 (1.3)	5.3 (1.1)	.24
Case 3	5.6 (1.8)	5.3 (1.6)	.20
Case 4	1.9 (1.2)	1.7 (1.1)	.14

Field vs. Experimental Findings

Compare and Contrast Designs

- Field study (Murrie et al., 2008; 2009)
  - Attorneys select experts (mostly)
  - Score differences could be due to **adversarial allegiance** *or* **selection effects**
- Experiment
  - Randomly assign experts to sides (*no* selection)
  - Any effects we observe cannot be selection effects

Compare and contrast findings (PCL-R)

	Field	Experiment
Mean difference	6.0	3.0
State 6.0+ higher	40%	30%
Defense 6.0+ higher	6%	11%

- Selection likely accounts for *some*, but not all of the effect observed in the field

Compare and contrast (Static-99R)

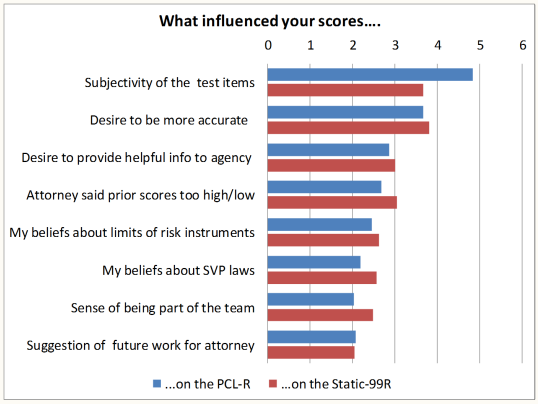
	Field	Experiment
Mean difference	0.5	0.3
State 2 SEM+ higher	16%	18%
Defense 2 SEM + higher	4%	10%

- Selection likely accounts for *some*, but not all of the effect observed in the field

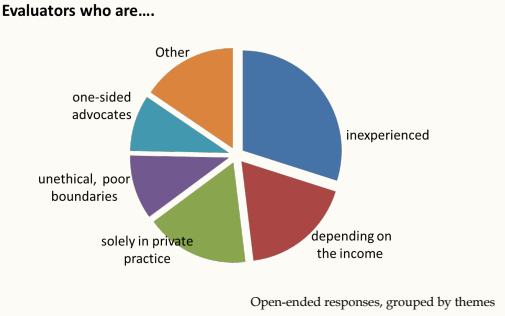
What did participants think about allegiance?

After the study and debriefing

- Participants left with their own scoresheets and the "correct" scores
- Follow-up, online survey
  - (for additional CEUs)
- 60% response rate
- Divided evenly between defense and prosecution



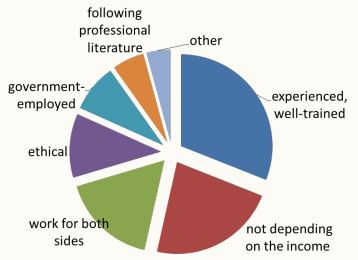
Who is *most* vulnerable to allegiance?



Open-ended responses, grouped by themes

Who is *least* vulnerable to allegiance?

Evaluators who are….



Open-ended responses, grouped by themes

Allegiance is a problem. For others.

Participants who….	tended to name *these* evaluators…	…as most vulnerable to allegiance effects.
Worked for state facilities		Private practice evaluators
Were more experienced		Inexperienced evaluators
Were older		"Younger" "Novice" or "Less mature" evaluators
Worked in academic settings		Evaluators who lacked training, especially reliability training

"Bias Blind Spot" (Pronin, 2007)

- We recognize bias in human judgment … except when that bias is our own.
- Because:
  1. We rely on introspection to screen for bias
     - …but bias is usually non-conscious
  2. We assume our perceptions directly reflect reality ("naive realism")
     - …so anyone who perceives differently must be biased

Other contributing factors:

- Outright "Hired-Gun" behaviors
  - (probably uncommon)
- Common Cognitive Errors
  - Expectancy Effects
  - Anchoring
  - Suggestion Effects
  - Confirmation Bias
  - Motivated Reasoning

How might we reduce allegiance?

- Structural changes:
  - "Neutral experts"
    - Often recommended, but bring new challenges
  - "Blinded" referrals
    - Borrowed from research methods
  - Term limited evaluators
    - Borrowed from accounting/ auditing
- Clinician Changes:
  - Improved Evaluator training and oversight
  - Self scrutiny as habit, and professional priority

Thank you:

- NSF, Law and Social Sciences program
- Adelle Forth
- Eric Masden
- "Blue Ribbon Panel"
- Pilot-testers
- Study Consultants

For comments or more info:

- Murrie@Virginia.edu

Are Forensic Experts Biased by the Side That Retained Them?

Daniel C. Murrie, Marcus T. Boccaccini, Lucy A. Guarnera and Katrina A. Rufino

Psychological Science 2013 24: 1889 originally published online 22 August 2013
DOI: 10.1177/0956797613481812

General Article

Psychological Science
24(10) 1889–1897

DOI: 10.1177/0956797613481812
pss.sagepub.com

Are Forensic Experts Biased by the Side That Retained Them?

Daniel C. Murrie[1], Marcus T. Boccaccini[2], Lucy A. Guarnera[1], and Katrina A. Rufino[2]

[1]Institute of Law, Psychiatry, and Public Policy, University of Virginia, and [2]Department of Psychology and Philosophy, Sam Houston State University

Abstract

How objective are forensic experts when they are retained by one of the opposing sides in an adversarial legal proceeding? Despite long-standing concerns from within the legal system, little is known about whether experts can provide opinions unbiased by the side that retained them. In this experiment, we paid 108 forensic psychologists and psychiatrists to review the same offender case files, but deceived some to believe that they were consulting for the defense and some to believe that they were consulting for the prosecution. Participants scored each offender on two commonly used, well-researched risk-assessment instruments. Those who believed they were working for the prosecution tended to assign higher risk scores to offenders, whereas those who believed they were working for the defense tended to assign lower risk scores to the same offenders; the effect sizes (*d*) ranged up to 0.85. The results provide strong evidence of an allegiance effect among some forensic experts in adversarial legal proceedings.

Keywords

forensic science, forensic assessment, forensic psychology, bias, risk assessment, adversarial allegiance

Received 9/17/12; Revision accepted 2/16/13

Recently, the National Research Council (NRC, 2009) warned that the accuracy and reliability of many popular forensic-science techniques are unknown, that error rates are rarely acknowledged, and that forensic scientists are prone to bias because they are not independent of the parties requesting their services. Emerging research has clearly documented subjectivity and bias even in the forensic-science procedures that courts have tended to consider most reliable, such as analyses of DNA (Dror & Hampikian, 2011) and fingerprints (Dror & Cole, 2010). Thus, the NRC urged further research on the cognitive and contextual biases that influence forensic experts.

The NRC report did not specifically address mental-health experts or forensic psychological evaluations. But psychological evaluations—like other forensic-science procedures—are often admitted as evidence or presented via expert testimony in adversarial legal proceedings. Indeed, evaluations by mental-health experts influence decisions as grave as death sentences (*Barefoot v. Estelle*, 1983) and indefinite civil confinement (*Kansas v. Hendricks*, 1997). Therefore, recent concerns regarding forensic science raise questions about whether forensic psychological evaluations might suffer similar problems of unreliability and bias.

How reliable are forensic psychologists and psychiatrists when they are retained as experts in adversarial legal proceedings? For more than a century, courts and legal scholars have lamented apparent bias among medical experts (Bernstein, 2008; Hand, 1901; Mnookin, 2008; Wigmore, 1923). Likewise, practicing judges and attorneys have complained that experts sacrifice objectivity for advocacy (e.g., Krafka, Dunn, Johnson, Cecil, & Miletich, 2002). But little psychological research has investigated what we call *adversarial allegiance* (Murrie et al., 2009), the presumed tendency for experts to reach conclusions that support the party who retained them. Psychology's delay in investigating adversarial allegiance

Corresponding Author:
Daniel C. Murrie, Institute of Law, Psychiatry, and Public Policy, UVA Box 800660, Charlottesville, VA 22908-0660
E-mail: murrie@virginia.edu

is disappointing, because psychologists are uniquely suited to explore reliability and bias in decision making.

Field Studies of Risk Instruments Suggest, but Do Not Prove, Adversarial Allegiance

Recently, we investigated adversarial allegiance by examining civil commitment proceedings for sex offenders, also known as sexually-violent-predator (SVP) trials. SVP trials provide an ideal context for studying the possibility of adversarial allegiance, because court decisions depend largely on weighing testimony from opposing experts. Twenty states and the federal system have SVP laws, which allow them to identify sexual offenders whom they consider likely to reoffend and confine them indefinitely after their incarceration (*Kansas v. Hendricks*, 1997). SVP proceedings routinely involve forensic psychologists and psychiatrists who are retained by opposing sides, conduct risk assessments of the same offender, and consider the same data, often using the same instruments. So we could study adversarial allegiance in SVP proceedings by comparing the scores that defense-retained and prosecution-retained evaluators assigned to offenders using popular risk-assessment instruments (Murrie, Boccaccini, Johnson, & Janke, 2008; Murrie et al., 2009).

Scores on risk instruments are an ideal metric to measure expert opinions because (a) experts routinely administer these instruments to inform legal proceedings, and (b) dozens of studies have documented strong interrater agreement when clinicians score these instruments in research and practice contexts that are *not* adversarial. For example, Hare's (2003) Psychopathy Checklist–Revised (PCL-R), an instrument that relies on clinical interview and review of records, is widely used in forensic assessments of risk for violence or sexual violence (Skeem, Polaschek, Patrick, & Lilienfeld, 2011). The PCL-R manual reports strong interrater agreement (intraclass correlation, or ICC = .87; Hare, 2003). Indeed, most (92%) pairs of scores from trained raters who score the same offender differ by fewer than 2 points (Gacono & Hutton, 1994), even though PCL-R scores can range from 0 to 40.

However, in a small sample of SVP proceedings that featured PCL-R scores from defense-retained and prosecution-retained evaluators, the ICC for opposing evaluators was .42, which indicated that less than half of the variance in PCL-R scores could be attributed to the offenders' true standing on the PCL-R (Murrie et al., 2009). Moreover, the average PCL-R score from prosecution experts was 24, whereas the average score from defense experts was only 18 (Cohen's d = 0.78). The PCL-R may be especially vulnerable to this allegiance effect because it requires clinicians to make inferences about an offender's personality and emotions (e.g., lack of guilt or remorse, superficial charm). The adversarial-allegiance effect was smaller (d = 0.34) for the Static-99 (Hanson & Thornton, 2000), a highly structured measure scored from file information about criminal history that requires less subjective judgment.

These field studies (Murrie et al., 2008; Murrie et al., 2009) strongly suggest adversarial allegiance, in that prosecution-retained evaluators assigned higher scores and defense-retained evaluators assigned lower scores to the same offenders. But we cannot draw firm conclusions from these field studies alone, because they investigated scores from experts selected by attorneys. Conceivably, attorneys could have chosen specific experts because they perceived the experts *already* had attitudes or scoring tendencies conducive to their case. Or perhaps attorneys consulted many experts, but arranged testimony only from those whose opinions were most supportive of their case. For example, a defense attorney might retain several evaluators to examine a client, but request testimony only from the evaluator who assigned the lowest risk scores. Thus, the apparent allegiance in field studies might reflect selection effects, whether in terms of which expert an attorney selected to perform an evaluation or which findings an attorney selected to present at trial.

Understanding Adversarial Allegiance Requires a True Experiment

Field studies raise an important question that can be answered only with a true experiment. Is apparent allegiance due simply to attorneys choosing evaluators who have preexisting attitudes that favor their side, or to attorneys calling only experts with the most favorable findings to testify in court (selection effects)? Or do evaluators, once retained and promised payment by one side, tend to form opinions that favor that side (allegiance effects)? If an experiment using random assignment failed to find allegiance effects, it would suggest that the apparent allegiance in the field is due primarily to one or both of these selection effects. But if an experiment using random assignment *did* find allegiance effects, it would suggest that being retained and paid by one side in an adversarial system may compromise objectivity among experts.

To answer this question, we recruited more than 100 experienced forensic psychologists and psychiatrists, provided 2 days of in-person training on risk instruments from established experts, had them meet with an attorney, and then paid them to score risk instruments for up to four offenders. We deceived participants to believe they were performing a large-scale, paid forensic consultation. But unbeknownst to participants, they all received

exactly the same four offender files, and each participant was randomly assigned to believe that he or she was working for either the prosecution or the defense.

Method

Participants

We sent recruitment correspondence to a broad group of practicing forensic evaluators, offering "gold standard" training (and continuing-education credits) on the two most commonly used measures in sex-offender risk assessments: the PCL-R and Static-99R (Helmus, Thornton, Hanson, & Babchishin, 2012). This training was offered at no cost to participants who could commit to returning a few weeks later to spend 1 day scoring offenders at a pay rate typical of forensic consultation ($400). We received more than 100 applications from practicing, doctoral-level forensic clinicians.

Of the 118 clinicians who participated in the risk-measure training, 108 returned to score files for the experiment.[1] Five who scored cases did not pass a manipulation check (i.e., they could not identify which side had retained them), and 4 expressed some suspicion that the cover story of scoring cases for a forensic consultation was a sham (see the Debriefing section). So we report results for the 99 participants (49 ostensibly retained by the defense, 50 ostensibly retained by the prosecution) who accepted the manipulation and believed they were scoring cases for one side of an adversarial process.

Participants (60% female, 40% male) came from 15 states. Most (88%) reported having doctoral degrees in psychology (Ph.D. or Psy.D.). Others reported having a medical degree (7%) or another type of doctoral degree (5%). Most (84%) reported that they had experience conducting forensic evaluations, and most (75%) reported that they had experience conducting sex-offender risk assessments. About half (51%) had used the PCL-R in practice, and about half (49%) had used the Static-99R in practice.

Training

The participants attended a single 2-day training. The first 1.5 days (14 hr) involved training on the PCL-R, conducted by an internationally known expert who had coauthored one version of the Psychopathy Checklist and provided many formal PCL-R workshops. The final half-day of training (4 hr) focused on the Static-99R. Our goal was not to train participants to a predetermined level of reliability (a common practice in validity studies) because evaluators in the field are never required to demonstrate a specified level of reliability before accepting cases. Rather, we provided training to ensure that all participants had, at a minimum, completed the type of high-quality workshop that is offered to professionals in the field. Many evaluators cite workshop training as evidence of their qualifications to score risk measures for SVP cases (Rufino, Boccaccini, Hawes, & Murrie, 2012), although it is possible that some evaluators administer these measures after receiving less formal training. Regarding deception at the training stage, participants were informed only that the training and subsequent scoring were funded by an "out-of-state agency" that wanted to ensure that all participants had rigorous training before they scored offender files.

Deception and experimental manipulation: scoring cases for the prosecution or defense

Participants returned about 3 weeks later to score offender files. They were randomly assigned[2] to either a prosecution-allegiance or a defense-allegiance group and were deceived to believe that they were a part of a formal, large-scale forensic consultation paid for by either a public-defender service or a specialized prosecution unit that prosecutes SVP cases. Immediately after arrival, participants met for 10 to 15 min with a confederate (a former SVP attorney) who posed as an attorney for either the public-defender service or the specialized prosecution unit. The same attorney played both roles, but followed a slightly different script (see the Supplemental Material available online) depending on whether the participant had been randomly assigned to the defense or the prosecution.

The attorney addressed the defense-allegiance participants with statements that are typical of many defense attorneys (e.g., "We try to help the court understand that the data show not every sex offender really poses a high risk of reoffending"). Likewise, he addressed participants in the prosecution-allegiance condition with statements that are typical of prosecutors (e.g., "We try to help the court understand that the offenders we bring to trial are a select group whom the data show are more likely than other sex offenders to reoffend"). In both conditions, he asked participants to score the offenders using the two risk instruments. He also hinted at the possibility of future opportunities for paid consultation.

Participants were led to believe that, as a group, they were reviewing and scoring cases from a large cohort. But in truth, all participants scored the same four case files, which we selected to span the range from low risk to high risk. Each set of case materials was authentic (i.e., from an actual SVP case). The files included de-identified, but real, court, criminal, and correctional records. Specifically, these included real police investigation and

arrest documents; victim and witness statements; plea, judgment, and sentencing documents from court; presentence investigation reports; criminal-history summary documents; prison intake and case-summary documents; prison placement documents; and prison disciplinary records. Prison records also included some material from routine psychological assessments performed by the prison's sex-offender treatment program, that is, results from the Personality Assessment Inventory (Morey, 1991) and a clinical interview (similar in content to a PCL-R interview) conducted by treatment staff. Again, all of these records were real, but de-identified, material unique to each of the four cases. Finally, each file also included a realistic transcript of a fabricated PCL-R interview that we wrote to correspond to that file's records. The fabricated PCL-R interview transcripts were cosmetically altered to appear as if they were part of the original records.

The four offender files were selected to be representative of SVP cases generally. One sex offender had adult victims, whereas three had child victims. All had been convicted of multiple sexual offenses. After the participants reviewed a case file,[3] they scored the PCL-R and Static-99R.

Measures

Psychopathy Checklist–Revised. Hare's (2003) PCL-R is a 20-item measure of interpersonal, emotional, and behavioral traits, which clinicians score on the basis of an offender's records and a clinical interview. PCL-R items are rated on a scale from 0 to 2, with higher scores reflecting a higher level of the psychopathic trait; these scores are summed to yield a Total score that can range from 0 to 40. Although forensic evaluators usually emphasize PCL-R Total scores in reports or testimony, PCL-R items are divided into two factors: Factor 1 consists of an Interpersonal facet and an Affective facet, and Factor 2 (Social Deviance) consists of an Impulsive Lifestyle facet and an Antisocial Behavior facet.

The PCL-R is the most widely used and well-researched measure of psychopathy, a personality construct characterized by a self-serving interpersonal style, shallow emotions, an unstable lifestyle, and antisocial behavior. Although it was not originally developed for risk assessment, ample research suggests that PCL-R scores correspond with violence and recidivism. For example, meta-analyses have found that PCL-R Total scores tend to be moderately associated with antisocial behavior (Leistico, Salekin, DeCoster, & Rogers, 2008), including sexual violence (Hawes, Boccaccini, & Murrie, 2013). Thus, the measure has become widely used in assessments of risk for violence or sexual violence, and courts routinely admit expert testimony regarding PCL-R scores (DeMatteo & Edens, 2006).

The PCL-R manual (Hare, 2003) reports strong agreement among independent raters for PCL-R Total scores (ICC = .87), at least outside of adversarial legal proceedings. But the manual also reveals that interrater agreement tends to be stronger for Factor 2 items that relate to antisocial behavior (e.g., criminal versatility, juvenile delinquency) and weaker for Factor 1 items (e.g., failure to accept responsibility, glibness/superficial charm), which may require more clinical inference.

Static-99R. The Static-99R is an actuarial risk-assessment instrument designed to predict sexual recidivism among sex offenders (Helmus et al., 2012). Composed of 10 items that address an offender's age and prior living arrangements, as well as several aspects of his offense history, the Static-99R is scored on the basis of file review. According to the Static-99 Clearinghouse (n.d.), the Static-99 (and now the Static-99R) is "the most widely used sex offender risk assessment instrument in the world, and is extensively used in the United States, Canada, the United Kingdom, Australia, and many European nations." It is widely accepted in legal proceedings, given its strong empirical relation to important outcomes and strong evidence of validity and reliability. For example, the Static-99 score is among the best-known predictors of sexual recidivism, and a meta-analysis of more than 60 studies found a mean predictive effect (*d*) of 0.67 (Hanson & Morton-Bourgon, 2009). A recent review of rater-agreement coefficients found a median rater-agreement value of .90 (Hanson & Morton-Bourgon, 2009), suggesting that the Static-99 and Static-99R meet or exceed commonly accepted standards for reliability in psychological measures. Compared with PCL-R items, Static-99R items (e.g., age at release, any male victims) appear fairly straightforward and require less clinical inference to score.

Clinician attitudes. One potential explanation for any allegiance effects we might observe would be preexisting differences in clinicians' attitudes (i.e., if participants assigned to score files for the prosecution tended to have a harsher perspective on sexual offenders than participants assigned to score files for the defense). So, although we randomly assigned participants to the prosecution and defense conditions, we nevertheless had participants complete two additional measures that allowed us to check whether participants in the two conditions were similar in their attitudes regarding sexual offenders.

We asked participants to complete a five-item questionnaire at the end of the scoring day, to avoid revealing that their attitudes and scoring patterns were the focus of study. The questionnaire asked them to rate the extent to which restrictive policies for sex offenders (e.g., SVP laws) are necessary and reasonable. For example, one item read, "Laws that allow states to civilly commit

potentially dangerous sex offenders who have completed their sentences are reasonable strategies to protect people in the community" (1 = *strongly disagree*, 5 = *strongly agree*). Internal consistency for this attitudes measure was .79. We also asked participants (at the end of PCL-R training) to report their best estimate of the typical PCL-R Total score among offenders who have committed sexually violent crimes against (a) adults and (b) children.

Debriefing

After participants completed the presumed forensic consultation, we performed a manipulation check, in which a member of the research team met privately with each participant. The researcher asked about the participant's understanding of study goals, and then asked explicitly whether the participant was suspicious about any additional or hidden study goals. The 4 participants who conveyed any degree of suspicion (ranging from vague suspicion to more specific guesses about alternate study goals) were excluded from subsequent data analysis, as were the 5 who could not identify which side retained them. The researcher then described the experimental manipulation and the true study goals. Although all participants had the option of withdrawing their data from the study, none did so. All received the payment ($400) and continuing-education credits originally promised.

Results

Overall, the risk scores assigned by prosecution and defense experts showed a clear pattern of adversarial allegiance. As expected, allegiance effects were stronger for the PCL-R, a measure that requires more subjective clinical judgment, than for the Static-99R, a measure that requires less clinical judgment (see Table 1). For the PCL-R Total score, independent-samples t tests indicated that prosecution-retained evaluators assigned significantly higher scores than defense-retained evaluators for Case 1, $t(94) = 4.15$, $p < .001$; Case 2, $t(94) = 3.73$, $p < .001$; and Case 3, $t(97) = 2.71$, $p = .008$; but not Case 4, $t(62) = -0.33$, $p = .97$. Cohen's d for the three cases with significant effects ranged from 0.55 to 0.85, and were similar in magnitude to effects (d = 0.63–0.83) documented in a sample of actual SVP proceedings (Murrie et al., 2009). The one case for which the PCL-R Total

Table 1. Differences Between Risk-Measure Scores From Evaluators Randomly Assigned and Paid to Score Cases for the Prosecution or the Defense

	Prosecution		Defense		Effect size	
Score and case	*M*	*SD*	*M*	*SD*	Cohen's *d*	95% confidence interval
PCL-R Total						
Case 1	16.64	3.50	13.41	4.10	0.85***	[0.43, 1.26]
Case 2	26.53	4.32	23.22	4.37	0.76***	[0.35, 1.17]
Case 3	26.40	4.69	24.00	4.14	0.55**	[0.14, 0.94]
Case 4	7.81	4.09	7.84	3.36	–0.01	[–0.32, 0.31]
PCL-R Factor 1 (Interpersonal/Affective)						
Case 1	11.22	2.60	8.95	3.20	0.78***	[0.36, 1.18]
Case 2	8.34	2.72	6.51	2.95	0.65**	[0.23, 1.05]
Case 3	11.91	2.80	11.27	2.52	0.24	[–0.15, 0.63]
Case 4	4.74	3.30	4.60	2.66	0.05	[–0.44, 0.54]
PCL-R Factor 2 (Social Deviance)						
Case 1	3.86	1.68	3.13	1.60	0.44*	[0.04, 0.85]
Case 2	15.61	2.26	14.45	2.19	0.52**	[0.11, 0.93]
Case 3	12.26	2.36	10.65	2.00	0.73***	[0.33, 1.14]
Case 4	2.58	1.45	2.98	1.79	–0.25	[–0.74, 0.25]
Static-99R						
Case 1	4.46	0.85	4.06	1.05	0.42*	[0.01, 0.82]
Case 2	5.56	1.35	5.27	1.05	0.24	[–0.16, 0.64]
Case 3	5.62	1.81	5.29	1.57	0.20	[–0.20, 0.59]
Case 4	1.85	1.21	1.69	1.11	0.14	[–0.35, 0.64]

Note: Evaluators scored cases using the Psychopathy Checklist–Revised (PCL-R; Hare, 2003) and the Static-99R (Helmus, Thornton, Hanson, & Babchishin, 2012). Statistical significance of the difference between conditions was determined using independent-samples t tests (two-tailed). For the four cases, ns were as follows—Case 1: $n = 96$; Case 2: $n = 96$; Case 3: $n = 99$; Case 4: $n = 64$.
*$p \leq .05$. **$p \leq .01$. ***$p \leq .001$.

scores did not show an allegiance effect was one we had selected to be unusually low in psychopathy;[4] this case received unusually low scores both from prosecution-retained (M = 7.81) and defense-retained (M = 7.84) evaluators.

Adversarial-allegiance effects were evident for both Factor 1 (Interpersonal/Affective) and Factor 2 (Social Deviance) scores from the PCL-R, as detailed in Table 1. In terms of absolute value, Factor 1 effects were larger than Factor 2 effects in two of the three cases with Total score allegiance effects, which is consistent with findings that Factor 1 items tend to require more subjective judgment to score (Rufino, Boccaccini, & Guy, 2011). For Case 3, however, there was a significant effect for Factor 2 scores (d = 0.73, p < .001), but not Factor 1 scores (d = 0.24, p = 24). Examination of the Factor 1 facets for Case 3 indicated that there was some evidence for an allegiance effect for Facet 2 (Affective traits) scores, $t(97)$ = 1.94, p = .06, d = 0.39, 95% confidence interval (CI) = [–0.01, 0.79], but not Facet 1 (Interpersonal traits) scores, $t(97)$ = 0.08, p = .94, d = 0.01, 95% CI = [–0.38, 0.41].

For the Static-99R, a more structured measure, prosecution-retained evaluators tended to assign higher scores than defense-retained evaluators in each of the four cases (see Table 1), but the difference was large enough to reach statistical significance for only Case 1 (d = 0.42, p = .05). The effect sizes across these four cases (ds = 0.14, 0.20, 0.24, and 0.42) were similar to, although somewhat smaller than, the effect sizes (d = 0.29–0.37) reported across 27 actual SVP cases (Murrie et al., 2009).

Differences among pairs of prosecution- and defense-retained evaluators

In court, judges and juries would never consider risk-instrument scores that have been averaged across many experts. Rather, they usually hear expert testimony about risk scores from two experts: one called by each opposing side. Moreover, because all test scores are influenced to some extent by random measurement error, it is unrealistic to expect two experts to assign exactly the same score in every case. Small score differences may be trivial, even if they are in the direction of allegiance. The mean scores in Table 1 do not provide any information about how often, if ever, one might expect large, nontrivial differences in risk scores within pairs of opposing experts.

Therefore, we conducted a series of follow-up analyses to examine how likely it was that a randomly selected prosecution-retained evaluator and a randomly selected defense-retained evaluator would assign scores that were so different that they could not be explained by expected random measurement error. We considered the difference between a pair of scores to be meaningful if it was more than twice the standard error of measurement (SEM) for the risk instrument. The SEM is the amount that experts' scores for the same offender could be expected to differ as a result of random measurement error. Given a normal curve, one would expect only about 32% of difference scores to be larger than the SEM, and only about 4% to be more than twice as large as the SEM (i.e., > 2 SEM units). In the absence of adversarial allegiance, prosecution-retained evaluators would be expected to assign scores that are more than twice the SEM higher than the scores of defense-retained evaluators in about 2% of cases, and vice versa.

For each of the four cases, we calculated a difference score for each possible pairing of prosecution- and defense-retained evaluators. This process yielded approximately 2,400 difference scores for each measure, for each case. We then calculated the percentage of difference scores that were more than twice the SEM in the direction of allegiance (prosecution's score > defense's score) and the percentage that were more than twice the SEM in the opposite direction (see Table 2). The SEM for the PCL-R is about 3.0 points, and the SEM for the Static-99R is about 1.0 point.

The findings in Table 2 show two clear effects. First, more than 20% of the score pairings for each case led to a score difference that was more than twice the SEM, although only about 4% of score pairings in research contexts lead to score differences this large. There were four instances in which more than 35% of the score pairings led to differences that were greater than 2 SEMs:

Table 2. Percentage of Opposing Evaluator Pairs Whose Difference in Risk Scores Was Greater Than Twice the Standard Error of Measurement

Score and case	Prosecution's score > defense's score	Defense's score > prosecution's score
PCL-R		
Case 1	29%	4%
Case 2	33%	7%
Case 3	28%	9%
Case 4	13%	12%
Static-99R		
Case 1	18%	7%
Case 2	20%	12%
Case 3	28%	21%
Case 4	20%	18%

Note: Evaluators scored cases using the Psychopathy Checklist–Revised (PCL-R; Hare, 2003) and the Static-99R (Helmus, Thornton, Hanson, & Babchishin, 2012).

Cases 2 (40%) and 3 (37%) for the PCL-R and Cases 3 (49%) and 4 (38%) for the Static-99R. Second, most large (i.e., > 2 SEM) differences were in the direction of adversarial allegiance, with the prosecution-retained evaluator assigning higher scores and the defense-retained evaluator assigning lower scores. This pattern was especially clear for the PCL-R. For the three cases with clear PCL-R allegiance effects, 28% or more of all possible score pairings led to a score difference of more than 2 SEMs in the direction of allegiance. Again, score differences greater than 2 SEMs in one direction (e.g., prosecution's score > defense's score) should occur in only about 2% of cases, according to rater-agreement values from nonadversarial-research contexts. Between 4% and 9% of PCL-R score pairings in the three cases with clear allegiance effects led to large differences in the opposite direction, which is also more than the 2% expected on the basis of nonadversarial research, but these differences clearly were not as common as large differences in the direction of allegiance (≥ 28%).

Potential explanations for allegiance effects

One possible alternate explanation for our findings is that, despite random assignment, evaluators assigned to score for the prosecution maintained harsher attitudes toward sex offenders or had different types of clinical experience than did those assigned to score for the defense. But we found no evidence for this alternate explanation. Prosecution- and defense-retained evaluators did not differ in their ratings on our five-item measure of support for restrictive sex-offender policies, $t(97) = 0.07$, $p = .95$, $d = 0.02$; their estimate of the typical PCL-R Total score among sex offenders with adult victims, $t(93) = 0.51$, $p = .62$, $d = 0.10$; or their estimate of the typical PCL-R Total score assigned to sex offenders with child victims, $t(93) = 0.25$, $p = .80$, $d = 0.05$. Likewise, prosecution- and defense-retained evaluators did not differ in the percentage who had used the Static-99R in practice (52% vs. 45%), $\chi^2(1, N = 99) = 0.50$, $p = .48$, odds ratio = 1.33. Those assigned to score for the prosecution were somewhat more likely (62%) to have used the PCL-R in practice than were those assigned to score for the defense (41%), $\chi^2(1, N = 99) = 4.45$, $p = .04$, odds ratio = 2.36, but this is a difference that would actually reduce the likelihood of observing an allegiance effect because participants with more experience tended to assign lower PCL-R scores (reported previously by Guarnera, Murrie, Boccaccini, & Rufino, 2012).

Participants with higher scores on the attitude measures also tended to assign higher scores in some cases, but these effects were similar in size and direction for prosecution- and defense-retained evaluators (Guarnera et al., 2012). We could find only one instance in which an attitude or experience measure might help explain an allegiance effect. Recall that the strongest Static-99R allegiance effect occurred in Case 1 ($d = 0.42$). A two-way analysis of variance on Static-99R scores revealed a statistically significant interaction between condition and prior use of the Static-99R in practice, $F(1, 91) = 4.38$, $p = .04$. Specifically, there was a clear allegiance effect for evaluators who had not used the Static-99R in practice ($d = 0.71$, 95% CI = [0.12, 1.29]), but no evidence of an effect for those who had used the Static-99R in practice ($d = 0.00$, 95% CI = [–0.12, 0.12]). However, there was no evidence of a similar interaction for Static-99R scores from other cases, or for PCL-R scores from any case. In short, we could find no variables that seemed to explain the allegiance effects we observed overall.

Discussion

Results from this study underscore recent concerns about forensic sciences (NRC, 2009)—and raise concerns specific to forensic psychology—by demonstrating that some experts who score ostensibly objective assessment instruments assign scores that are biased toward the side that retained them. In the field, some apparent adversarial allegiance may result from selection effects (i.e., a savvy attorney selects experts who are predisposed to the attorney's perspective or presents input only from experts who favor the attorney's perspective), but our results suggest that even without selection effects, the pull of adversarial proceedings tends to influence opinions by paid forensic experts.

Of course, there was considerable variability in scores even from evaluators assigned to the same side, and certainly not every evaluator produced scores consistent with adversarial allegiance. But the systematic score differences among opposing experts could not be explained by chance, random measurement error, or preexisting differences between the experimental groups.

This evidence of allegiance was particularly striking because our experimental manipulation was less powerful than the forces experts are likely to encounter in most real cases. For example, our participants spent only about 15 min with the retaining attorney, whereas experts in the field may have extensive contact with retaining attorneys over weeks or months. Our participants formed opinions on the basis of files only, and they all reviewed identical files, whereas experts in the field may elicit different information by seeking different collateral sources or interviewing offenders in different ways. Therefore, the pull toward allegiance in this study was relatively weak compared with the pull typical of most cases in the field. Consequently, the large group differences provide compelling evidence for adversarial allegiance.

Our study could not identify the mechanisms responsible for the allegiance effect. We do not know whether the effect was more attributable to the initial conversation with an attorney, a sense of team loyalty, the monetary payment, or the promise of future work. We do not know the role of confirmation bias, anchoring, or other potentially important cognitive mechanisms. Of course, the role of each mechanism may have varied by participant, and not all participants demonstrated an allegiance effect. Future research is needed to disentangle the roles of these mechanisms and to identify evaluator characteristics that are associated with adversarial allegiance.

Although this study addressed only one kind of evaluation (i.e., assessment of risk for sexual recidivism), there is little reason to believe that this is the only kind of forensic psychological evaluation or forensic-science procedure vulnerable to allegiance effects. Indeed, the evidence of allegiance effects in the case of structured, ostensibly objective instruments that usually reveal strong interrater agreement leaves us even more concerned about the possibility of allegiance effects in the case of procedures that are less structured or less guided by scoring rules. Many forensic-science procedures rely heavily on subjective judgment (e.g., matching bite marks, hair fibers, or tire treads; NRC, 2009), as do many opinions psychologists offer in court (e.g., assigning diagnoses or assessing emotional injury). Our findings underscore the need for research on the cognitive and procedural biases that may facilitate adversarial allegiance, as well as the need for research on potential interventions to reduce allegiance. Indeed, our findings suggest that there may be opportunities to improve forensic psychological practice, broader forensic-science practice, and even legal policy and procedures in ways that might better promote scientific objectivity and reduce adversarial allegiance.

Author Contributions

D. C. Murrie and M. T. Boccaccini designed the study. D. C. Murrie drafted the introduction and the Method and Discussion sections. M. T. Boccaccini performed data analysis and drafted the Results section. L. A. Guarnera and D. C. Murrie had primary responsibility for arranging and overseeing the experiment, whereas M. T. Boccaccini and K. A. Rufino collected and coded the data. All authors reviewed and edited the final manuscript submitted for publication.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This research was supported by the National Science Foundation Law & Social Science Program (Award SES 0961082).

Supplemental Material

Additional supporting information may be found at http://pss.sagepub.com/content/by/supplemental-data

Notes

1. Of the 10 clinicians who failed to return for scoring case files, most explained that they were absent because they had been called to court to provide testimony as part of their professional practice.
2. To reduce the possibility of researchers' expectations influencing the results, we kept three of the four researchers blind to participants' assignment to conditions (inevitably, the third author, who managed the random assignment, was aware).
3. The order of administration was randomized for three of the four cases. Pilot testing suggested that most participants would be able to score three files in one day, but that some might be unable to complete four. Therefore, we provided the first three offender files to participants in a randomized order, to ensure that we would have similar, sufficient *n*s for these three cases. A fourth case was provided to all participants last, with the understanding that time constraints might preclude many participants from completing it.
4. We included this unusual case for exploratory purposes because we hypothesized that there may be some floor effect to adversarial allegiance. That is, we wondered whether some offenders might be so low in psychopathy that evaluators would score these offenders similarly regardless of the side that retained them. This seemed to be the case. However, because this exploratory case was the last file provided to participants (see note 3), and was completed by fewer participants than the other cases were (see Table 1), it is conceivable that some of the difference in results was attributable to these other factors.

References

Barefoot v. Estelle, 463 U.S. 880 (1983).

Bernstein, D. E. (2008). Expert witnesses, adversarial bias, and the (partial) failure of the Daubert Revolution. *Iowa Law Review*, *93*, 101–137.

DeMatteo, D., & Edens, J. F. (2006). The role and relevance of the Psychopathy Checklist-Revised in court: A case law survey of U.S. courts (1991-2004). *Psychology, Public Policy, and Law*, *12*, 214–241. doi:10.1037/1076-8971.12.2.214

Dror, I. E., & Cole, S. A. (2010). The visit in "blind" justice: Expert perception, judgment, and visual cognition in forensic pattern recognition. *Psychonomic Bulletin & Review*, *17*, 161–167. doi:10.3758/PBR.17.2.161

Dror, I. E., & Hampikian, G. (2011). Subjectivity and bias in forensic DNA mixture interpretation. *Science and Justice*, *51*, 204–208. doi:10.1016/j.scijus.2011.08.004

Gacono, C., & Hutton, H. (1994). Suggestions for the clinical and forensic use of the Hare Psychopathy Checklist–Revised (PCL-R). *International Journal of Law and Psychiatry*, *17*, 303–317.

Guarnera, L. A., Murrie, D. C., Boccaccini, M. T., & Rufino, K. (2012, March). *Do attitudes affect psychopathy scores evaluators assign to sexual offenders in Sexually Violent Predator proceedings?* Paper presented at the annual meeting

of the American Psychology-Law Society, San Juan, Puerto Rico.

Hand, L. (1901). Historical and practical considerations regarding expert testimony. *Harvard Law Review*, *15*, 40–58.

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis. *Psychological Assessment*, *21*, 1–21. doi:10.1037/a0014421

Hanson, R. K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior*, *24*, 119–136. doi:10.1023/A:1005482921333

Hare, R. D. (2003). *The Hare Psychopathy Checklist–Revised: Second edition*. Toronto, Ontario, Canada: Multi-Health Systems.

Hawes, S. W., Boccaccini, M. T., & Murrie, D. C. (2013). Psychopathy and the combination of psychopathy and sexual deviance as predictors of sexual recidivism: Meta-analytic findings using the Psychopathy Checklist—Revised. *Psychological Assessment*, *25*, 233–243. doi:10.1037/a0030391

Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2012). Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders: Revised age weights. *Sexual Abuse: Journal of Research and Treatment*, *24*, 64–101. doi:10.1177/1079063211409951

Kansas v. Hendricks, 521 U.S. 346 (1997).

Krafka, C., Dunn, M. A., Johnson, M. T., Cecil, J. S., & Miletich, D. (2002). Judge and attorney experiences, practices, and concerns regarding expert testimony in federal civil trials. *Psychology, Public Policy, and Law*, *8*, 309–332. doi:10.1037/1076-8971.8.3.309

Leistico, A. R., Salekin, R. T., DeCoster, J., & Rogers, R. (2008). A large-scale meta-analysis relating the Hare measures of psychopathy to antisocial conduct. *Law and Human Behavior*, *32*, 28–45. doi:10.1007/s10979-007-9096-6

Mnookin, J. (2008). Expert evidence, partisanship, and epistemic confidence. *Brooklyn Law Review*, *73*, 587–611.

Morey, L. C. (1991). *Personality Assessment Inventory: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Murrie, D. C., Boccaccini, M. T., Johnson, J. T., & Janke, C. (2008). Does interrater (dis)agreement on Psychopathy Checklist scores in sexually violent predator trials suggest partisan allegiance in forensic evaluations? *Law and Human Behavior*, *32*, 352–362. doi:10.1007/s10979-007-9097-5

Murrie, D. C., Boccaccini, M. T., Turner, D., Meeks, M., Woods, C., & Tussey, C. (2009). Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation? *Psychology, Public Policy, and Law*, *15*, 19–53. doi:10.1037/a0014897

National Research Council, Committee on Identifying the Needs of the Forensic Science Community. (2009). *Strengthening forensic science in the United States: A path forward*. Washington, DC: National Academies Press.

Rufino, K. A., Boccaccini, M. T., & Guy, L. S. (2011). Scoring subjectivity and item performance on measures used to assess violence risk: The PCL-R and HCR-20 as exemplars. *Assessment*, *18*, 453–463. doi:10.1177/1073191110378482

Rufino, K. A., Boccaccini, M. T., Hawes, S., & Murrie, D. C. (2012). When experts disagreed, who was correct? A comparison of PCL-R scores from independent raters and opposing forensic experts. *Law and Human Behavior*, *36*, 527–531. doi:10.1037/h0093988

Skeem, J., Polaschek, D., Patrick, C., & Lilienfeld, S. (2011). Psychopathic personality: Bridging the gap between scientific evidence and public policy. *Psychological Science in the Public Interest*, *12*, 95–162. doi:10.1177/1529100611426706

Static-99 Clearinghouse. (n.d.). *Static-99/Static-99R*. Retrieved from http://www.static99.org/

Wigmore, J. (1923). *A treatise on the Anglo–American system of evidence in trials at common law: Including the statutes and judicial decisions of all jurisdictions of the United States and Canada*. Boston, MA: Little, Brown.

Psychology, Public Policy, and Law
2009, Vol. 15, No. 1, 19–53

1076-8971/09/$12.00 DOI: 10.1037/a0014897

RATER (DIS)AGREEMENT ON RISK ASSESSMENT MEASURES IN SEXUALLY VIOLENT PREDATOR PROCEEDINGS
Evidence of Adversarial Allegiance in Forensic Evaluation?

Daniel C. Murrie
University of Virginia

Marcus T. Boccaccini, Darrel B. Turner, Meredith Meeks, and Carol Woods
Sam Houston State University

Chriscelyn Tussey
University of Virginia

Actuarial risk assessment measures are often admitted in court, partly because strong psychometric properties such as interrater agreement suggest that they increase reliability and reduce subjectivity in forensic evaluation. But how strong is rater agreement when raters are retained by opposing sides in adversarial legal proceedings? The authors review sexual offender civil commitment cases in which opposing evaluators reported scores on the STATIC-99, the Minnesota Sex Offender Sex Offender Screening Tool—Revised (MnSOST–R), or the Psychopathy Checklist—Revised (PCL–R) for the same individual. Differences between scores from opposing evaluators were often greater than expected based on rater agreement values reported in the instrument manuals and research literature. Score differences were often in a direction that supported the party who retained each evaluator. Rater agreement was stronger for the STATIC-99, intraclass correlation coefficient ([ICC]A,1) = .64; than for the MnSOST–R, ICC(A,1) = .48; and the PCL–R, ICC(A,1) = .42. STATIC-99 scores appeared less influenced by adversarial allegiance. Overall, however, results raise concern that an evaluator's adversarial allegiance could influence some assessment instrument scores in forensic evaluation.

Keywords: sexually violent predator, sex offender civil commitment, allegiance, bias, psychopathy

In recent years, public concern about sexual offenders has prompted states to adopt a variety of laws and policies, including postincarceration civil commitment of sexually violent predators, that attempt to protect the community from high-risk sexual offenders (LaFond, 2005). Of course, distin-

Daniel C. Murrie and Chriscelyn Tussey, Institute of Law, Psychiatry and Public Policy, University of Virginia School of Medicine; Marcus T. Boccaccini, Darrel B. Turner, Meredith Meeks, and Carol Woods, Department of Psychology, Sam Houston State University.

We gratefully acknowledge the Special Prosecution Unit for their assistance in this study specifically and for their commitment to external academic research generally. Any opinions conveyed in this article are those of the authors and do not necessarily reflect the opinions of the Special Prosecution Unit.

Correspondence concerning this article should be addressed to Daniel Murrie, Institute of Law, Psychiatry and Public Policy, University of Virginia School of Medicine, P. O. Box 800660, Charlottesville, VA 22908-0660. E-mail: Murrie@Virginia.edu

guishing offenders at relatively higher or lower risk requires some form of reliable risk assessment. Therefore, clinicians and administrators in forensic and correctional settings increasingly rely on actuarial risk assessment instruments (ARAIs; Janus & Prentky, 2003) designed to estimate the risk of recidivism among sexual offenders. National surveys reveal that the vast majority of states use a sex-offender-specific ARAI at some point in sex offender supervision (Interstate Commission for Adult Offender Supervision [ICAOS], 2007). At least 30 states reported using the STATIC-99 (Hanson & Thornton, 1999) specifically. Another popular sex offender risk measure, the Minnesota Sex Offender Sex Offender Screening Tool—Revised ([MnSOST–R]; Epperson et al., 1998), has been adopted by 7 state systems (ICAOS, 2007) and more than 20% of sex offender treatment programs in the United States (McGrath, Cumming, & Bouchard, 2003).

Using ARAIs not only is common practice but also is sometimes mandated by law. The Virginia statute (Va. Code. Ann. § 37.2-903) delineating procedures related to civil commitment of certain sexual offenders as sexually violent predators ([SVPs] a process described later) specifically requires the Department of Corrections to administer to all sexual offenders a particular ARAI (i.e., STATIC-99; Hanson & Thornton, 1999) and refer for a subsequent clinical evaluation any inmates who score above a certain total. Indeed, there appear to be few, if any, psychological assessment instruments more ingrained into law and policy than sex-offender-specific ARAIs.

Generally, actuarial risk assessment relies on explicit rules that specify which risk factors are examined, how those risk factors are scored, and how the scores are mathematically combined to yield an objective estimate of risk (Monahan, 2006). ARAIs for sex offender recidivism, such as the STATIC-99 and the MnSOST–R, were developed by following samples of released sexual offenders and documenting the observed recidivism rates. Researchers identified risk factors (usually data easily retrieved from records, such as age and previous offenses) that were statistically related with recidivism. They also documented recidivism rates among subgroups of the sample with specific numbers of risk factors (e.g., of offenders with X of the identified risk factors, Y% reoffended over Z years). Thus, the premise of ARAIs is that clinicians can observe the number of predefined risk factors present in the offender they evaluate and estimate the likelihood that the offender will recidivate on the basis of the observed recidivism rate in the risk measure's development sample.

In many respects, the movement to create and adopt ARAIs is a positive development. ARAIs tend to yield more accurate risk estimates than unstructured clinical judgments (Grove & Meehl, 1996; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Hanson, Morton, & Harris, 2003; Hanson, & Morton-Bourgon, 2007), and they provide a clear basis for decisions that may have dire consequences for individual offenders and potential victims in the community. One potential advantage of actuarial risk measures is that they should reduce clinician subjectivity in risk judgments and thus increase the likelihood of a fair and objective forensic evaluation (Janus & Prentky, 2003). The perception that actuarial risk measures reduce bias and subjectivity is particularly important in adversarial legal contexts such as SVP proceedings, in

which opposing sides retain forensic evaluators to conduct a formal risk assessment of the same individual.

Actuarial Risk Assessment Instruments in Sex Offender Civil Commitment

SVP laws allow states to identify sexual offenders perceived to be at high risk for repeated sexual offenses and civilly commit them after their incarceration to protect potential victims and provide treatment to the offender (for detailed descriptions of forensic evaluations in SVP proceedings, see Campbell, 2007; Doren, 2002; Jackson & Richards, 2008; Miller, Amenta, & Conroy, 2005). Most SVP commitment laws mirror criteria that the U.S. Supreme Court set forth in *Kansas v. Hendricks* (Kansas v. Hendricks, 117 S. Ct. 2072 (1997)) and require four elements for commitment. Usually, the committed individual must (a) have a history of sexual offending, (b) manifest a mental abnormality (sometimes defined as a mental disorder or personality disorder), (c) manifest a volitional impairment rendering him less able to control his sexual behavior, and d) pose significant risk for future sexual offending (Miller et al., 2005).

Proceedings for offenders facing civil commitment as SVPs routinely involve forensic evaluators, retained by opposing sides, who conduct risk assessments of the same offender, often using the same ARAIs to do so. ARAIs play a prominent role in sex offender civil commitment proceedings because they are the primary method by which evaluators assess the criterion of risk of future offending (Wollert, 2006). One national survey revealed that nearly all surveyed SVP evaluators administered ARAIs to assess risk of sexual recidivism (Jackson & Hess, 2007). Likewise, guidelines from professional groups (Association for the Treatment of Sex Abusers, 2001) and professional texts (Doren, 2002; Jackson & Richards, 2008) recommend that evaluators use ARAIs. ARAIs are routinely admitted as evidence during SVP trials (Janus & Prentky, 2003), because they reflect current practice standards and because there appears to be a general consensus that they have a strong empirical basis and adequate psychometric properties such as interrater reliability (Doren, 2002, 2006; but cf. Campbell, 2007).

In Florida, Levenson (2004a) found that offenders whom evaluators recommended for civil commitment had significantly higher scores on ARAIs than offenders who were evaluated, but not recommended, for commitment. For example, offenders recommended for commitment had a mean total MnSOST–R score of 10, compared with the released offenders, who had a mean score of 3. The effect size for this difference is large (Cohen's $d = 1.47$),[1] with similarly large effects for the STATIC-99 ($d = 1.31$) and Hare's (1991, 2003) Psychopathy Checklist—Revised (PCL–R; $d = 1.06$), another measure that is also often used in sex offender civil commitment evaluations. These three measures were the strongest predictors of whether evaluators recommended civil commitment, outpredicting offender and offense characteristics such as victim age, number of victims, and number of sex offense arrests. In short, although ARAIs are rarely the

[1] Effect sizes were calculated from descriptive statistics from Levenson (2004a; see Table 1).

sole criteria for initiating civil commitment proceedings, ARAIs appear to play a substantial role in influencing who is civilly committed.

Actuarial Risk Assessment and Rater Agreement

Overall, research studies report strong rater agreement values for ARAIs (Doren, 2004, 2006). For example, the STATIC-99 manual summarized good agreement values, although using three different measures of interrater agreement. Regarding intraclass correlation coefficients (ICCs) specifically, researchers have reported values ranging from .85 to .90 (Barbaree, Seto, Langton, & Peacock, 2001; Hanson, 2001; Harris et al., 2003) for the STATIC-99. The MnSOST–R also appears to demonstrate adequate agreement. Indeed, the most recent MnSOST–R technical paper (Epperson et al., 2003) offers more detailed information than is typically available regarding rater agreement for ARAIs:

> In the Minnesota reliability study, the singular ICC for the 10 raters was .80 for consistency of ratings and .76 for absolute agreement of ratings, indicating that the ratings of individual raters were reasonably reliable, particularly give the harsh conditions for the raters. The Florida reliability study better reflected the conditions under which the MnSOST–R is typically scored in real-life situations, and this study yielded higher reliability coefficients: ICC = .87 for relative agreement of ratings and .86 for absolute agreement of ratings. (p. 23)

Independent researchers using the MnSOST–R have reported similar rater agreement values (e.g., an ICC of .80; Barbaree et al., 2001). Thus, nothing in the available research suggests that trained raters cannot achieve adequate levels of interrater agreement on ARAIs such as the STATIC-99 and the MnSOST–R (Doren, 2006).

Rater Agreement in Adversarial Legal Proceedings

Strong rater agreement values in research contexts are important because these demonstrate that clinicians *can* consider the same case information and similarly score an instrument. Strong rater agreement values in adversarial legal proceedings are perhaps even more important because these demonstrate whether clinicians indeed *do* consider the same case information and similarly score an instrument when doing so has important consequences. So how strong is rater agreement when evaluations are conducted to provide information for adversarial legal proceedings? Levenson (2004b) examined SVP evaluation reports in Florida for offenders who had been evaluated twice by petitioner-retained evaluators. She identified 281 offenders with two STATIC-99 scores, 224 with two MnSOST–R scores, and 69 with two PCL–R scores. The ICC for absolute agreement was .85 for both the STATIC-99 and the MnSOST–R, and it was .84 for the PCL–R. Levenson's findings suggest a high level of agreement in the field when evaluators work on the same side of the case.

Only one published study has addressed rater agreement on a forensic assessment instrument as scored by opposing evaluators in adversarial legal proceedings (Murrie, Boccaccini, Johnson, & Janke, 2008). This study addressed Hare's (1991, 2003) PCL–R, a measure widely used in forensic assessments of risk for violence and sexual violence (see, e.g., DeMatteo & Edens, 2006). This

clinician-administered measure relies on clinical interview and review of records. Clinicians score the measure by assigning a 0, 1, or 2 to each of 20 items. Although the PCL–R scoring manual provides specific guidance for scoring, there remains room for some subjectivity (Campbell, 2006) or inference (Hare, 2003), particularly on items addressing the offender's interpersonal and emotional style (e.g., superficial charm, lack of remorse).

Researchers (Murrie et al., 2008) examined 23 SVP trials in which opposing evaluators reported PCL–R total scores for the same individual. For the PCL–R total score, the single-evaluator ICC for absolute agreement was .39, which was well below the strong levels of agreement observed for the PCL–R in research contexts (usually above .85), the ICC of .84 reported by Levenson (2004b) for SVP evaluations performed for the same side, and published test–retest values for the PCL–R (approximately .60 over a 2-year period; Rutherford, Cacciola, Alternman, McKay, & Cook, 1999).

On average, there was a 7.81-point ($SD = 6.85$) difference in scores from opposing evaluators. Considering score differences with respect to the standard error of measurement (*SEM*; approximately ± 3 points for the PCL–R), difference scores were greater than 2 *SEM*s (i.e., >6.0) for 61% of the cases. Finally, score differences were usually in a direction that supported the party who retained their services. In other words, scores from the petitioner (prosecution) evaluator tended to be higher than scores from the respondent (defense) evaluator (Cohen's $d = 1.03$; Murrie et al., 2008). After examining several other potential explanatory factors—and finding that none were sufficient to account for the results—these score differences appeared best attributed to *adversarial allegiance*, or the pull for forensic evaluators in adversarial proceedings to reach opinions that support the party who retained them.

Although results from the study of opposing PCL–R scores in SVP trials reflect only a small sample of instrument scores from one type of forensic evaluation in one state, the study results raise provocative questions. For example, results heighten concerns about ethical practice and the evidentiary value of the PCL–R as administered by privately retained evaluators. For reasons related to both science and SVP policy, we might also wonder whether the influence of adversarial allegiance extends to other clinical decisions that involve some degree of subjective clinician judgment (e.g., diagnosis) or to other assessment instruments that presumably involve less subjective judgment in scoring.

Therefore, an important second step in investigating adversarial allegiance is to examine scores from opposing evaluators on the sex-offender-specific ARAIs. These instruments are often used to "screen in" or "screen out" offenders as suitable for further consideration under SVP civil commitment statutes (Va. Code. Ann. § 37.2-903), they appear to influence whom evaluators recommend for SVP civil commitment (Levenson, 2004a), and they are usually presented at SVP trials by expert witnesses. Therefore, a finding that scores on the ARAIs were influenced by the party who retained the evaluator would raise both ethical and practical questions.

However, there are several reasons why ARAIs such as the STATIC-99 and MnSOST–R may be less vulnerable than the PCL–R to any effects of adversarial allegiance. For example, because the PCL–R requires an interview—which is structured but allows some room for evaluator differences—it is conceivable that

evaluators may seek or elicit information differently, depending on whether they are retained by prosecution or defense. Conversely, the examinee might behave differently with a defense-retained evaluator versus a prosecution-retained evaluator, thereby contributing to the score differences reported by opposing evaluators. However, for ARAIs scored primarily from records, this possibility is less likely.

Probably more important is that the PCL–R includes several items (e.g., superficial charm, dishonesty, lack of empathy) that require an evaluator to draw inferences from an offender's behavior during an interview. These inferences, which inevitably involve some subjective judgment, might be subtly influenced by adversarial allegiance, as well as by other evaluator biases or idiosyncrasies (see Boccaccini, Turner, & Murrie, 2008).

In contrast to the PCL–R, the STATIC-99 and MnSOST–R primarily involve coding straightforward demographic and historical information (e.g., age, prior offense data) available through records alone. Indeed, both measures were developed by examining groups of known recidivists and nonrecidivists and coding information typically available from correctional records. The measures are often completed by correctional staff and do not require a clinical interview (although interviews are acceptable as additions to record review). For many of the items on these actuarial measures (e.g., age, offense victims), there seems to be little room for evaluators to disagree.

Purpose of the Present Study

We examined rater agreement for the STATIC-99 and the MnSOST–R, as administered by opposing evaluators in SVP proceedings. As in the study of PCL–R scores (Murrie et al., 2008), if we were to find that evaluators demonstrated rater agreement on the STATIC-99 and MnSOST–R that was similar to rater agreement in the research literature, we would suspect that actuarial risk measures are relatively immune to any influence of adversarial allegiance. If we were to find rater agreement values poorer than those reported in the literature yet observe score differences that were unsystematic, we would suspect a general lessening in rater agreement from research to real-world settings. Finally, if we were to find poorer rater agreement values than those reported in the literature and observe scores that systematically differed in a direction consistent with the opposing sides that retained the evaluators (i.e., petitioner's experts reported higher scores, whereas respondent's experts reported lower scores), we would suspect that adversarial allegiance played some part in the poorer rater agreement.

A second goal of examining rater agreement values for the STATIC-99 and MnSOST–R in SVP proceedings was to compare whether these measures reveal rater (dis)agreement values similar to those Murrie et al. (2008) reported for the PCL–R. We have supplemented the Murrie et al. (2008) PCL–R data with scores from more recent depositions and trials, which provided 12 new sets of opposing PCL–R scores, for a total of 35 cases with opposing scores. We report PCL–R findings from this larger sample here, which allows us to view rater agreement values for ARAIs alongside rater agreement values for the interview-based PCL–R. Had we sampled from another jurisdiction or another type of trial, it would not be clear whether differences in rater agreement for the PCL–R versus

the actuarial measures were attributable to differences in the instruments themselves or to differences in the context from which we sampled. Examining these three popular risk measures (i.e., the STATIC-99, the MnSOST–R, and the PCL–R) as administered in the same "population" of SVP trials allows us to better compare the degree to which each appears vulnerable to the effects of adversarial allegiance under the same circumstances.

If we were to find that rater agreement on ARAIs appears stronger in adversarial legal contexts than rater agreement for the PCL–R, this finding might suggest that relying on highly structured measures (i.e., those that minimize the role of interview and clinical inference) could reduce the influence of adversarial allegiance in forensic evaluation. However, if we were to find that opposing evaluators show poor rater agreement on highly structured actuarial measures also, this finding would suggest that structured assessment measures alone are not enough to minimize the influence of adversarial allegiance in forensic evaluation. To better facilitate these comparisons between the ARAIs and the PCL–R, we present updated analyses of PCL–R rater agreement, including the recent cases. We also conducted additional analyses using a generalizability theory framework to quantify the amount of variance in risk scores that was attributable to the side that retained the evaluator, as opposed the offenders being evaluated. In most studies, researchers assume that any variance in scores not captured by the rater agreement coefficient (ICC) is random error, which usually is not true. In this study, we used generalizability theory to quantify the proportion of variance in scores that was attributable to the side for which the evaluation was performed (petitioner or respondent).

Method

Context for the Present Study

Civil commitment proceedings for offenders facing commitment as SVPs provided the opportunity to examine scores from risk measures (i.e., the STATIC-99, MnSOST–R, and PCL–R) as administered by opposing evaluators in adversarial legal proceedings. In Texas, SVP procedures begin when a multidisciplinary team (MDT) receives notice that a sexual offender is within 16 months of scheduled release. The MDT determines whether the inmate has two qualifying sexual offenses and may then refer the inmate to the "the department," featuring representatives from state criminal justice and mental health agencies who commission an assessment for behavioral abnormality. These commissioned assessments occur on a contract basis with independent (i.e., not current employees of the correctional system) doctoral-level evaluators. To establish such a contract, evaluators must demonstrate relevant experience and training with sexual offender assessment and/or treatment, as well as qualification to administer popular risk measures (i.e., the STATIC-99, MnSOST–R, and the PCL–R). Contracted assessment reports typically describe a review of records, clinical interviews, a review of risk factors, and an overall risk estimate (see Amenta, 2005).

After the department reviews a completed, contracted evaluation, department staff decide whether the offender manifests a behavioral abnormality (typically pedophilia, another paraphilia, antisocial personality disorder, or psychopathy; Amenta, 2005). If so, they refer the offender to the Special Prosecution Unit, Civil

Division (SPU), which has typically had the resources to select approximately 15 offenders per year (about one fifth of the cases they review) for whom to initiate civil commitment proceedings. Again, the evaluations on which many of the decisions up to this point are based are not solicited directly by the petitioner[2] for purposes of trial. They are third-party evaluations that the state department of corrections used to screen possible candidates for referral to the SPU for potential civil commitment. However, evaluators understand that their evaluations and expert testimony may be required for those cases that proceed to trial. Evaluators probably also understand that the SPU considers the original evaluator's report as one source of data when deciding which offenders to pursue for civil commitment. During eventual civil commitment proceedings, it is the petitioner who calls the original evaluator to serve as a witness, and the petitioner uses the report from this evaluator as evidence in support of civil commitment. As the case moves toward trial, the petitioner sometimes retains additional expert evaluators (e.g., a psychiatrist or psychologist) who may perform another evaluation. Therefore, in some trials, the petitioner calls more than one expert to testify and may have more than one set of evaluation data (including scores from ARAIs).

Once the petitioner gives notice that they are initiating civil commitment proceedings against a particular inmate, the inmate (now considered a respondent in the civil commitment proceedings) secures defense counsel, which is almost always through a state-sponsored agency offering legal defense for indigent inmates. The defense counsel for the respondent typically arranges for a second-opinion evaluation by a mental health professional. Often, as in many legal contexts, defense counsel may invite more than one evaluator to review case materials and offer preliminary opinions before hiring an evaluator for the full evaluation. The resulting evaluations are defense evaluations in that the evaluators were retained by the respondent for the purpose of defending against civil commitment. Unlike the original evaluations, which always result in a written report, the respondent's evaluators rarely produce a written report. Rather, the evaluator usually presents findings (including ARAI scores) only in deposition and trial testimony. It is important to emphasize that both the original evaluator and the respondent-retained evaluator have access to essentially the same collateral materials. Both receive the same case file of correctional and law enforcement records, often including STATIC-99 and MnSOST–R protocols scored by correctional staff.

Data Sources

In November 2007, we collected offender information and risk scores from three types of documents. First, the SPU allowed access to its files for each offender they had pursued for civil commitment ($N = 72$). Each of these files contained transcripts of depositions from expert witnesses who had evaluated the offender. Second, we received permission to search a database of case information and risk scores for all offenders who had been referred to the SPU between September 1999 and September 2006, when the SPU stopped using this database.

[2] In civil commitment proceedings, *petitioner* is roughly analogous to prosecution and *respondent* is roughly analogous to defense.

This database contained STAIC-99, MnSOST–R, and PCL–R total scores from the initial petitioner evaluator for 64 of the 72 offenders. Finally, we reviewed trial transcripts provided by the SPU. Trial transcripts were only available for cases in which the committed offenders had filed an appeal (41 of 72 offenders).

Offender Sample

Of the 72 offenders pursued for commitment by the SPU, 80.6% ($n = 58$) were civilly committed by a jury, 6.9% ($n = 5$) opted for a bench trial and were civilly committed by a judge, and 8.3% ($n = 6$) did not go to trial because they agreed to the conditions of civil commitment. Risk scores were also available for the only offender (1.4%) who was found by a jury to *not* meet civil commitment and for 2 offenders (2.8%) whom the SPU began pursuing for commitment but then stopped proceedings. Offenders were all male and identified as Caucasian ($n = 42$, 58.3%), Hispanic ($n = 16$, 22.2%), or African American ($n = 14$, 19.4%).

The Results section provides detailed information about the extent to which instrument scores were available for offenders, including when offenders had multiple scores for the same instrument from opposing or nonopposing evaluators. Overall, 38 (52.8%) of the offenders had scores from both petitioner and respondent evaluators for at least one instrument. There were 23 (31.9%) offenders who had opposing evaluator scores for all three measures. Three offenders had opposing evaluator scores for the STATIC-99 and MnSOST–R only, 1 had opposing evaluator scores for the PCL–R and STATIC-99 only, 1 had opposing evaluator scores for the PCL–R and MnSOST–R only, and 10 had opposing evaluator scores for the PCL–R only.

Evaluator Sample

Twenty-one evaluators provided at least one risk measure score for 1 of the 72 offenders. All evaluators in the sample had a doctoral degree; 20 had degrees in psychology and 1 had an MD. Fifteen evaluators performed evaluations for the state; they performed from 1 to 19 each, with a mean of 5.67 ($SD = 6.30$) evaluators. Ten different evaluators performed evaluations for the respondent; they performed 1 to 16 each, with a mean of 5.10 ($SD = 5.02$) evaluators. Only 4 evaluators provided at least one evaluation for both the petitioner and respondent.

Measures

MnSOST–R. The MnSOST–R (Epperson et al., 1998) was developed in collaboration with the Minnesota Department of Corrections to assess risk for sexual reoffense among men who committed at least one sexual offense against an unrelated victim. The measure includes 16 items; 12 are scored on the basis of historical information typically available through files, and 4 are coded using information related to the offender's incarceration for his index offense. The instrument yields a score that can range from −14 to 30 and falls within one of three risk levels (Epperson et al., 2003), although some have also examined an arrangement with six risk levels (Barbaree et al., 2001). Hanson and Morton-Bourgon (2007) examined the MnSOST–R in their recent analysis of sexual recidivism prediction and found fairly strong predictive validity (mean Cohen's

$d = 0.72$; eight studies). However, others have emphasized limitations of the MnSOST–R, including problems with item selection and poor performance in samples with low base rates of reoffending (Vrieze & Grove, 2008; Wollert, 2002, 2003). Regarding interrater reliability in research settings, Epperson et al. (2003) reported ICC values (absolute agreement, single rating) from .76 to .86. However, there is no available research documenting interrater agreement in adversarial legal contexts.

STATIC-99. The STATIC-99 (Hanson & Thornton, 2003) is an actuarial risk measure comprising 10 items, most of which are risk factors that can be scored as present or absent. Total scores can range from 0 to 12, so that individuals are assigned to one of seven risk categories ranging from 0 (*lowest risk*) to ≥6 (*highest risk*). In a document created for the Canadian government, Hanson and Morton-Bourgon (2007) reported that, across more than 40 studies, STATIC-99 scores were moderately strong predictors of sexual recidivism specifically (mean Cohen's $d = 0.70$; 42 studies) and of violent recidivism (sexual or nonsexual) generally (mean Cohen's $d = 0.58$; 25 studies). A more recent meta-analysis of published studies found a similar overall effect for the STATIC-99, although effects from STATIC-99 authors tended to be larger than those from other researchers (Blair, Marcus, & Boccaccini, 2008). Regarding interrater reliability, researchers have reported ICCs of .85 to .90 (e.g., Barbaree et al., 2001; Hanson, 2001; Harris, Rice, et al., 2003). As with the MnSOST–R, we could find no research documenting interrater agreement in adversarial legal contexts.

PCL–R. The PCL–R (Hare, 1991, 2003) is a 20-item checklist that requires a review of records and a structured interview to complete. The rater assigns a score of 0 (*not present*), 1 (*possibly present*), or 2 (*definitely present*) to quantify the degree to which the interviewee manifests particular psychopathy criteria. Large-scale meta-analyses support the PCL–R's predictive relationship to a variety of antisocial behaviors (Leistico, Salekin, DeCoster, & Rogers, 2008), and sexual reoffense in particular (mean d of 0.25 in Hanson & Morton-Bourgon, 2005). Within research contexts, reliability values for the PCL–R are generally quite strong. The test manual (Hare, 2003) reports interrater reliability values for a single evaluator, ICC(1), ranging from .86 for male inmates to .88 for male forensic psychiatric patients.

Results

MnSOST–R

Of the 72 offenders, 49 had one MnSOST–R score from a petitioner expert, 17 had two MnSOST–R scores from petitioner experts, and 6 had no MnSOST–R score from a petitioner expert. A total of 27 cases had at least one MnSOST–R score from a petitioner expert and a respondent expert. Twenty-one offenders had one MnSOST–R score from a respondent expert, 6 had two MnSOST–R scores from respondent experts, and 45 had no MnSOST–R score from a respondent expert. There were 3 offenders who had MnSOST–R scores from two petitioner and two respondent experts.

Agreement between evaluators on the same side of the case. Table 1 provides rater agreement statistics for cases in which two evaluators working on

Table 1
Rater Agreement for Risk Scores When Two Evaluators From the Same Side of the Case Provide Separate Scores for the Same Offender

Measure and side	*n*	ICC(A,1)	95% Confidence		Mean difference (and *SD*)[a]
			Lower	Upper	
MnSOST–R					
Petitioner	17	.63	.24	.85	2.59 (3.54)
Respondent	6	.00	−.99	.78	4.00 (2.97)
STATIC-99					
Petitioner	15	.84	.60	.95	0.53 (0.83)
Respondent	6	.95	.74	.99	0.17 (0.41)
PCL–R					
Petitioner	13	.24	−.38	.69	6.37 (5.90)
Respondent	7	.88	.46	.98	2.57 (1.40)

Note. ICC(A,1) = intraclass correlation coefficient for a absolute agreement for a single rater; MnSOST–R = Minnesota Sex Offender Sex Offender Screening Tool—Revised; PCL–R = Psychopathy Checklist—Revised.
[a] Mean difference scores based on the absolute value of the difference in scores between the two raters.

the same side of the case reported separate MnSOST–R scores. The single-evaluator, absolute agreement ICC(A,1) was in the moderate-to-poor range for the 17 cases with multiple-petitioner MnSOST–R scores, ICC(A,1) = .63, and extremely low for the six cases with two respondent scores, ICC(A,1) = .00. The average difference in scores for evaluators on the same side of the case was 2.59 for petitioner experts and 4.00 for respondent experts. The largest difference between petitioner experts was 13.00 points, although 6 of 17 cases featured perfect agreement between evaluators. The largest difference between respondent experts was 8.00 points, with perfect agreement in none of the cases.

Minimum, maximum, and average discrepancy datasets. Of the 27 cases with MnSOST–R scores from both petitioner and respondent experts, 8 had two petitioner MnSOST–R scores and 6 had two respondent MnSOST–R scores. There were three cases with two petitioner and two respondent MnSOST–R scores. Thus, 11 of these 27 cases had at least three MnSOST–R scores for the same offender. We considered three strategies for collapsing these data so that we could calculate agreement using all 27 offenders, with one petitioner score and one respondent score per offender. First, we selected the pair of scores (one state, one respondent) that would result in the smallest absolute difference between the two scores (minimum discrepancy). Second, we selected the pair of scores that would result in the largest absolute difference between the two scores (maximum discrepancy). Third, we used the average of the two scores from the same side for all cases with two scores per side (average discrepancy).

We present agreement analysis results separately for each of these three approaches to examining discrepancy. Our rationale for reporting effects for these three approaches was that no one approach can be singled out as providing the most representative measure of agreement on the instrument. Although the average discrepancy approach may appear to be the least biased, it is also the least ecologically valid. In reality, results from multiple evaluators are never combined

in court to produce a single averaged score. The minimum and maximum dataset analyses provide lower and upper limits for agreement in the dataset, with the average discrepancy values falling in between.

Agreement between opposing evaluators: Mean scores. Our first approach for examining rater agreement was to consider whether there was a statistically significant difference in MnSOST–R scores from petitioner and respondent experts in the 27 cases with opposing scores. Results from these analyses are summarized in Table 2. MnSOST–R scores from petitioner experts were significantly higher than those from respondent experts, regardless of how we handled cases with multiple scores from experts on the same side of the case. Effect sizes (Cohen's *d*) for these comparisons tended to be in the moderate to strong range (*d*s = 0.70 to 0.95) and in the direction expected by adversarial allegiance (higher scores from petitioner experts and lower scores from respondent-retained experts).

Agreement between opposing evaluators: Difference scores. Our second approach for examining rater agreement was to consider whether scores from petitioner and respondent experts tended to differ to a greater extent that we would expect according to the *SEM*. We calculated a difference score for each offender by subtracting the respondent expert's score from the petitioner expert's score. Difference scores with a positive value indicated that the petitioner expert assigned a higher MnSOST–R score, whereas a negative score indicated that the respondent's expert assigned a higher score.

Assuming a normal distribution, approximately 98% of difference scores between two evaluators should fall within 2 *SEM*s of one another. Although difference scores greater than 2 *SEM*s apart should be unusual, they might be

Table 2
Difference in Risk Measure Scores From State and Respondent Experts

Measure and sample	*M* (and *SD*) for: Petitioner	Respondent	*t*	Cohen's *d*
MnSOST–R (n = 27)				
Minimum discrepancy	8.00 (4.32)	5.81 (3.90)	4.05**	0.70
Maximum discrepancy	9.10 (4.57)	4.93 (4.25)	5.14**	0.95
Average discrepancy	8.90 (4.32)	5.37 (3.94)	4.97**	0.85
STATIC-99 (n = 27)				
Minimum discrepancy	4.74 (1.29)	4.30 (1.70)	1.85	0.29
Maximum discrepancy	4.85 (1.43)	4.25 (1.73)	2.21*	0.37
Average discrepancy	4.80 (1.33)	4.28 (1.71)	2.09*	0.34
PCL–R (n = 35)				
Minimum discrepancy	23.61 (80.8)	18.63 (6.49)	4.05**	0.68
Maximum discrepancy	24.89 (9.01)	18.29 (6.65)	5.08**	0.83
Average discrepancy	24.25 (8.23)	18.46 (6.54)	4.83**	0.78

Note. Minimum, maximum, and average discrepancies refer to how the discrepancy score was calculated when multiple scores were assigned and two scores were available from the same side of the case. Degrees of freedom for paired samples *t* tests are as follows: PCL–R = 34, STATIC-99 = 26, and MnSOST–R = 26. Cohen's *d*s for paired samples *t* tests were calculated using procedures recommended by Dunlap, Cortina, Vaslow, and Burke (1996). MnSOST–R = Minnesota Sex Offender Sex Offender Screening Tool—Revised; PCL–R = Psychopathy Checklist—Revised.
* $p < .05$. ** $p < .01$.

tolerable if about half reflected higher scores from the respondent and half reflected higher scores from the petitioner. However, a consistent pattern of difference scores greater than 2 *SEM*s, and in the same direction, would suggest adversarial allegiance.

The MnSOST–R technical manual does not report an *SEM* value for the measure's total score. Moreover, the manual does not report a standard deviation for the total score, making it impossible to calculate an *SEM* value from the manual. However, Langton et al. (2007) reported both a rater agreement coefficient (.83) and standard deviation value (5.60) for the MnSOST–R total score in a sample of more than 350 offenders. We used these values to calculate an *SEM* of 2.30 for the MnSOST–R total score. We then examined the extent to which the scores from petitioner and respondent experts differed by more than 2 *SEM* units (4.60).

Using the dataset of minimum discrepancy scores, we found that the mean difference score was 2.89 ($SD = 3.70$), with values ranging from −3.00 to 10.00. Of the 27 difference scores, 10 (37.0%) were greater than 2 *SEM* units apart in the direction of adversarial allegiance, and none were greater than 2 *SEM* units in the opposite direction. Using the dataset of maximum discrepancy scores, we found that the mean difference score was 4.18 ($SD = 4.23$), with values ranging from −4.00 to 11.40. Of the 27 difference scores, more than half ($n = 14$; 51.8%) were 2 or more *SEM* units apart in the direction of adversarial allegiance, and none were greater than 2 *SEM* units in the opposite direction. With the dataset of average discrepancy scores, the mean difference score was 3.53 ($SD = 3.69$), with values ranging from −3.00 to 10.5. Of the 27 difference scores, 11 (40.7%) were 2 or more *SEM* units apart in the direction of adversarial allegiance, and none were greater than 2 *SEM* units in the opposite direction.

Agreement between opposing evaluators: ICCs. ICCs are the most commonly used metric for describing rater agreement from risk measures. When there are two risk scores for each offender, the ICC is the proportion of variance in the set of scores that is attributable to the people being evaluated. The remaining variance is considered to be error. In most studies, researchers assume that all of this error variance is due to random error, which usually is not true. In the present study, it is possible to quantify the proportion of variance that is attributable to the side for which the evaluation was performed (state or respondent). Specifically, we were able to use generalizability theory analyses (see Brennan, 2001; Shalverson & Webb, 1991) to quantify three sources of variance in our set of MnSOST–R scores: variance attributable to the individual being evaluated (ICC), to the side for which the evaluator was retained, and to other (nonspecified) sources of error.

Researchers can calculate different ICCs depending on whether the researcher is interested in absolute agreement or consensus agreement (see McGraw & Wong, 1996). Coefficients for consensus are only concerned with covariation in scores. Consensus coefficients consider whether the evaluators generally agree about who warrants higher scores and who warrants lower scores, but the absolute values of the scores do not matter. Therefore, an evaluator who assigned a MnSOST–R score of 0 to Offender A and 5 to Offender B would show a high level of consensus agreement with an evaluator who assigned a MnSOST–R score of 10 to Offender A and 15 to Offender B, although these two evaluators assigned

very different scores to the same offender. Coefficients for absolute agreement consider both covariation and the specific value of the test score to be important for gauging agreement.[3] Differences in the specific value of the score are considered to be error in the calculation of absolute agreement coefficients. It is important to use the absolute agreement coefficients when measuring rater agreement on risk measures, because differences in specific scores become important in court (Murrie et al., 2008). Specific scores (or ranges of scores) from actuarial measures are used to identify the probability of an offender reoffending. A score difference of 1 point may have important practical results. For example, the offender may be screened in or out in systems that use specific cutoff scores for civil commitment screening approaches (e.g., Va. Code. Ann. § 37.2-903), or an evaluator may report to court a different level of risk and a different probability of reoffending, on the basis of a 1-point score difference.

ICCs can also be calculated for a single rater or for multiple raters. However, the multiple-rater ICC is only appropriate when scores from all evaluators are averaged or combined together for decision-making purposes. In court, scores from multiple evaluators are not presented to the decision maker as an average, especially when scores come from opposing evaluators. Thus, the ICCs reported in this study are for absolute agreement and a single rater, ICC(A,1).

We used the SPSS 15 VARCOMP procedure to estimate the proportion of variance in opposing MnSOST–R scores that was attributable to the offender, the side for which the evaluator performed the evaluation, and for other sources of error. Both offenders and evaluators were treated as random effects in the analysis of variance model that was used to estimate the variance components.[4] The ICC calculated using this method is identical to the ICC reported by SPSS under scale analysis. The advantage of the VARCOMP approach is that it also allows for calculating the effects for other sources of variance.

Results from the generalizability theory analyses are summarized in Table 3. ICCs(A,1) ranged from .38 to .48 for the MnSOST–R. These values indicate that less than half of the variance in the set of MnSOST–R scores could be attributed to the offenders' true level of risk as measured by the MnSOST–R.[5]

Of the remaining variance, anywhere from 19% to 30% was attributable to the side for which the evaluation was performed. In the maximum discrepancy dataset, nearly as much variance was attributable to the side for which the

[3] The coefficient calculated for absolute agreement is often referred to as the *index of dependability* in generalizability theory analyses (Brennan, 2001), although it is often reported as an ICC (e.g., in SPSS).

[4] The side for which the evaluation was performed could be treated as a random or fixed effect in the generalizability theory analyses. The ICCs calculated by SPSS for a two-way random effects model and a two-way mixed effects model are identical; what differs is how the coefficient is interpreted (Norusis, 2003). Our reasons for treating evaluators as a random effect were (a) to ensure that evaluator differences were considered to be error and (b) because there are other types of adversarial situations to which these findings may apply, such as those involving true prosecution experts, court-appointed experts, or treating experts.

[5] We use the term *true score* here for convenience, but we point out that ICCs are perhaps best understood in a generalizability theory framework, which focuses on universe scores, as opposed to a more restrictive classical test theory approach, which focuses on true scores (see Brennan, 2001; Shalverson & Webb, 1991; Shalverson, Webb, & Rowley, 1989).

Table 3
Intraclass Correlations (ICCs) for Opposing Evaluator Risk Scores and the Proportion of Variance Attributable to Adversarial Allegiance

Measure and sample	ICC(A,1)	Proportion of variance (%) attributable to:		
		Offender	Side of case	Other error
MnSOST–R (n = 27)				
Minimum discrepancy	.48	48	19	33
Maximum discrepancy	.38	38	30	32
Average discrepancy	.44	44	26	30
STATIC-99 (n = 27)				
Minimum discrepancy	.64	64	3	33
Maximum discrepancy	.58	58	5	37
Average discrepancy	.62	62	4	34
PCL–R (n = 35)				
Minimum discrepancy	.42	42	18	40
Maximum discrepancy	.40	40	25	35
Average discrepancy	.42	42	23	35

Note. ICC(A,1) = intraclass correlation coefficient for a single rater and absolute agreement (see McGraw & Wong, 1996). Minimum, maximum, and average discrepancies refer to how the score was assigned when multiple scores were assigned and two scores were available for the same side of the case. MnSOST–R = Minnesota Sex Offender Sex Offender Screening Tool—Revised; PCL–R = Psychopathy Checklist—Revised.

evaluation was performed (30%) as was attributable to the offender's risk level, as measured by the instrument (38%). Finally, anywhere from 30% to 33% of the variance in scores was attributable to unmeasured or random sources of error. Unexplained variance is provided in the analyses through an interaction term (Person × Evaluator). This variance can be thought of as a combination of systematic error variance that cannot be estimated because of the study design (e.g., variance attributable to individual psychologist) and random measurement error. Factors contributing to this value could be differences among the individual evaluators (e.g., training, experience, or methods of interpreting data from records), variability in the information available to score the measure (although this possibility is unlikely in our sample, in which both petitioner and respondent evaluators receive the same record base to review), or other sources of systematic or random measurement error.

Do selection effects account for low levels of agreement in cases with opposing scores? The fact that only 27 of the 72 offenders had opposing MnSOST–R scores raises concerns about the representativeness of those 27 offenders. One possible explanation for the low levels of agreement that we observed might be that these 27 cases were the ones in which the initial petitioner evaluation scores were unusually high. Did respondent-retained evaluators perhaps only administer the MnSOST–R evaluation when they perceived that the score from the petitioner's evaluator was excessively high? Because most of the 72 offenders had an initial petitioner MnSOST–R score, it was possible to examine this potential explanation by comparing the petitioner evaluator MnSOST–R scores for offenders who did and did not have a respondent

MnSOST–R scores. If the 27 offenders represented a select group of offenders with unusually high MnSOST–R scores, then petitioner MnSOST–R scores should be higher for those who underwent a respondent evaluation compared with those who did not.

Independent samples *t* tests revealed statistically significant differences in MnSOST–R scores, but in the opposite direction of the pattern expected. The 27 offenders with both petitioner and respondent MnSOST–R scores had significantly lower MnSOST–R scores from the initial petitioner evaluators than did the 39 offenders without a respondent evaluation. This finding applied to all three sets of discrepancy scores: For minimum discrepancy, $t(64) = 1.91$, $p = .06$, $d = 0.51$; for maximum discrepancy, $t(64) = 2.15$, $p = .04$, $d = 0.54$; and for average discrepancy, $t(64) = 2.11$, $p = .04$, $d = 0.53$. Thus, it seems unlikely that selection effects (i.e., unusually high MnSOST–R scores from original petitioner evaluators) accounted for the score differences we observed between opposing evaluators.

STATIC-99

Of the 72 offenders, 51 had one STATIC-99 score from a petitioner expert, 15 had two STATIC-99 scores from petitioner experts, and 6 had no STATIC-99 score from a petitioner expert. A total of 27 cases had at least one STATIC-99 score from a petitioner expert and a respondent expert. Twenty-one offenders had one STATIC-99 score from a respondent expert, 6 had two STATIC-99 scores from respondent experts, and 45 had no STATIC-99 score from a respondent expert. Two offenders had STATIC-99 scores from two petitioner and two respondent experts.

Agreement between evaluators on the same side of the case. Table 1 provides rater agreement statistics for cases in which two evaluators working on the same side of the case reported separate STATIC-99 scores. ICC(A,1) was strong between the two experts on the same side of the case for the 15 cases with two petitioner expert scores (.84) and 6 cases with two respondent expert scores (.95). The average difference in STATIC-99 scores for evaluators on the same side of the case was well below 1 point (.53 for petitioner experts, .17 for respondent experts). The largest difference between petitioner experts was 3.00 points, with perfect agreement in 9 of 12 cases. The largest difference between respondent experts was 1 point, with perfect agreement in the other five cases.

Agreement between opposing evaluators: Mean scores. Table 2 reports the results of paired-samples *t* tests examining whether there was a statistically significant difference in STATIC-99 scores from petitioner and respondent experts in the 27 cases with opposing scores. As with the MnSOST–R, we conducted analyses using separate sets of minimum, maximum, and average discrepancy scores for cases in which two STATIC-99 scores were available from experts on the same side of the case. Results from these analyses are summarized in Table 2. Across the 27 cases, differences scores tended to be in the direction expected by adversarial allegiance (higher scores from petitioner experts), but these differences were only moderate in size (Cohen's *d* range = .29 to .37) and

were only large enough to reach statistical significance in the maximum and average discrepancy datasets.

Agreement between opposing evaluators: Difference scores. Although the STATIC-99 authors do not report the *SEM* for the measure, they do report rater agreement coefficients and standard deviation values, which can be used to calculate an *SEM* value. Harris, Phenix, Hanson, and Thorton (2003; revised scoring procedures) reported an ICC of .87 for the STATIC-99 total score. Hanson and Thorton (2003; the STATIC-2002 report) reported that the standard deviation for the STATIC-99 across nine samples and more than 2,000 offenders is 1.9. With these two values, the *SEM* for the STATIC-99 total score is 0.68 points. We then examined the extent to which the scores from petitioner and respondent experts differed by more than 2 *SEM* units (1.37). We calculated a difference score for each offender by subtracting the respondent expert's score from the petitioner expert's score. Difference scores with a positive value indicated that the petitioner expert assigned a higher STATIC-99 score, whereas a negative score indicated that the respondent's expert assigned a higher score.

With the minimum discrepancy scores, we found that the mean difference score was 0.44 ($SD = 1.25$), with values ranging from −3.00 to 4.00. Of the 27 difference scores, 4 (14.8%) were greater than 2 *SEM* units apart in the direction of adversarial bias, and only 1 (3.7%) was greater than 2 *SEM* units in the opposite direction (score of −3.00). Using maximum discrepancy scores, we found a mean difference score of 0.59 ($SD = 1.39$), with values ranging from −3.00 to 4.00. Of the 27 difference scores, 5 (18.5%) were greater than 2 *SEM* units apart in the direction of adversarial allegiance, and only 1 was greater than 2 *SEM* units in the opposite direction. Using the average discrepancy scores, we found a mean difference score of 0.52 ($SD = 1.29$), with values ranging from −3.00 to 4.00. Of the 27 difference scores, 4 (14.8%) were greater than 2 *SEM* units apart in the direction of adversarial bias, and only 1 was greater than 2 *SEM* units in the opposite direction.

Agreement between opposing evaluators: ICCs. ICC(A,1) was in the .60 range for the STATIC-99, regardless of how we dealt with cases in which there were multiple scores from evaluators on the same side of the case. These values indicate that slightly more than half of the variance in the set of STATIC-99 scores could be attributed to the offenders' true level of risk as measured by the STATIC-99. In contrast to the other measure we examined, only 3% to 5% of the variance in STATIC-99 scores was attributable to the side for which the evaluation was performed. Finally, 33% to 37% of the variance in scores was attributable to other sources of error.

Do selection effects account for low levels of agreement in cases with opposing scores? As in the analyses of the MnSOST–R, we considered the possibility that the 27 cases for which we had opposing scores might systematically differ from the other cases in which opposing scores were not available. Perhaps respondent-retained evaluators only administered the STATIC-99 when they suspected that the score reported by the petitioner-retained evaluator was inappropriately high.

Therefore, we used independent samples *t* tests to compare the initial petitioner STATIC-99 scores for offenders who did have a STATIC-99 score from a

respondent expert to those that did not have a STATIC-99 score from a respondent expert. As with the MnSOST–R, findings from the independent samples *t* tests revealed that the 27 offenders with both a petitioner and a respondent STATIC-99 score had significantly lower STATIC-99 scores from the initial petitioner evaluators than the 39 offenders without a respondent evaluation. This finding applied to all three sets of discrepancy scores: For minimum discrepancy, $t(64) = 2.38$, $p = .02$, $d = 0.59$; for maximum discrepancy, $t(64) = 2.40$, $p = .02$, $d = 0.60$; and for average discrepancy, $t(64) = 2.43$, $p = .02$, $d = 0.61$. In other words, there was no evidence to support the hypothesis that rater disagreement between opposing evaluators was due to a selection effect in which respondent evaluators administered the STATIC-99 because the petitioner evaluator's score appeared unusually high.

PCL–R

Of the 72 offenders, 55 had one PCL–R score from a petitioner expert, 13 had two PCL–R scores from petitioner experts, and 4 had no PCL–R score from a petitioner expert. A total of 35 cases had at least one PCL–R score from a petitioner expert and a respondent expert. Thirty offenders had one PCL–R score from a respondent expert, 7 had two PCL–R scores from respondent experts, and 35 had no PCL–R score from a respondent expert. No case had two PCL–R scores from both petitioner and respondent experts. There were two cases with respondent PCL–R scores but no petitioner scores, and 33 cases with petitioner PCL–R scores but no respondent PCL–R scores.

Agreement between evaluators on the same side of the case. Table 1 provides rater agreement statistics for cases in which two evaluators retained by the same side of the case reported separate risk scores. For the 13 cases with PCL–R scores from two petitioner experts, ICC(1,A) was poor (.24). The difference in scores for these cases ranged from 0 to 17 points, with an average difference of 6.37 points. Five of the 13 scores differed by $\geq$10 points. Petitioner experts reported identical scores in two cases. For the 7 cases with PCL–R scores from respondent experts, ICC(A,1) was strong (.88), with difference scores ranging from 0 to 4 points ($M = 2.57$). The largest difference between respondent experts was 4.00 points, with identical scores in none of the cases.

Of the 35 cases with PCL–R scores from both petitioner and respondent experts, 9 had two petitioner PCL–R scores and 7 had two respondent PCL–R scores. Thus, 16 of the 35 offenders had two PCL–R scores from one of the two sides in the case.

Agreement between opposing evaluators: Mean scores. Table 2 reports the results of paired-samples *t* tests examining whether there was a statistically significant difference in PCL–R scores from petitioner and respondent experts in the 35 cases with opposing scores. PCL–R scores from petitioner experts were significantly higher than those from respondent experts, regardless of how we handled cases with multiple scores from experts on the same side of the case. Cohen's *d* values for these differences were medium in size, ranging from .68 (minimum discrepancy dataset) to .83 (maximum discrepancy dataset).

Agreement between opposing evaluators: Difference scores. The PCL–R manual (Hare, 2003) reports that the *SEM* for the PCL–R total score is approx-

imately 3 points. Assuming a normal distribution, approximately 98% of difference scores between two evaluators should fall within 2 *SEMs*, or 6 points. We calculated a difference score for each of the 35 offenders with opposing PCL–R scores by subtracting the respondent expert's score from the petitioner expert's score. Difference scores with a positive value indicated that the petitioner expert assigned a higher PCL–R score, whereas a negative score indicated that the respondent's expert assigned a higher score.

Using the minimum discrepancy scores, we found that the mean difference score was 4.97 ($SD = 7.26$), with values ranging from −8.00 to 25.20. Of the 35 difference scores, 13 (37.1%) were 6.00 or greater, in the direction of adversarial allegiance. Only two discrepancy scores were greater than 2 *SEM* units in the opposite direction (scores of −7.00 and −8.00), indicating that the respondent expert reported a higher PCL–R score than the petitioner expert. Using the dataset of maximum discrepancy scores, we found that the mean difference score was 6.60 ($SD = 7.68$), with values ranging from −8.00 to 25.20, which is more than twice as large as the *SEM* reported in the PCL–R manual (Hare, 2003). Of the 35 difference scores, 17 (48.6%) were 6.00 or greater. Once again, only two discrepancy scores were greater than 2 *SEM* units in the opposite direction. Using the dataset of average discrepancy scores, we found that the mean difference score was 5.79 ($SD = 7.08$), with values ranging from −8.00 to 25.20, with 14 (40.0%) difference scores greater than 6.00 and only one (2.9%) greater than 2 *SEM* units in the opposite direction.

Agreement between opposing evaluators: ICCs. ICC(A,1) was in the .40 range for the PCL–R, regardless of how we dealt with cases in which there were multiple scores from evaluators on the same side of the case (see Table 3). These values indicate that less than half of the variance in the set of PCL–R scores could be attributed to the offenders' true level of psychopathy as measured by the PCL–R. Of the remaining variance, 18% to 25% was attributable to the side for which the evaluation was performed. Finally, 35% to 40% of the variance in scores was attributable to other sources of error.

Do selection effects account for low levels of agreement in cases with opposing scores? As in the analyses of the STATIC-99 and MnSOST–R, we considered the possibility that the 35 cases for which we had opposing scores might systematically differ from the other cases in which opposing scores were not available. Perhaps respondent-retained evaluators only administered the PCL–R when they suspected that the score reported by the petitioner-retained evaluator was inappropriately high. The t tests revealed similar PCL–R scores for the 35 offenders with opposing PCL–R scores and 33 with only a petitioner PCL–R score. This finding applied to all three sets of discrepancy scores: For minimum discrepancy, $t(66) = 0.11$, $p = .91$, $d = 0.03$; for maximum discrepancy, $t(66) = 0.27$, $p = .79$, $d = 0.07$; and for average discrepancy, $t(66) = 0.06$, $p = .95$, $d = 0.02$. Again, there was no evidence to support the hypothesis that rater disagreement between opposing evaluators was due to a selection effect in which respondent evaluators administered the PCL-R because the score from the original petitioner evaluator appeared unusually high.

Discussion

The goal of this study was to investigate interrater agreement on ARAIs as scored by forensic evaluators who were retained by opposing sides in adversarial legal proceedings. We studied SVP trials because these proceedings routinely involve opposing forensic evaluators who have administered and scored the same ARAIs after evaluating the same offender, using essentially the same record base. Comparing ARAI rater agreement values from opposing evaluators in these adversarial SVP proceedings with ARAI rater agreement values from the research literature allows us to form some impressions about the pull of adversarial proceedings. Indeed, rater agreement values such as ICCs can be used as a metric to estimate the effects of adversarial allegiance, particularly if generalizability theory is used to estimate the amount of variance in test scores attributable to the side retaining the forensic evaluator. Finally, apart from the issue of adversarial allegiance, examining rater agreement values in SVP trials provides new data on the field reliability (Wood, Nezworski, & Stejskal, 1996) of popular risk measures; that is, their reliability as applied in routine practice outside of the research context.

Agreement Between Evaluators Retained by the Same Side

As detailed in Table 1, there were several instances in which rater agreement was lower than expected, even among raters retained by the same side. For example, the ICC(A,1) value for the PCL–R as administered by respondent-retained evaluators was .88—in the range we would expect based on research studies—but the ICC(A,1) value for petitioner-retained evaluators was poor (.24). Conversely, for the MnSOST–R, the ICC(A,1) value for petitioner-retained evaluators was higher (.63) than the unusually poor agreement among respondent-retained evaluators (.00). Rater agreement values for the STATIC-99—ICCs(A,1) of .84 and .95— were noticeably stronger and comparable with those reported in the research literature.

These rater agreement values for evaluators on the same side should be interpreted quite cautiously, given the small number of cases contributing to these analyses (*n*s = 6–17), in which a few score differences can have a tremendous impact on overall agreement values. There is also a possibility that, in these cases, the legal team requested a second evaluation specifically because results from their first evaluator appeared questionable (a selection effect that would inflate rater disagreement).[6]

Nevertheless, the strong agreement from same-side evaluators for the STATIC-99 is noteworthy compared with the poorer agreement for the other measures. The differences may be attributable to the nature of the items on each measure. The 10 STATIC-99 items require knowledge of offender demographics, offense history, and some minimal data about offense victims (i.e., gender, whether victim was related to offender, whether victim was a stranger to offender). The MnSOST–R items also rely on knowledge of an offender's basic criminal history. However, some require additional detailed knowledge of the sexual offense (e.g., "Was force or threat of force used?"), which may be less clear in the records and require more subjective inference by the evaluator. The MnSOST–R also requires some knowledge of the

[6] Unfortunately, we were not able to examine this possibility.

offender's incarceration experience (e.g., disciplinary infractions, drug treatment, sex offender treatment). One research group (Barbaree et al., 2001) observed:

> Scoring the MnSOST-R requires careful reading of extensive manual material, a relatively large amount of training of the coders, and a high degree of diligence among the coders. Our coders were trained over a period of 1 full working day. The developers of the MnSOST-R have provided more comprehensive scoring guidelines and examples than provided by the developers of the RRASOR or the Static-99. Nevertheless, we found the MnSOST-R to be the most difficult of the actuarial measures to code. The MnSOST-R items are very specific and the clinical file material available was not always exactly pertinent to the items as described. In contrast, the . . . Static-99 [was] straightforward to code and score. (p. 513)

The PCL–R, to a greater extent than that of the two ARAIs, relies on evaluator clinical skills to elicit information and draw inferences about an offender's personality and interpersonal style. Research reveals that it is certainly possible for trained raters in the field to arrive at adequate interrater reliability when they participate in the same training and have access to the same interview and records (Gacano & Hutton, 1994). The only other study of agreement for same-side SVP evaluators also found a higher level of agreement for both the PCL–R and MnSOST–R (Levenson, 2004b). Our same-side rater agreement values are based on so few cases—and perhaps atypical cases—that we cannot conclude that the field reliability of the MnSOST–R and PCL–R is uniformly poor across all contexts. However, we do conclude that their field reliability cannot be assumed and that more research into their field reliability is essential.

Agreement Between Evaluators Retained by Opposing Sides

As with the rater agreement values for evaluators on the same side of the case, our rater agreement values from evaluators on opposing sides revealed weaker rater agreement in the field, as compared with research studies. Depending on which approach we took to analyzing opposing scores (i.e., minimum, maximum, or average discrepancies), the ICC(A,1) values fell near .42 for the PCL–R, .44 for the MnSOST–R, and .62 for the STATIC-99. These values are much lower than the same-side agreement values for SVP evaluation scores in Florida (Levenson, 2004b).[7]

Of course, if agreement between evaluators on the same side of a case is poor, we should certainly be cautious about concluding that poor agreement between opposing evaluators suggests adversarial allegiance. Therefore, an important

[7] One possible explanation for the difference between the ICCs for opposing (Murrie et al., 2008) and nonopposing (Levenson, 2004b) scores is that the researchers used different ICC equations. Researchers examining the opposing evaluator agreement reported an ICC for a single evaluator (Murrie et al., 2008), whereas the nonopposing study reported an ICC for the score averaged across two evaluators (J. S. Levenson, personal communication, July 14, 2008). The multiple-evaluator coefficients are relevant when the score that is used in court is a single score averaged across all evaluators. However, because scores reported in court are not averaged across raters, the single-evaluator ICC is the most directly relevant for scores reported in court. The ICC for a single evaluator would be .72 for PCL–R in the sample of nonopposing evaluators (applying the Spearman–Brown prophecy formula to data from Levenson 2004b), which is still widely discrepant from the .39 from opposing evaluators. The single-evaluator ICCs for the MnSOST–R and STATIC-99 among the nonopposing SVP evaluators (Levenson, 2004b) are both .73.

research question is whether the disagreement between opposing raters was unsystematic, which would simply suggest weaker agreement in the field than in research studies, or systematic, which might suggest adversarial allegiance *if* petitioner-retained evaluators tended to assign higher risk scores and respondent-retained evaluators tended to assign lower risk scores.

Score differences appeared systematic, regardless of whether we used pairs of scores that produced the smallest or largest differences between opposing evaluators. In other words, although evaluators on the same side disagreed at times, these differences did not explain the poor agreement between opposing evaluators. As detailed in Table 2, the mean petitioner scores were higher than mean respondent scores for every measure, using minimum and maximum discrepancy scores.

There were some important differences among measures. The difference between petitioner and respondent scores were in the medium-to-large and large ranges for the PCL–R (ds = 0.68–0.83) and the MnSOST–R (ds = 0.70–0.95). Substantial proportions of score differences were greater than 2 *SEM* units, in the direction consistent with adversarial allegiance. For the STATIC-99, scores also differed in the direction consistent with adversarial allegiance, but these differences were considerably smaller in size (ds = 0.29–0.37), and few were greater than 2 *SEM* units.

Estimating the Sources of Score Variance

One contribution of this study is the application of generalizability theory to better estimate the sources of error variance in ARAI scores. We estimated the proportion of variance attributable to three sources: (a) the offenders being evaluated, (b) the "side" (petitioner or respondent) that retained the evaluator, and (c) other unidentified sources of error. For two measures, a sizeable portion of this variance was attributable to the side that retained the evaluator: 18–25% of the variance in PCL–R scores and 19–30% of the variance in MnSOST–R scores, depending on how we handled data from cases with more than two scores per side. For the STATIC-99, only 3–5% of variance was attributable to the side retaining the evaluator. Thus, although none of the measures produced ICC values in the desired .80 range, the reasons for poor ICCs might differ for each measure. For the STATIC-99, adversarial allegiance accounted for only a small amount of the error variance, suggesting that random error or other sources of systematic error (e.g., scoring tendencies of individual evaluators, idiosyncratic interpretation of record information) are more likely to be responsible for the modest ICC. For the PCL–R and MnSOST–R, adversarial allegiance accounted for a substantial amount of the error variance, although the amount of unexplained error variance for the PCL–R was similar to that of the STATIC-99, suggesting that factors other than adversarial allegiance also contribute to the low ICC for the PCL–R.

Ideally, the variance in test scores attributable to adversarial allegiance would be near zero. Values farther away from zero might be tolerable if the rater disagreement appeared to be random; for example, if higher scores were produced by petitioner evaluators in about half of the cases and respondent evaluators in the other half. However, our analyses suggested that this was not the case in our sample. If the differences are not random, we still want the proportion of variance

attributable to adversarial allegiance to be as small as possible. From a purely psychometric perspective, if rater agreement coefficients, ICC(A,1), should be at least .80 for forensic instruments (Heilbrun, 1992), then no more than 20% of the variance can be attributable to opposing evaluators. However, when the amount of variance attributable to opposing evaluators was in the 20% range in the present study (see Table 2), the mean score differences between opposing evaluators were still moderate to large in size (Cohen's d range = 0.68–0.78) and probably larger than the field will (or should) tolerate. Our findings for the STATIC-99 appeared much closer to ideal, with about 5% of the variance attributable to adversarial allegiance. This amount of variance translated into statistically significant but smaller differences between opposing evaluators (Cohen's d = 0.29–0.37).

Analyses also revealed that a sizeable proportion of the variance in scores was attributable to other sources of error, beyond the variance attributable to the offenders and the side retaining the evaluator. Of course, *error*, in the psychometric sense, simply refers to other sources of variance that we could not examine. These sources may be random (unexplainable) or systematic (potentially explainable depending on study design). Examples of systematic sources of variance may include the following: variance attributable to specific evaluators (perhaps some evaluators are inclined to assign higher scores in ambiguous cases, while others lean towards lower scores); variance due to whether an interview was conducted (in a few cases, offenders declined to participate in interview, and evaluators scored instruments entirely by record, although our data do not allow us to examine this as a variable); or variance due to difference in the quality of records available for review (although likely influential in many forensic evaluations, this variance is likely minimal in this study, in which evaluators on each side received the same database of records). However, these sources of variance should further divide the unexplained error variance (i.e., see the last column in Table 3), not variance due to offenders being evaluated.

Implications for Sex-Offender Risk Assessment

Regarding sex-offender-specific ARAIs, results from this study tend to support the use of STATIC-99 over the MnSOST–R. Evaluators using the STATIC-99—whether on the same side of a case (see Table 1) or on opposing sides (Tables 2 and 3)—tended to demonstrate much stronger interrater agreement than evaluators using the MnSOST–R. Our stronger reliability data for the STATIC-99 appear consistent with the many studies that support use of the STATIC-99 (see Hanson & Morton-Bourgon, 2007, for a review). Likewise, our findings regarding poor reliability for the MnSOST–R appear consistent with studies that report weaker predictive validity results for the MnSOST–R (Bartosh, Garby, Lewis, & Gray, 2003) or emphasize problems with the measure (Vrieze & Grove, 2008; Wollert, 2002, 2003). Because there appear to be few advantages to combining actuarial measures (e.g., Seto, 2005), our results suggest that administering the less reliable MnSOST–R, in addition to the more reliable STATIC-99, may be unnecessary (at best) or misleading (at worst).

Regarding the PCL–R, these results extend the findings from an initial study of PCL–R agreement in adversarial SVP proceedings (Murrie et al., 2008). Our small-sample results are probably not sufficient to make definitive recommenda-

tions for widescale practice, but results do suggest that caution is in order when considering PCL–R results in SVP proceedings, if a meaningful portion of the variance in PCL–R scores is attributable to the side retaining the evaluator. The PCL–R has been popular in sex offender risk assessments because research data tend to reveal a relation between PCL–R scores and sexual reoffense (Quinsey, Rice, & Harris, 1995; Rice, Harris, & Quinsey, 1990). For example, a comprehensive meta-analysis identified psychopathy as one of the strongest predictors of sex offender recidivism (Hanson & Morton-Bourgon, 2005), although the predictive power of the PCL–R appears more attributable to the impulsive/unstable lifestyle than to the interpersonal/emotional style associated with psychopathy (Knight & Guay, 2006).

However, even this well-known research base on PCL–R scores and sexual recidivism must be considered more cautiously in light of our findings. Research studies describe the predictive validity of PCL–R scores as scored in research contexts, usually by trained raters. They do not describe the predictive validity of the PCL–R as scored by evaluators in adversarial legal proceedings. If PCL–R reliability is poorer in adversarial proceedings, as our findings suggest (see also Boccaccini, Turner, & Murrie, 2008; and Murrie et al., 2008), then evaluators in adversarial proceedings should be cautious in assuming that the (potentially less reliable) PCL–R scores they assign carry the same predictive validity with respect to sexual recidivism. Future research must examine how well the PCL–R *as administered in the field* predicts sex offender reoffense. At least one recent field validity study found that the PCL–R, as administered by SVP screening evaluators in routine practice, bore no relationship to sexual reoffenses (Boccaccini, Murrie, & Caperton, 2008).

Broadly, results suggest that evaluators in SVP proceedings (and the attorneys who scrutinize their testimony) should be more attuned to the possible influence of adversarial allegiance. This holds true even for tasks that appear to involve little subjectivity, such as scoring ARAIs. SVP proceedings are explicitly adversarial and rely primarily on expert testimony as a source of evidence. SVP evaluations feature a number of complex, debatable considerations about which well-qualified and thoughtful professionals may sometimes disagree: for example, assigning paraphilia diagnoses to offenders who may not be forthcoming about sexual interests or experiences, drawing inferences from ambiguous criminal records, and using actuarial instruments based on group data from offenders in one context to draw inferences about a single offender in another context. Perhaps, then, it is not surprising if the adversarial system pulls opposing SVP evaluators toward different perspectives on debatable issues (Murrie, Boccaccini, & Turner, in press).

Implications for Addressing Adversarial Allegiance in Practice

One recent survey asked forensic clinicians to share their opinions about bias among experts in their field. Most participants reported that experts view themselves "bias free" and able to "compensate for any biases they might have" (Commons, Miller, & Gutheil, 2004, p. 73). Our findings, although from a small sample of unique trials in one state, bode for a more humble perspective. We could identify no reason to believe that evaluators in our sample were any more

vulnerable to adversarial allegiance than other clinicians in adversarial proceedings; many had decades of experience, and several had advanced qualifications (e.g., diplomate status, supervisory and mentoring roles). Likewise, we could identify no reason to believe that SVP evaluations pull for adversarial allegiance more than other adversarial proceedings. Thus, our results underscore the cautions about objectivity that authorities offer (Brodsky, 1991; Rogers, 1987; Shuman & Greenberg, 2003) and underscore the need for clinicians to use practical self-monitoring approaches to reduce the influence of adversarial allegiance (Borum, Otto, & Golding, 1993; Murrie & Warren, 2005).

It is important to emphasize that score differences in the direction of adversarial allegiance may not have been intentional. Our study cannot address, for example, whether respondent-retained evaluators who reviewed original evaluation results purposefully worked to arrive at lower risk scores. Likewise, our study cannot address whether petitioner-retained evaluators tended to lean toward giving a higher score when an offender was on the border between a higher and a lower score. Social science research consistently suggests that even unintentional behaviors tend to be self-serving. For example, we can expect about two thirds of scoring and data recording errors to favor the position of the person responsible for the errors (Rosenthal, 1978). Likewise, the adversary legal system pulls forensic evaluators in ways they may not immediately recognize (Applebaum, 1998).

However, it is also important to emphasize that not all disagreements among clinicians in the adversarial system reflect adversarial allegiance. To take one concrete example, a respondent-retained evaluator may (and indeed, should) note an error in a previous petitioner-scored STATIC-99, which results in a lower total score when corrected. This respondent-retained evaluator is thorough but not necessarily biased. Of course, it would be important for the same respondent-retained evaluator to correct *all* scoring errors, even if doing so yielded a higher risk score. Evaluators who are cautious about adversarial allegiance need not be reluctant to vigorously advocate for their opinions and need not "preclude forceful representation of the data and reasoning upon which a conclusion . . . is based" (Committee on Ethical Guidelines for Forensic Psychologists, 1991, p. 664). Evaluators, however, do need to examine closely for the degree to which their opinions—even routine instrument scoring decisions—may have been subtly influenced by adversarial allegiance. Opposing attorneys will do the same.

Implications for Addressing Adversarial Allegiance in Research

Our findings suggested that some instruments (the STATIC-99 in this study) might be less susceptible to rater disagreement in the context of adversarial proceedings than other instruments (the PCL–R and the MnSOST–R in this study). Research should continue to investigate agreement on these instruments and others in the context of adversarial proceedings and to identify factors that make forensic assessment instruments more or less susceptible to disagreement between opposing evaluators.

Although this study examined interrater agreement on specialized actuarial risk measures in one unique legal context (i.e., SVP trials), questions about the reliability of assessment instruments in adversarial proceedings are relevant to

many forms of forensic assessment. First, poor agreement likely limits the observed validity of an instrument. A finding that instrument scores vary by adversarial side raises questions about which scores are most accurate. For example, if we were to find that scores on risk measures consistently differed as scored by prosecution versus as scored by defense, it would become important to examine the predictive validity of the instruments as scored by each side. Of course, such a study would almost never be possible, because there is rarely a sufficient sample of evaluations from each side. Our findings of disagreement between opposing evaluators raises questions about the extent to which recidivism findings from well-controlled research studies, with high rater agreement, generalize to risk scores reported in court. Which side's expert, if either, is more like the type of systematically trained research assistant who scores risk measures for a research study? Again, we emphasize the need for studies that address the field validity of measures used in adversarial proceedings.

Another important research implication to emphasize is that adversarial allegiance does not explain all of the error variance in ARAI and PCL–R scores in our sample. As can be seen in Table 3, 30% to 40% of the variance in scores could not be accounted for by either true standing on the measure or adversarial allegiance. Some recent research suggests that the idiosyncratic evaluation and scoring tendencies of individual evaluators may explain a portion of this variance. A few simple studies document differences in terms of how often evaluators reach opinions supportive of trial competence (Murrie, Boccaccini, Zapf, Warren, & Henderson, 2008) or legal sanity (Murrie & Warren, 2005). Although these studies could not address all possible explanations for the differences in evaluator opinions, these studies do suggest that forensic evaluators are probably not interchangeable. Rather, differences in the evaluators themselves likely contribute to differences in their patterns of opinions, and these evaluator differences warrant further study.

More relevant to this study of forensic assessment instruments in SVP proceedings, recent research suggests that evaluator differences can account for some of the variability in PCL–R scores, even outside the pull of adversarial legal proceedings (Boccaccini, Turner, & Murrie, 2008). This study examined PCL–R scores for more than 300 offenders who were evaluated by one of 20 different experts during the first step of Texas's SVP evaluation process. Differences between evaluators accounted for approximately 30% of the variance in PCL–R scores, with the average PCL–R scores given by two of the most prolific evaluators differing by nearly 10 points.

A carefully designed experiment might allow researchers to calculate the relative effects of offender characteristics, evaluator characteristics, and adversarial allegiance on evaluator opinions or measure scores. The ideal study would have (a) a large pool of evaluators, (b) randomly assigned to perform some evaluations for the prosecution and an equal number for the defense, and (c) opposing scores for every case. For many reasons, this ideal study is virtually impossible in most real-world contexts. For example, the adversarial system rarely encourages random assignment of evaluators to cases, and policies protecting attorney work–product allow discretion about disclosing evaluator scores and opinions. Given the challenges to real-world experimental studies, investi-

gating adversarial allegiance will likely continue to require naturalistic, and imperfect, designs such as the present study.

Despite the practical challenges to researching adversarial allegiance in forensic evaluation, it is important for researchers to attempt these studies, particularly with cooperating jurisdictions or systems. Some jurisdictions may be understandably wary of facilitating this type of study. However, any short-term drawbacks are likely outweighed by the potential for substantial practical benefits in the long term. Studies that identify the nature and extent of adversarial allegiance lay the groundwork for improvements that minimize the influence of this adversarial allegiance. In SVP proceedings, for example, minimizing adversarial allegiance increases the likelihood that limited commitment resources are devoted to those offenders who most closely match commitment criteria and pose the greatest risk to the community.

Implications for Addressing Adversarial Allegiance in Law and Policy

In some respects, we should consider cautiously any policy implications from this study. After all, data are based on a sample of 72 cases from one state, which is only one of the nearly 20 states with SVP laws. Any policy implications must be tentative, recognizing that only additional research can reveal whether findings from this study are typical of other jurisdictions as well.

On the other hand, this study (along with the Murrie et al., 2008 and Boccaccini, Turner, & Murrie, 2008 studies of the same jurisdiction) offers the only available case data on an understudied but important question. The jurisdiction from which these data are drawn appears similar to other jurisdictions in terms of how SVP screening, evaluations, and trials take place.[8] There is no reason to expect stronger or weaker patterns of adversarial allegiance in this jurisdiction as compared with others. Therefore—although it is important not to overvalue these findings—it also seems important not to underestimate the degree to which these findings may carry implications for other jurisdictions, at least until other research specific to those jurisdictions becomes available.

Perhaps the most straightforward and narrow question these results raise is: How might systems reduce the impact of adversarial allegiance on test scores? It may be reasonable to require SVP evaluators to participate in uniform instrument training or formally demonstrate additional instrument expertise and correct scoring. This approach seems most appropriate if we assume that poor rater agreement results primarily from generally poor scoring practices that, perhaps inadvertently, drift in the direction of adversarial allegiance. If we assume that scoring in the direction of adversarial allegiance is more intentional than inadvertent, mandated training is probably of less help. In either case, systems may benefit from outside consultation to better understand the nature and extent of scoring problems related to adversarial allegiance in their jurisdiction.

---

[8] Texas is unique among the states with SVP laws in that offenders civilly committed as SVPs in Texas face outpatient commitment only, albeit with extremely rigorous restrictions. Inpatient commitment is not an option. However, there is no reason to believe that a unique commitment arrangement would lead to unique findings regarding adversarial allegiance when the precommitment process is similar to the process in most other states.

Even if systems make some progress in reducing the influence of adversarial allegiance on instrument scores, it remains important to ask: To what extent should correctional policies and court decisions be influenced by scores on instruments, if instrument scores are influenced by adversarial allegiance? Findings demonstrating that certain instruments are vulnerable to the effects of adversarial allegiance, whereas others are more robust, should probably deter correctional systems from using the more vulnerable instruments. In court, there may be good reason to challenge the admissibility of particular instruments if research evidence continues to suggest that their scores are influenced by adversarial allegiance. At a minimum, courts should view with a healthy skepticism any results from the PCL–R or MnSOST–R as scored by retained evaluators in adversarial proceedings.

Of course, instruments are only vulnerable to the effects of adversarial allegiance if the evaluators who score them are vulnerable.[9] Therefore, the broader policy questions involve how we might minimize the influence of adversarial allegiance on evaluators. Certainly these questions are not new. For decades, observers have complained—although usually through anecdotes and impressions rather than empirical data—of bias or partisanship by expert witnesses. There is no shortage of proposed solutions, but each proposed solution to the problem of adversarial allegiance brings problems of its own (see Mnoonkin, 2008, for an excellent review).

For example, one popular suggestion to reduce adversarial allegiance among expert witnesses is to rely only on "neutral experts" appointed by the court rather than either litigating party. A well-qualified, court-appointed neutral expert may in fact solve the problem of adversarial allegiance influencing instrument scores by providing the court with more trustworthy data (i.e., something closer to the "real" scores). However, as Mnoonkin (2008) emphasized, neutral experts may be problematic in those cases in which scientific data is insufficiently developed, under dispute, or open to legitimate differences of opinion. In these instances, the neutral expert can rarely convey to the court the nature and extent of the scientific dispute or clinical ambiguity. The jury may simply follow the expert's summary opinion (particularly because the expert is perceived as neutral and trustworthy), bypassing much of the deliberation and evidence weighing that occurs when juries consider contrasting opinions from opposing experts. In short, neutral experts may make the juries' job easier, but their presence makes less likely the careful consideration of opposing perspectives that our adversarial system of justice prioritizes. Particularly in SVP proceedings—in which the necessary science is generally underdeveloped and there remains much room for reasonable clinicians to disagree (Murrie et al., in press)—we should probably be leery of interventions that reduce opportunities for courts to consider contrasting data or contrasting interpretations of ambiguous data.

[9] For this reason, we do not argue that abandoning all instruments is a solution to the problem of adversarial allegiance influencing test scores. Even if some instruments are vulnerable to adversarial allegiance, there is no reason to believe that unaided, unstructured clinical judgment is *less* vulnerable to adversarial allegiance.

In contrast to the court-appointed "neutral expert" role, policies in the jurisdiction that we studied may enhance the visibility of adversarial allegiance. For example, defense counsel is free to consult with potential evaluators, asking them to review case materials and even provide preliminary opinions, before retaining them to formally evaluate the respondent. Consistent with *United States v. Alvarez* (1975), the court is therefore not exposed to opinions offered by any "discarded experts" whom attorneys screen for possible participation in a case but ultimately do not retain for a full evaluation and testimony.It is important to note that these discarded experts might have offered opinions that are quite concordant with one another or with the petitioner's evaluator. In other jurisdictions (consistent with *United States ex rel. Edney*) prosecutors are allowed access to experts whom the defense has decided not to use. Our study could not identify which respondent evaluators were the only ones consulted on the case and whose selection was the result of extensive screening; nor would the jurors, judge, or petitioner be privy to this information. Thus, we again emphasize that our results may not generalize to all forensic opinions in adversarial proceedings; rather, results reflect only opinions visible at the point of deposition or trial testimony.

On reviewing the research findings that suggest adversarial allegiance, it is tempting to favor an *Edney*-like arrangement in which the petitioner could query all of the respondent's consulting experts rather than only the one who provided a favorable opinion. Under such an arrangement, we might observe better agreement among more evaluators and less apparent adversarial allegiance. However, such a change would make it even more difficult for the respondent to pursue a defense against civil commitment.[10] The *Alvarez* court suggested that an attorney "must be free to make an informed judgment with respect to the best course for the defense without the inhibition of creating a potential government witness" (*United States v. Alvarez,* 1975).[11] In SVP proceedings, counsel for the respondent may be reluctant to seek an evaluation, or even a consultation, if results are not protected from discovery. However, seeking an evaluation is likely a crucial defense strategy in a legal proceeding that revolves primarily around expert opinion. In summary, many of the approaches that appear likely to reduce adversarial allegiance are those that may also reduce defense counsel's options to explore reasonable defense strategies.

Overall, the problem of score differences in the direction of adversarial allegiance is probably intertwined with several policies that reflect deeply held values about the benefits of due process and an adversary system of justice. A core value of the adversary system is the opportunity for opposing sides to put forward their best case. Therefore, there is nothing inherently problematic about attorneys searching for clinicians who are most likely to interpret the available data in a manner favorable to their case. Indeed, in a field such as SVP risk assessment,

[10] In Texas, almost all cases have been decided in favor of civilly committing the respondent.

[11] To be clear, both *Alvarez* and *Edney* applied to criminal justice proceedings, but the implications are similar in quasi-criminal proceedings such as SVP trials. The American Bar Association's (1989) Criminal Justice Mental Health Standards also favor an *Alvarez* ruling, except in cases in which it is clear that defense counsel has consulted many experts specifically to make them unavailable to the prosecution (see Melton et al., 2007, for an excellent overview of issues related to *Edney* and *Alvarez*).

which features many issues on which reasonable professionals may disagree, we can expect attorneys to seek and favor evaluators whose positions on these debatable issues are most conducive to the attorney's case strategy. What is striking about our findings, however, is that opposing evaluators differed not simply on scientifically debatable issues but also on the scores they derived from ostensibly objective assessment instruments, which have clear scoring guidelines and strong rater agreement values in nonadversarial research contexts. For this reason, we suspect that the problem of score differences in the direction of adversarial allegiance primarily reflects a problem with evaluators much more than it reflects a problem with legal policy or practice.

Therefore, our results also have implications for the policies that guide forensic mental health professionals. The Specialty Guidelines for Forensic Psychologists (Committee on Ethical Guidelines for Forensic Psychologists, 1991), which are currently under extensive revision, give some attention to adversarial allegiance and the need for objectivity. However, these guidelines are considered "aspirational" rather than regulatory in the strictest sense. Indeed, the field of forensic mental health assessment is not currently regulated by any clearly recognized professional standard of care, but there is recent movement toward articulating clearer standards of practice (Heilbrun, DeMatteo, Marczyk, & Goldstein, 2008). As this movement continues, it will be important to consider the issue of adversarial allegiance.

However, regulating adversarial allegiance, even within the mental health fields, is a delicate matter. Because many scientific or clinical issues that arise at trial are those in which well-qualified professionals may genuinely and reasonably disagree (Mnoonkin, 2008), overzealous efforts to addressed adversarial allegiance could do more harm than good, both to clinicians and the courts that rely on their input. Nevertheless, it may be helpful to have some mechanism to address clinician behaviors that, in a manner linked to adversarial allegiance, clearly and substantially deviate from firmly established guidelines (e.g., scoring procedures delineated in instrument manuals or formal diagnostic criteria delineated in the *Diagnostic and Statistical Manual of Mental Disorders* [4th ed., text revision; American Psychiatric Association, 2000]). Melton and colleagues (2007) suggested negligent misdiagnosis as one possible ground for malpractice litigation addressing forensic assessment. Heilbrun and colleagues (2008) elaborated that clinicians might commit negligent misdiagnosis "by obtaining so little information that mistakes are far more likely and plausible alternatives are not tested" (p.14). Similarly, clinicians who seek information only to support an adversarial position or dismiss information contradictory to their position may be similarly negligent in the service of adversarial allegiance.

Conclusion

Results from our small sample strongly suggest that scores on some popular measures widely used in legal proceedings may be influenced by adversarial allegiance. There appeared to be less interrater disagreement on the STATIC-99 than on the MnSOST–R and the PCL–R, which suggests that some measures may be more vulnerable than others to the pull of the adversarial system. Therefore, we recommend further research—both by scholars and systems—to investigate the

potential influence of adversarial allegiance in forensic evaluation and expert testimony. We also recommend that forensic evaluators, and the courts that consider their input, examine carefully the ways in which adversarial dynamics may have influenced their evaluation procedures, instrument scoring, and opinion formation.

References

Amenta, A. (2005). *The assessment of sexual offenders for civil commitment proceedings: An analysis of report content.* Unpublished doctoral dissertation, Sam Houston State University.

American Bar Association. (1989). *ABA criminal justice mental health standards*. Washington, DC: Author.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

Applebaum, P. S. (1998, March). *The medicalization of "judicial decision making."* Lecture presented at the Judicial Training Conference on Mental Health Testimony in the Courtroom sponsored by Mental Health Legal Advisors Committee and Flaschner Judicial Institute. Retrieved August 15, 2008, from http://www.mass.gov/mhlac/applebaum.htm

Association for the Treatment of Sex Abusers. (2001). *Civil commitment of sexually violent predators.* Position paper retrieved from http://www.atsa.com/ppcivilcommit.html

Barbaree, H. E., Seto, M. C., Langton, C. M., & Peacock, E. J. (2001). Evaluating the predictive accuracy of six risk assessment instruments for adult sex offenders. *Criminal Justice and Behavior, 28,* 490–521.

Bartosh, D. L., Garby, T., Lewis, D., & Gray, S. (2003). Differences in the predictive validity of actuarial risk assessments in relation to sex offender type. *International Journal of Offender Therapy and Comparative Criminology, 47,* 422–438.

Blair, P., Marcus, D., & Boccaccini, M. T. (2008). Is there an allegiance effect for assessment instruments? Actuarial risk assessment as an exemplar. *Clinical Psychology: Science and Practice, 15,* 346–360.

Boccaccini, M. T., Murrie, D. C., & Caperton, J. (2008, March 5–8). *Predicting recidivism with the MnSOST-R, PAI, PCL-R, and STATIC-99 in a statewide sex offender sample.* Paper presented at the 2008 conference of the American Psychology-Law Association, Jacksonville, FL.

Boccaccini, M. T., Turner, D. T., & Murrie, D. C. (2008). Do some evaluators report consistently higher or lower scores on the PCL-R?: Findings from a statewide sample of sexually violent predator evaluations. *Psychology, Public Policy, and Law, 14,* 262–283.

Borum, R., Otto, R., & Golding, S. (1993). Improving clinical judgment and decision making in forensic evaluation. *Journal of Psychiatry & Law, 21,* 35–76.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer.

Brodsky, S. L. (1991). *Testifying in court: Guidelines and maxims for the expert witness.* Washington, DC: American Psychological Association.

Campbell, T. (2006). The validity of the Psychopathy Checklist—Revised in adversarial proceedings. *Journal of Forensic Psychology Practice, 6,* 43–53.

Campbell, T. W. (2007). *Assessing sex offenders: Problems and pitfalls* (2nd ed.). Springfield, IL; Charles C Thomas.

Committee on Ethical Guidelines for Forensic Psychologists. (1991). Specialty guidelines for forensic psychologists. *Law and Human Behavior, 15,* 655–665.

Commons, M. L., Miller, P. M., & Gutheil, T. G. (2004). Expert witness perceptions of

bias in experts. *Journal of the American Academy of Psychiatry and the Law, 32,* 70–75.

DeMatteo, D., & Edens. J. F. (2006). The role and relevance of the Psychopathy Checklist—Revised in court: A case law survey of U.S. courts (1991–2004). *Psychology, Public Policy, and Law, 12,* 215–241.

Doren, D. M. (2002). *Evaluating sex offenders: A manual for civil commitment and beyond.* London: Sage.

Doren, D. M. (2004). Stability of the interpretative risk percentages for the RRASOR and Static-99. *Sexual Abuse: A Journal of Research and Treatment, 16,* 25–36.

Doren, D. M. (2006). Inaccurate arguments in sex offender civil commitment proceedings. In A. Schlank (Ed.), *The sexual predator: Law and public policy—Clinical practice* (Vol. III). Kingston, NJ: Civic Research Institute.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures. *Psychological Methods, 1,* 170–177.

Epperson, D. L., Kaul, J. D., Goldman, R., Hout, S., Hesselton, D., & Alexander, W. (1998). *Minnesota Sex Offender Screening Tool—Revised (MnSOST-R).* St. Paul: Minnesota Department of Corrections. Available online at http://www.psychology.iastate.edu

Epperson, D. L., Kaul, J. D., Hout, S., Goldman, R., Hesselton, D., & Alexander, W. (2003). *Minnesota Sex Offender Screening Tool—Revised (MnSOST-R) technical paper: Development, validation, and recommended risk level cut scores.* St. Paul: Minnesota Department of Corrections. Available online at http://www.psychology.iastate.edu

Gacano, C. B., & Hutton, H. E. (1994). Suggestions for the clinical and forensic use of the Hare Psychopathy Checklist—Revised. *International Journal of Law and Psychiatry, 17,* 303–317.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2,* 293–323.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19–30.

Hanson, R. K. (2001). *Note on the reliability of STATIC-99 as used by the California Department of Mental Health evaluators.* Unpublished report. Sacramento: California Department of Mental Health.

Hanson, R. K., Morton, K. E., & Harris, A. J. R. (2003). Sexual offender recidivism risk: What we know and what we need to know. In R. A. Prentky, E. S. Janus, & M. C. Seto (Eds.), *Annals of the New York Academy of Sciences: Vol. 989. Sexually coercive behavior: Understanding and management* (pp. 154–166). New York: New York Academy of Sciences.

Hanson, R. K., & Morton-Bourgon, K. E. (2005). The characteristics of persistent sexual offenders: A meta-analysis of recidivism studies. *Journal of Consulting and Clinical Psychology, 73,* 1154–1163.

Hanson, R. K., & Morton-Bourgon, K. E. (2007). *The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis.* Ottawa, Ontario, Canada: Department of the Solicitor General of Canada.

Hanson, R. K., & Thornton, D. (1999). *Static-99: Improving actuarial risk assessments for sex offenders* (User Report 99–02). Ottawa, Ontario, Canada: Department of the Solicitor General of Canada.

Hanson, R. K., & Thornton, D. (2003). *Notes on the development of the Static-2002* (User Report 2003–10). Ottawa, Ontario, Canada: Office of the Solicitor General of Canada.

Hare, R. D. (1991). *The Hare Psychopathy Checklist—Revised.* Toronto, Ontario, Canada: Multi-Health Systems.

Hare, R. D. (2003). *The Hare Psychopathy Checklist—Revised* (2nd ed.). Toronto, Ontario, Canada: Multi-Health Systems.

Harris, A., Phenix, A., Hanson, R. K., & Thornton, D. (2003). *STATIC-99 coding rules revised—2003.* Ottawa, Ontario, Canada: Office of the Solicitor General of Canada. Retrieved from www.static99.org

Harris, G. T., Rice, M. E., Quinsey, V. L., Lalumiere, M. L., Boer, D., & Lang, C. (2003). A multisite comparison of actuarial risk instruments for sex offenders. *Psychological Assessment, 15,* 413–425.

Heilbrun, K. (1992). The role of psychological testing in forensic assessment. *Law and Human Behavior, 16,* 257–272.

Heilbrun, K., DeMatteo, D., Marczyk, G., & Goldstein, A. M. (2008). Standards of practice and care in forensic mental health assessment: Legal, professional, and principles-based considerations. *Psychology, Public Policy, and Law, 14,* 1–26.

Interstate Commission for Adult Offender Supervision. (2007, April). *Sex offender assessment information survey 4–2007.* Retrieved December 29, 2008, from http://www.interstatecompact.org/Link Click.aspx?fileticket=jLoQPVeviaQ%3d&tabid=105&mid=431

Jackson, R. L., & Hess, D. T. (2007). Evaluation for civil commitment of sex offenders: A survey of experts. *Sexual Abuse: A Journal of Research and Treatment, 19,* 425–448.

Jackson, R. L., & Richards, H. J. (2008). Evaluations for the civil commitment of sexual offenders. In R. L. Jackson (Ed.), *Learning forensic assessment* (pp. 183–209). New York: Routledge.

Janus, E. S., & Prentky, R. (2003). Forensic use of actuarial risk assessment with sex offenders: Accuracy, admissibility, and accountability. *American Criminal Law Review, 40,* 1443–1499.

Kansas v. Hendricks, 117 S. Ct. 2072 (1997).

Knight, R. A., & Guay, J. (2006). The role of psychopathy in sexual coercion against women. In C. Patrick (Ed.), *Handbook of psychopathy* (pp. 512–532). New York: Guilford Press.

LaFond, J. Q. (2005). *Preventing sexual violence: How society should cope with sex offenders.* Washington, DC: American Psychological Association.

Langton, C. M., Barbaree, H. E., Seto, M. S., Peacock, E. J., Harkins, L., & Hansen, K. T. (2007). Actuarial assessment for reoffense among adult sex offenders: Evaluating the predictive accuracy of the STATIC-2002 and five other instruments. *Criminal Justice and Behavior, 34,* 37–59.

Leistico, A. R., Salekin, R. S., DeCoster, J., & Rogers, R. (2008). A large-scale meta-analysis relating Hare measures of psychopathy to antisocial conduct. *Law and Human Behavior, 32,* 28–45.

Levenson, J. S. (2004a). Sexual predator civil commitment: A comparison of selected and released offenders. *International Journal of Offender Therapy and Comparative Criminology, 48,* 638–648.

Levenson, J. S. (2004b). Reliability of sexually violent predator civil commitment criteria in Florida. *Law and Human Behavior, 28,* 357–368.

McGrath, R. J., Cumming, G. F., & Bouchard, B. L. (2003). *Current practices and trends in sexual abuser management: The Safer Society 2002 Nationwide Survey.* Brandon, VT: Safer Society Press.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30–46.

Melton, G. B., Petrila, J., Poythress, N. G., Slobogin, C., & Otto, R. K. (2007). *Psycho-*

logical evaluations for the courts: A handbook for mental health professionals and lawyers (3rd ed.). New York: Guilford Press.

Miller, H. A., Amenta, A. E., & Conroy, M. A. (2005). Sexually violent predator evaluations: Empirical evidence, strategies for professionals, and research directions. *Law and Human Behavior, 29,* 29–54.

Mnoonkin, J. (2008). Expert evidence, partisanship, and epistemic confidence. *Brooklyn Law Review, 73,* 587–611.

Monahan, J. (2006). A jurisprudence of risk assessment forecasting harm among prisoners, predators, and patients. *Virginia Law Review, 92,* 391–435.

Murrie, D. C., Boccaccini, M., Johnson, J., & Janke, C. (2008). Does interrater (dis)agreement on Psychopathy Checklist scores in sexually violent predator trials suggest partisan allegiance in forensic evaluation? *Law and Human Behavior, 32,* 352–362.

Murrie, D. C., Boccaccini, M. T., & Turner, D. T. (in press). Ethical challenges in sex-offender civil commitment evaluations: Applying imperfect science in adversarial proceedings. In A. Schlank (Ed.), *The sexual predator* (Vol. 4). Kingston, NJ: Civic Research Institute.

Murrie, D. C., Boccaccini, M. T., Zapf, P. A., Warren, J. I., & Henderson, C. E. (2008). Clinician variation in findings of trial competence. *Psychology, Public Policy, & Law, 14,* 179–193.

Murrie, D. C., & Warren, J. I. (2005). Clinician variation in rates of legal sanity opinions: Implications for self-monitoring. *Professional Psychology: Research and Practice, 36,* 519–524.

Norusis, M. (2003). *SPSS 12.0 statistical procedures companion.* Upper Saddle, NJ: Prentice Hall.

Quinsey, V. L., Rice, M. E., & Harris, G. T. (1995). Actuarial prediction of sexual recidivism. *Journal of Interpersonal Violence, 10,* 85–105.

Rice, M. E., Harris, G. T., & Quinsey, V. L. (1990). A follow-up of rapists assesses in a maximum-security psychiatric facility. *Journal of Interpersonal Violence, 5,* 435–448.

Rogers, R. (1987). Ethical dilemmas in forensic evaluations. *Behavioral Sciences & the Law, 5,* 149–160.

Rosenthal, R. (1978). How often are our numbers wrong? *American Psychologist, 33,* 1005–1008.

Rutherford, M., Cacciola, J. S., Alterman, A. I., McKay, J. R., & Cook, T. G. (1999). The 2-year test-retest reliability of the Psychopathy Checklist—Revised in methadone patients. *Assessment, 6,* 285–291.

Seto, M. C. (2005). Is more better? Combining actuarial risk scales to predict recidivism among adult sex offenders. *Psychological Assessment, 17,* 156–167.

Shalverson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Shalverson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44,* 922–932.

Shuman, D. W., & Greenberg, S. A. (2003). The expert witness, the adversary system, and the voice of reason: Reconciling impartiality and advocacy. *Professional Psychology: Research & Practice, 34,* 219–224.

United States v. Alvarez, 519 F. 2d 1036 (3rd Cir. 1975).

Vrieze, S. I., & Grove, W. M. (2008). Predicting sex offender recidivism: I. Correcting for item overselection and accuracy overestimation in scale development. II. Sampling error-induced attenuation of predictive validity over base rate information. *Law and Human Behavior, 32,* 266–278.

Wollert, R. (2002). The importance of cross-validation in actuarial test construction:

Shrinkage in the risk estimates for the Minnesota Sex Offender Screening Tool—Revised. *Journal of Threat Assessment, 2,* 87–102.

Wollert, R. (2003). Additional flaws in the Minnesota Sex Offender Screening Tool—Revised: A response to Doren and Dow (2002). *Journal of Threat Assessment, 2,* 65–78.

Wollert, R. (2006). Low base rates limit expert certainty when current actuarials are used to identify sexually violent predators: An application of Bayes's Theorem. *Psychology, Public Policy, and Law, 12,* 56–85.

Wood, J., Nezworski, M., & Stejskal, W. (1996). The Comprehensive System for the Rorschach: A critical examination. *Psychological Science, 7,* 3–10.

Received August 20, 2008
Revision received October 23, 2008
Accepted November 24, 2008 ■

GLENN LANGENBURG

Glenn Langenburg is a certified latent print examiner at the Minnesota Bureau of Criminal Apprehension and most recently, became a Forensic Science Supervisor of its Drug Chemistry Section. Glenn also manages a private consulting business (Elite Forensic Services, LLC). He has experience with crime scenes and bloodstain pattern evidence and is certified as a general criminalist by the American Board of Criminalistics.

Glenn has a Ph.D. in Forensic Science from the University of Lausanne in Switzerland. His thesis, "A Critical Analysis and Study of the ACE-V Process," focused on decision-making and the application of ACE-V by fingerprint experts. He has lectured and hosted workshops nationally and internationally at forensic science conferences in the United States, Canada, and Europe on topics including *Daubert* issues, research, probabilistic approach, error rates, and fingerprint methodology. He has published numerous research articles in peer- reviewed journals.

Glenn had the privilege of serving the fingerprint community for 10 years as a member of SWGFAST (Scientific Working Group for Friction Ridge Analysis, Study, and Technology). He also co-hosts "The Double Loop Podcast," a weekly podcast on fingerprint topics.

Forensic Science Policy & Management: An International Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ufpm20

A Report of Statistics from Latent Print Casework

Glenn Langenburg[a], Flore Bochet[b] & Scott Ford[c]
[a] Minnesota Bureau of Criminal Apprehension, St. Paul, Minnesota
[b] Brigade de Police Technique et Scientifique, Police Cantonale, Geneva, Switzerland
[c] Tri County Regional Forensic Laboratory, Andover, Minnesota
Published online: 21 Jul 2014.

To cite this article: Glenn Langenburg, Flore Bochet & Scott Ford (2014) A Report of Statistics from Latent Print Casework, Forensic Science Policy & Management: An International Journal, 5:1-2, 15-37

To link to this article: http://dx.doi.org/10.1080/19409044.2014.929759

Forensic Science Policy & Management, 5(1–2):15–37, 2014

ISSN: 1940-9044 print / 1940-9036 online
DOI: 10.1080/19409044.2014.929759

A Report of Statistics from Latent Print Casework

Glenn Langenburg,[1] Flore Bochet,[2] and Scott Ford[3]
[1]Minnesota Bureau of Criminal Apprehension, St. Paul, Minnesota
[2]Brigade de Police Technique et Scientifique, Police Cantonale, Geneva, Switzerland
[3]Tri County Regional Forensic Laboratory, Andover, Minnesota

ABSTRACT Statistics were derived from casework from the Minnesota Bureau of Criminal Apprehension Latent Print Unit. These data represented a portion of the latent print casework completed in the 2003/2004 calendar years (N = 673 cases) and 2009/2010 calendar years (N = 885 cases). The 2003/2004 data revealed latent print recovery rates from various exhibits. Identifiable latent prints were recovered 13% of the time on firearms, 13% of the time on plastic bags, and no identifiable latent prints were recovered from fired or unfired ammunition. The processing of evidence prior to submission to the laboratory increased the chances of latent print recovery. Both data sets were explored for the rate at which identifiable latent prints were reported (61% of cases in 2003/2004 and 54% of cases in 2009/2010) and the rate at which identifiable latent prints were subsequently identified to an individual in the case (23% of cases in 2003/2004 and 25% of cases in 2009/2010). There was no noticeable difference for the identification rate in property crimes versus crimes against people.

The 2009/2010 data were explored for possible effects from analysts having access to contextual information or significant interaction and communication with police officers or prosecutors while working a case. We noted that 2% of cases in the data qualified for this condition—the majority of BCA-LPU cases are worked without contextual information or police interaction. Comparing high context/high interaction cases versus no context/no interaction cases, we found the latent print identification rates to be equal (21% versus 22%, respectively).

KEYWORDS Fingerprints, bias, statistics, recovery rates, firearms, ammunition

Received 13 March 2014;
accepted 27 May 2014.

Address correspondence to Glenn Langenburg, Minnesota Bureau of Criminal Apprehension, 1430 Maryland Avenue East, Saint Paul, Minnesota 55106. E-mail: glenn.langenburg@state.mn.us

Color versions of one or more figures in this article can be found online at www.tandfonline.com/ufpm.

INTRODUCTION

Finding a source for detailed and accurate fingerprint evidence from a crime lab can be difficult. While some sources have provided general trends for forensic service providers, proficiency testing results, or crime justice statistics (5; Peterson et al. 2013), few crime labs actually publish data from their case results. Elsewhere, we have reported data from a field study that focused on the volume of unrecovered evidence and its potential weight of evidence (Neumann et al.

2011), but that study did not examine elements such as recovery rates from various processing techniques, submission trends, AFIS use and success, etc. The aim of the present paper is to provide casework statistics, such as latent print recovery rates and rates of identification, that one would find in a fingerprint laboratory.

With respect to latent print recovery rates, recovery rates on firearms and ammunition in *actual casework* have been reported elsewhere (Barnum and Klasey 1997; Johnson 2010; Pratt 2012; Maldonado 2012). These sources noted consistent recovery rates of 11%, 12%, 10%, and 13%, respectively, for firearms or magazines from firearms, depending on the study. We wish to contribute to those data as well, while adding another layer of information by further subcategorizing our firearms, as was done by Pratt (2012). Recovery rates of latent prints from plastic bags from casework have not been reported to date.

A portion of the present paper was dedicated to the exploration of possible bias effects from significant interaction between the forensic analyst and the case investigator, or from analyst exposure to contextual information about the case—information which has nothing to do with the processing of the evidence. Much has been made of these interactions, and there is general concern for the influence it may have on the accuracy of the results from a crime lab (Kassin, Dror, and Kukucka 2013; Dror 2013; Dror and Hampikian 2011). Yet to date, no source has demonstrated that, in a crime lab that works a high volume of cases, these errors are frequent and exposure to contextual case information is to blame. Contrived research, anecdotal cases, and miscarriages of justice have showcased these dangers (Office of the Inspector General [OIG] 2006; Cole 2006; Dror and Charlton 2006). Yet, in comparable non-forensic, diagnostic testing domains, such as radiological diagnostic testing, there is considerable debate about the advantages and disadvantages of making patient clinical history available to the radiology technician to render an accurate and efficient assessment of the case (Potchen et al. 1979; Potchen et al. 2000; Loy and Irwig 2004; Dhingsa et al. 2004). Furthermore, some research in the forensic domain has pointed toward the benefits of information exchange between analysts and investigators (8; 9; 3; Roberts and Willmore 1993), while still acknowledging the pitfalls of bias effects. This has prompted some authors to argue that shielding a forensic analyst from case information or failing to consider the evidence in the context of the specific case may in fact lead to more error or missed opportunities to critically evaluate the evidence (1; Thornton 2010). They argue, generally, that forensic scientists should enter a professional dialogue with the investigator to develop an appropriate resource-conscious forensic strategy. This strategy can limit the examination and testing just to those evidential items which can impact the investigation.

In the midst of this debate, there has been a call for better quality assurance measures to prevent domain irrelevant information exchange between the analyst and the investigator (National Research Council 2009). These suggested measures have ranged from blinding the analyst from all domain irrelevant information in every case (Haber and Haber 2008) to a sequential unmasking approach, whereby case information is revealed ("unmasked") after critical decision making stages have been completed (Krane et al. 2008). In this scheme, the analyst will eventually have access to all the case information, but only after it cannot influence the analyst's decision. Other variations to these schemes have been proposed such as blind verification in select cases (Cole 2013) or evidence line-up/distractor sample approaches (Wells, Wilford, and Smalarz 2013). Typically, these quality assurance measures must be introduced by a case coordinator, who assigns the case, filters information, and acts as a liaison between the analyst and investigator. This approach raises some questions such as: 1) which information should be kept from an analyst? 2) what if contextual case information could help the analyst make more accurate, efficient, and informed decisions about the case? and 3) at what cost (both monetarily and in terms of benefits versus risks) do these changes bring? (Langenburg 2012).

The present paper explores these issues and identifies which cases may actually present the most danger of error from bias. This will give a clearer picture of what resources are required to address this issue or where best to concentrate efforts and quality assurance measures to limit bias effects.

Demographics of Minnesota and the BCA-LPU

The data in the present paper represent samplings from actual casework for the Minnesota Bureau of Criminal Apprehension Latent Print Unit (BCA-LPU). To properly assess these data, it is important to

understand how the BCA-LPU operates and what are the characteristics of BCA-LPU, the BCA in general, and the State of Minnesota. Before comparing data between agencies, it is important to ensure that what constitutes a "case," similar workflows, and similar processes are compared for a fair apples-to-apples comparison.

There are approximately 5.3 million people living in Minnesota (United States Census Bureau 2012). About 60% of the population (3 million people) live in the Minneapolis/St. Paul metropolitan area and suburbs, and the remainder of the population is spread throughout the mostly rural farmland or heavily wooded and lake abundant state.

The BCA-LPU is the latent fingerprint section for the State of Minnesota. The BCA-LPU services 87 counties. In actuality, since the two largest metropolitan areas in Minnesota, St. Paul and Minneapolis, have their own latent print units, the BCA-LPU does not routinely receive requests from these agencies. In effect, the BCA-LPU receives the cases from the greater metropolitan area and the rest of the State of Minnesota. The BCA-LPU is comprised of two laboratories: the headquarters laboratory in St. Paul and a regional laboratory in Bemidji. The St. Paul lab services the lower half of the state and the metropolitan area and the Bemidji lab services the upper half of the state, which is more rural and less populated. The BCA-LPU currently employs seven analysts (two in the satellite lab and five in the central headquarters). The range of experience of these analysts is from 4 years to 25 years in latent prints. The BCA-LPU is part of an accredited laboratory system, under ISO17025, and offers other testing services (e.g. DNA, firearms, etc.). All of the BCA-LPU analysts are certified latent print examiners by the International Association for Identification (IAI).

The BCA-LPU provides processing, comparison, and AFIS services. Analysts typically process their own evidence, perform photography of any identifiable latent prints, perform the comparisons, enter unidentified latent prints into AFIS, and write their reports. The BCA laboratory offers on a voluntary basis, the opportunity to join the BCA Crime Scene Team, which primarily assists local law enforcement when requested on homicides, kidnapping, officer-involved shootings, etc. Many of the BCA-LPU serve/have served on this team. This is relevant because, in those cases, the attending analyst often will also be the case-working fingerprint analyst. The authors anticipate that the readers will have mixed feelings about this. On the one hand, the attending analyst understands why and how the latent print evidence was collected and which evidence is most critical. On the other hand, there may be concern that this level of interaction, exposure to contextual information, and perhaps even emotional investment, may influence an analyst's decision in the case. The authors specifically wanted to explore that issue in this paper as well.

In the vast majority (over 99%) of the cases received by the BCA-LPU, the analysts receive case submissions from local law enforcement. These local police and sheriff departments have responded to a scene, collected (and possibly processed to some extent) evidence, and submitted it to BCA. The delay of evidence submitted to the BCA-LPU can vary from a few days to sometimes more than a year or more after the crime. Because the evidence is received by "Evidence Specialists," who take evidence into the BCA for all the forensic sections at the BCA, the BCA-LPU analysts rarely have contact with a submitter at the time of delivery. In the course of working the case, the analyst may have a need to contact the investigator with follow-up questions. These questions may occur at the beginning of the process (e.g. "which of these 100 items should I start processing first?" "this person does not have a fingerprint record against which to compare") or near the end of the process (e.g. "I have identified the suspect in the case several times, do I need to continue to compare all the remaining 20 latent prints to this suspect too?"). Often these questions help the analyst to allocate their time and resources effectively. The concern by some commentators is that in the course of those conversations, the potential to be exposed to biasing contextual information exists (Dror, Charlton, and Péron 2005; Mnookin 2010).

The BCA-LPU received approximately 1400 case submissions for the 2012 calendar year. This submission rate has steadily climbed over the last 10 years. The submission rate was around 1,000 to 1,100 cases ten years ago. A Bureau of Justice (BJS) survey in 2005 reported the median number of latent print examination requests in the U.S. was 909 cases for the 194 agencies that responded to the BJS survey (Durose 2008). This places the BCA-LPU slightly above those submission rates, and certainly these numbers have increased since 2005.

MATERIALS AND METHODS

For the present paper, two data sets were prepared by random sampling of completed BCA-LPU cases. The first data set, which focused on recovery rates, is referred to as the 2003/2004 data set. The second data set, which focused on rates of identification, impact of AFIS, and effect of exposure to case context information and interactions between forensic analysts and investigators, is referred to as the 2009/2010 data set.

The 2003/2004 set was prepared by sampling 673 cases from a 12-month period of cases worked by the BCA LPU in mid-2003 through mid-2004. At that time, the BCA LPU was working about 1,000 to 1,100 cases per year. This sample is about two-thirds of the cases worked in that time period. Specifically, the sample represented about 50% of the cases worked in the St. Paul laboratory, and about 70% of the cases worked in the Bemidji laboratory in this time period.

Data were collected through the use of a data sheet prepared for each case. At the end of the data collection period, the data were entered into a Microsoft Access (2003) database for analysis.

The 2009/2010 data set was prepared by sampling 885 cases from a 12-month period of cases worked by the BCA LPU in 2009 and 2010. There were approximately 1,200 cases per year received by the BCA in 2009 and 2010. This sampling represented approximately 75% of the cases worked in St. Paul and 30% of the cases worked in Bemidji. Caution is warranted when comparing the data from 2003/2004 to the data in 2009/2010; proportions should be compared to minimize sampling and population size differences.

The BCA codes a case during its submission based on the submitting officer's description of the case. For the 2003/2004 and 2009/2010 data sets, we pooled case types together to identify four classes of case type. These are:

1) Property crimes: includes burglary, theft, auto theft, fire investigation, forgery, fraud, stolen property, and vandalism.
2) Crimes against people: includes cases with death investigation, homicide, attempted homicide, robbery, criminal sexual conduct, assault, kidnapping, threats, stalking, hit and run, etc.
3) Drugs: includes controlled substances with possession, sale, or manufacture.
4) Weapons: includes cases with unlawful discharge or unlawful possession of a firearm.

In the 2009/2010 data, we assessed the level of interaction between the case analyst and the police/investigator(s)/prosecutor(s). We also assessed the amount of contextual information, such as police reports or investigative information, available to the analyst in the case. To collect these data, a worksheet was completed for each of the sampled cases by reviewing the case reports. We also reviewed the LIMS (Laboratory Information Management System), which tracks case information and would include such things as communiqués between the analysts and investigators, police reports available to the analyst at the time of the examination, and notes regarding the analysts' observations or decisions in a case.

We categorized the level of interaction as "high," "moderate," or "none/minimal." The level of interaction was deemed "high," "moderate," or "none/minimal" based on the following criteria:

- High = significant interaction between investigators or prosecutor, resulting from at least 3 phone calls, at least 3 email exchanges, or attendance at the crime scene.
- Moderate = 1-2 email or phone call exchanges between submitting officer(s), prosecutor(s), or investigator(s) typically where case information and details are exchanged.
- None/minimal = no recorded contact with submitting officer(s), prosecutor(s), or investigator(s), or minimal contact to clarify a case question (e.g., an e-mail to check the spelling or date of birth of a suspect, a phone call asking if the item had already been processed, etc.

This assignment was obviously a judgment call of the researchers. If there was any doubt, and any case information appeared to be exchanged with the analyst and the requesting parties, then the case was classified at a minimum as "moderate" interaction. We also considered the reading of case information to be a type of "interaction." If it was clear in the LIMS that the analyst had read considerable case information (high or moderate context report) then the level of interaction would be increased one level (i.e., "none/minimal" interaction was raised to

"moderate" if the analyst clearly read a detailed case report). Although it should be noted that it was only clear in 9 of 885 cases in the LIMS that the analyst had read the report. It was also possible that a detailed report was present, but it was not read by the analyst.

The amount of contextual information available to the case analyst was categorized as "high," "moderate," or "none/minimal." The level of contextual information was deemed "high," "moderate," or "none/minimal" based on the following criteria:

- High = significant case details were available in LIMS. Typically, in "high context" cases, officers have submitted detailed reports about the scene or the investigation. These reports may include investigator theories, detailed interviews with suspects, suspect statements, or details and observations made by investigators at the crime scene or during collection of the evidence. Cases where the analyst attended the crime scene were also deemed "high context."
- Moderate = short reports or details about the crime or investigation were provided by the investigator in addition to the standard submission forms required by BCA.
- None/minimal = no case details were provided at all, or only minor, domain relevant information, or required information for case submission were provided on standard BCA submission forms.

RESULTS AND DISCUSSION

Latent Print Submissions by Case Type

The BCA LPU received approximately 1,000 to 1,200 case assignments per year in the considered time frames for both data sets. Recent submission rates for 2011 and 2012 have increased by 20% to approximately 1400 per year.

The distribution of cases for both the 2003/2004 data set (recovery rate data set) and the 2009/2010 data set (conclusion rate data set) is shown in Figure 1 below. It can be seen that property crimes were the most common case type submissions for latent prints. The BCA-LPU received over four times as many property crimes as crimes against people or drug cases.

Care must be taken when comparing the two data sets in Figure 1. The two samples have different sizes, $N = 673$ and $N = 885$. A two-sample Z test for proportions can be used to assess the statistical significance of the difference in submissions between the two data sets. There was a significant increase ($Z = -4.18$; $p < 0.001$) for property crime cases from 2003/2004 to 2009/2010; 432 out of 673 cases (64%) in 2003/2004 were property crimes compared to 655 out of 885 cases (74%) in 2009/2010. Simultaneously, there was also a significant decrease ($Z = 3.15$; $p = 0.002$) in crimes against persons submissions for latent print analysis. The differences between the number of weapons and drugs submissions between the data sets were not statistically significant ($p > 0.05$). The shift in property

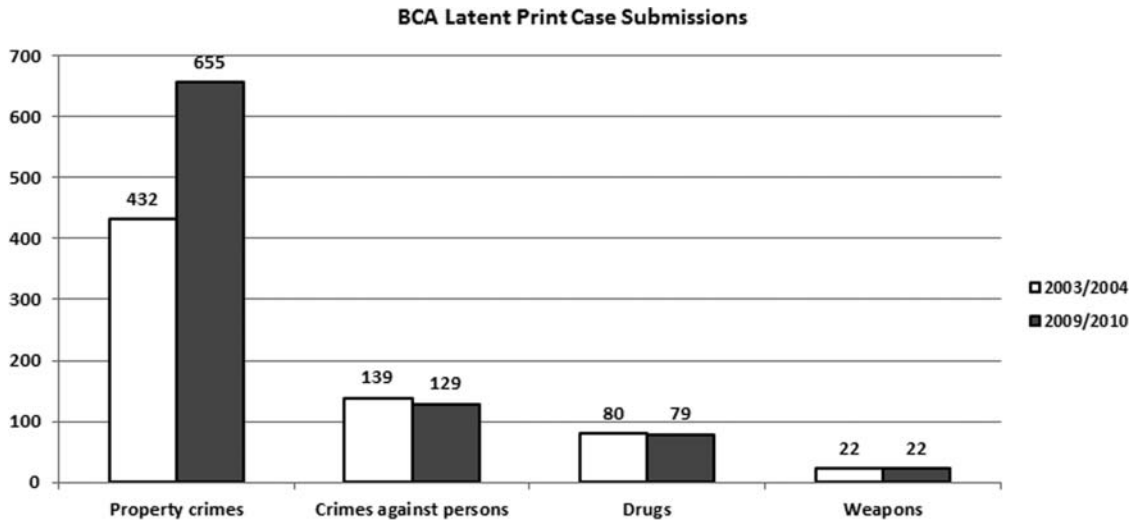


FIGURE 1 Comparison of case submissions requesting latent print examinations by category of case type.

crimes and crimes against people may be due to changes in the types of cases submitted for DNA analysis.

From 2003 to 2009, the BCA saw a significant increase in property crime cases submitted for DNA analysis. In 2003, the BCA received 1,714 DNA assignments; 224 (13%) were property crime cases. In 2009, the BCA received 3,407 DNA assignments; 907 (27%) were property crimes. The sheer volume of casework for DNA had doubled, but the proportion of property crimes for DNA analysis had also doubled. In many of these cases, latent prints were also being requested by the submitters, or a DNA analyst at the BCA would recommend latent print examinations to the submitter in lieu of, or sometimes in addition to, DNA examinations. This collateral effect is clearly seen in Figure 1, both in the increase in submissions, but also in the increased proportion of property crimes. We refer to this as the "DNA trickle down" effect.

It should also be noted that, in 2003, there were 15 BCA DNA analysts to work the 1,714 submissions. In 2009, when the number of DNA submissions doubled to 3,407, the number of BCA DNA analysts had also nearly doubled to 27. In 2003, the BCA LPU had 7 fingerprint analysts In 2009, the BCA LPU had 7 fingerprint analysts. Today at 300 more submissions annually than in 2009, the BCA LPU has 6 (and a half timer) fingerprint analysts.

While funding and backlog reduction funds (e.g., Coverdell grant) have been prioritized for DNA laboratories in the U.S., the same cannot be said for most latent print units. Unfortunately, the latent print sections have not received the benefit of funding and personnel to match their DNA counterparts. As a result, the increase in DNA testing requests has increased the burden on the latent print section without a commensurate investment in latent print personnel or resources.

Identifiable Latent Prints and Identification Rates

When determining the intrinsic value of latent print evidence, an analyst at the BCA-LPU will first note the presence of ridge detail, if any, observed on the exhibit. Then the analyst will determine its "suitability" (or in some agencies "value") for comparison. This is the analyst's judgment of the utility of the impression and the likelihood that they will be able to reach a definitive conclusion ("identification" or "exclusion"). Agencies will vary in how they apply this approach as noted by SWGFAST standards (Scientific Working Group on Friction Ridge Analysis Study and Technology [SWGFAST] 2013). BCA-LPU subscribes to Approach #2 as described in those standards, whereby most impressions are compared with the expectation that they can be identified when presented with the correct source exemplars, but not in all cases. In some cases, the correspondence may be insufficient and an "inconclusive" opinion due to the limited information in the latent print, may be rendered. For the non-technical reader, we have opted for the remainder of the paper to refer to these latent prints that have been deemed comparable by the analyst as "identifiable," although in actual practice at the BCA-LPU we use the term "suitable for comparison." Finally, it must be clarified, that this decision of "suitability" takes place before ever viewing the exemplars of any of the subjects in the case; it takes place during the analysis stage of the Analysis-Comparison-Evaluation-Verification (ACE-V) process (Langenburg and Champod 2011).

In the 2003/2004 data, we recorded if the analyst observed "any ridge detail." This would include cases where ridge detail was observed by the analyst, but not recovered due to the perceived inability to exploit the ridge detail. This question was not asked in 2009/2010, although cases from 2009 comprised the data set used in a previous study (Neumann et al. 2011) where the amount of unrecovered ridge detail was quantified and explored. In the 2009/2010 data, we were only concerned with the proportion of cases with identifiable latent prints. Lastly we examined the proportion of these cases where the identifiable latent prints resulted in "identification" decisions to either the victims or the suspects. The distinction between victim and suspect identifications was not made in the 2003/2004 data, but was explored in detail in the 2009/2010 data. These data are shown in Table 1 and they are further deconstructed by case type.

In the 2003/2004 data, 575 out of 673 (85%) cases had at least one item of evidence that bore some visible ridge detail for the analyst to evaluate for its potential "value." Of these 575 cases, 410 (410 out of 673 total cases = 61%) resulted in latent prints deemed "identifiable." Finally, for these 410 cases where suitable ridge detail was observed, 152 cases

TABLE 1 The Proportion of Cases with Identifiable Latent Prints and "Identification" Decisions are Compared Between the 2003/2004 and 2009/2010 Data Sets. Percentages Reported are Using the Total Number of Considered Cases (N = 673 and N = 885) as the Denominator

	2003/2004 (N = 673)			2009/2010 (N = 885)	
	Number of cases with any ridge detail observed	Number of cases with identifiable latent prints	Number of cases with "identification" reported	Number of cases with identifiable latent prints	Number of cases with "identification" reported
Property crime	398 (59%)	304 (45%)	101 (15%)	384 (43%)	167 (19%)
Drugs	53 (8%)	31 (5%)	17 (3%)	26 (3%)	14 (2%)
Weapons	14 (2%)	4 (<1%)	4 (<1%)	4 (<1%)	1 (<1%)
Crime against persons	110 (16%)	71 (11%)	30 (4%)	66 (8%)	40 (5%)
Total	575 (85%)	410 (61%)	152 (23%)	480 (54%)	222 (25%)

(152 out of 673 total cases = 23%) had at least one "identification" decision.

In the 2009/2010 data, 480 out of 885 (54%) cases bore at least one latent print deemed identifiable. If we compare this to the 2003/2004 data, we see there is a drop from 61% to 54% of cases with identifiable latent prints. This is a statistically significant decrease ($Z = 2.64$: $p = 0.008$), and may be due in part to the previously discussed "DNA trickle down" effect from increased property crime submissions for both DNA and latent prints. It may be possible that some of these exhibits selected for DNA testing may not have been the most appropriate or conducive for latent print evidence, but since the exhibit has been submitted for DNA, the officer requests latent print examination to be done anyway. There is no actual cost to the officer or prosecuting attorney and these decisions may not always be carefully considered. It may be one of the factors leading to some of the observed backlogs in crime labs (Durose 2008). Perhaps an approach closer to the "case assessment model" as proposed by Cook, et al. (1998) may lead to better screening and evidential choices. A discussion between the scientist and the investigator may allow for better choices when selecting which items to analyze, or which tests to perform, despite the potential risk of bias.

Recovery Rates From Various Exhibits

The rate of recovery of latent prints from various substrates and exhibit types was not explored in 2009/2010, therefore the data below only represent the 2003/2004 dataset. The cases were sorted into three categories:

1) Lifts only: these were cases where latent prints were recovered at the scene only by tape lifts or photographs. No exhibits to examine or process were submitted.
2) BCA processing: these cases required processing of exhibits by the BCA. They are the most time consuming due to the sequential application of different development techniques.
3) Submitting agency processed: these cases had exhibits that were processed by technicians prior to the submission to BCA. Processing may have occurred in the field or at the submitter's agency.

Figure 2 shows the relative proportions of cases where "lifts only," "submitter processing," or "BCA processing" was performed. Of the 673 reviewed cases, 330 cases (49%) were cases where only lifts were submitted from evidence technicians in the field, 288 cases (43%) were cases were BCA was required to process exhibits, and 55 cases (8%) were cases where the evidence technician did the processing before submitting the exhibit. Roughly speaking then, about half the cases submitted to BCA required no processing, while half the cases required some processing and/or photography.

When we examined the effect of processing by technicians prior to submission, we see in Figure 3, that the lift cases bore identifiable latent prints 77% of the time, while the submission of the exhibit only for BCA processing produced identifiable latent prints 41% of the time. Where the submitter performed processing of the exhibit prior to submission, identifiable latent prints were recovered 67% of the time. One of the explanations for this difference may be that the submitters processed many more items that were not

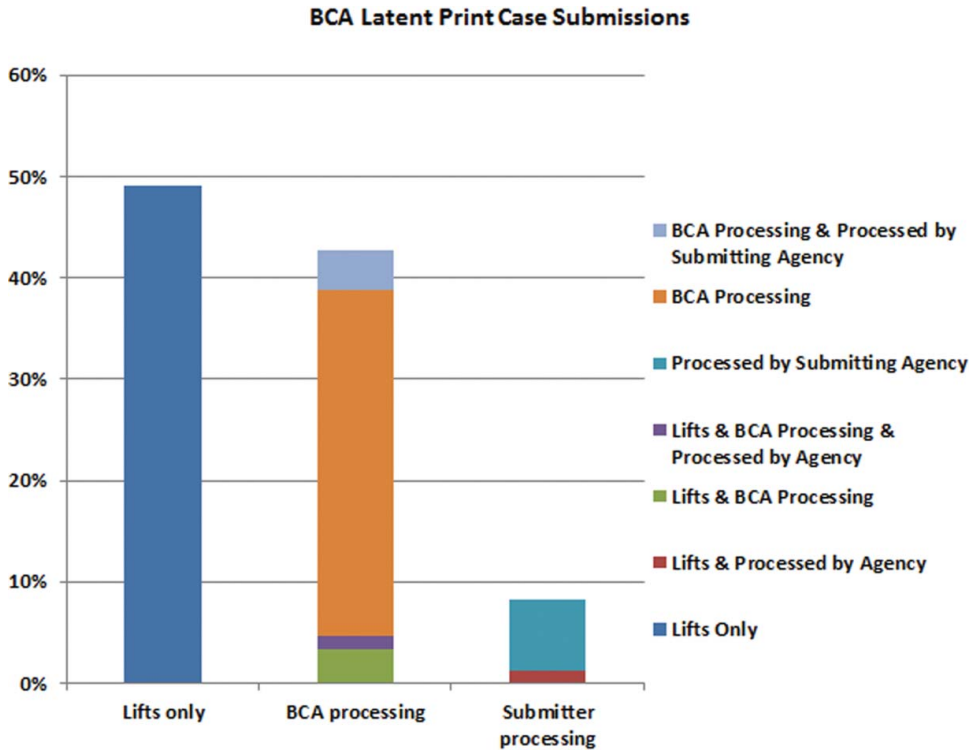


FIGURE 2 Distribution of cases in the 2003/2004 data set by level of pre-processing performed by submitters to BCA.

submitted; they only submitted those items where they observed some apparent ridge detail. The same would be true with respect to lifts. This may demonstrate an efficient selection of exhibits, both for the presence of useable ridge detail, but also for the purpose of choosing to process the exhibit in the first place. In other words, field technicians may be making good choices about what exhibits to process and which to submit.

Another explanation (not mutually exclusive) for the high recovery of identifiable latent prints from pre-processed exhibits is that preservation in the field, or after a relatively short time from the deposition of the latent print, may increase the recovery rates due to the fragility and volatility of latent print residues. While the crime lab may have premier equipment and expertise in the development of latent prints, these

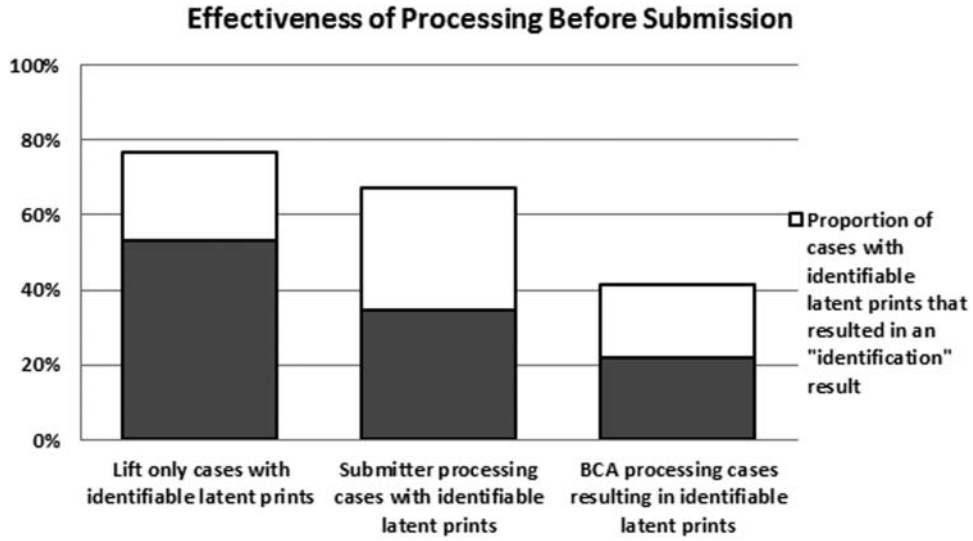


FIGURE 3 The percentage of cases that resulted in identifiable latent prints and the fraction of those cases with identifiable latent prints that resulted in an "identification" decision reported.

advantages may be lost when the evidence sits for several months before being processed due to delays in submission and case backlogs. Lastly there may be some potential loss of evidence during the collection, packaging, and transportation of the unprocessed evidence to the crime laboratory.

Figure 3 also shows the relative proportion of cases with identifiable latent prints which subsequently led to at least one "identification" decision in the case. It can be seen in Figure 3 that while lift cases produced identifiable latent prints 77% of the time, only about one-third (31%) of these cases led to an "identification." In the cases where the BCA processed the item or the submitter processed the item, identifiable latent prints were recovered about half the time (47% and 49% respectively). A possible explanation for this difference is, again, the relevance of the exhibit. In lift cases, lifts may often come from immovable objects in public places or with unrestricted access (doors, counters, windows, tables, vehicles, vending machines, Automatic Bank Teller Machines, etc.). Many individuals without relation to the crime could have touched these surfaces from which the lifts were generated. Whereas the choice to process an exhibit with cyanoacrylate, ninhydrin, etc. may be with an eye towards a very relevant object related to the crime, with limited access to a handful of individuals.

The BCA has four major protocols for processing evidence depending on the type of surface and latent print residue that may be deposited on the substrate. These processing protocols are: 1) non-porous (e.g. glass, plastic, metal, etc.); 2) porous (e.g. papers, checks, cardboard, etc.); 3) adhesive (duct tape, stickers, stamps, etc.); 4) blood processing (enhancement of visible ridge detail deposited with blood matrix). Figure 4 shows that in the 288 cases where BCA processing was required, non-porous processing (N = 206) is the most commonly used processing protocol at BCA, followed by porous processing (N = 51). Some cases (N = 29) required the use of multiple processing techniques. Typically, when porous processing or multiple techniques are required, these cases become more labor intensive and time-consuming. Also, most exhibits where tape is involved require multiple processes (i.e. non-porous and adhesive processing).

In the 2003/2004 data, we collected information about the recovery rates of identifiable latent prints from various non-porous exhibits. We did not look at recovery rates for porous, blood, or adhesive processing cases. We investigated latent print recovery rates for three categories of exhibits: 1) plastic bags, 2) firearms, and 3) ammunition for firearms (see Table 2).

In the 45 cases where plastic bags were submitted, 201 plastic bags exhibits were processed for latent prints. Twenty-six (26) identifiable latent prints were recovered from these 201 plastic bags. This is an average recovery rate of 13% for the plastic bags. It should be noted that the true recovery rate may actually be lower since some of these 26 identifiable latent prints

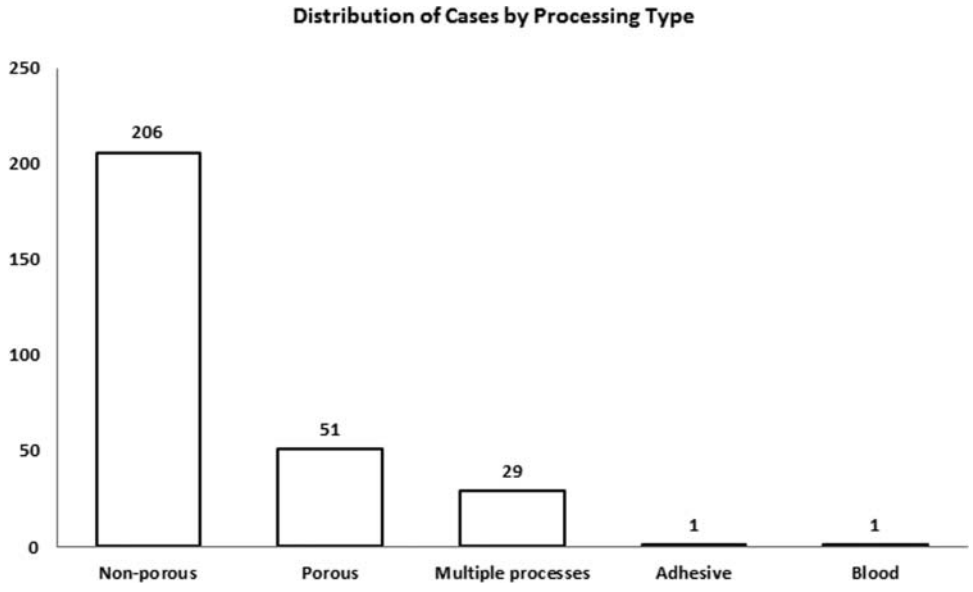


FIGURE 4 The distribution of the cases submitted to BCA (N = 288) for processing in the 2003/2004 data set (N = 673).

TABLE 2 Distribution of Cases and Exhibits that were Processed at BCA in the 2003/2004 Data Set

	Plastic bags	Firearms	Ammunitions
Number of cases with selected exhibit type	45	73	40
Number of exhibits processed	201	104	341
Number of identifiable latent prints	26	14	0
Identifiable latent print recovery rate	13%	13%	0%
Number of "identification" decisions reported	14	5	0

were found on the same bag. In other words, 13% recovery rate is likely an overestimate. In the 73 cases where firearms evidence was submitted, 104 firearms were processed for latent prints. Fourteen (14) identifiable latent prints were recovered from these 104 firearms. This is an average recovery rate of 13%. Again, this may be a slight overestimate if multiple latent prints were found on the same firearm, but these data are similar to other reported sources (0; 2; Pratt 2012). Finally, 40 cases were submitted for the processing of firearms ammunition. In 341 exhibits, no identifiable latent prints were recovered. This exceedingly low probability of success for latent print recovery on ammunition is also noted by the same aforementioned sources.

The low recovery rates from ammunition raises two important points. The first point is that given the low (non-existent) success of latent print processing techniques on ammunition, perhaps these exhibits should be going exclusively for DNA testing. Recovery of DNA from cartridges and cartridge cases, while still low and often involves mixtures or low-quantity DNA (1; Horsman-Hall et al. 2009), is still more successful on average than latent print processing. The second point is that these low recovery rates, in contrast to the constant success of our fictional TV counterparts, is likely contributing to the increased demand for what is referred colloquially by examiners as "negative testimony"—testimony in jury trials to address the question of why no identifiable latent prints were recovered from the exhibit(s).

We explored the sub-classification of plastic bag, firearms, and ammunition exhibits as shown in Tables 3, 4, and 5. The 201 plastic bag exhibits consisted of plastic bags of various types and size (see Figure 5). The 104 firearms exhibits consisted of pistols, revolvers, shotguns, and rifles (see Figure 6). The 341 ammunition exhibits consisted of various caliber fired and unfired ammunition. In the plastic bag category, it can be seen that Ziploc bags and garbage bags were the most successful for latent print recovery. This is likely due to the larger surface area and generally smoother surface of these exhibits. In the firearms category, recovery rates were higher for rifles over shotguns. Revolvers gave the highest recovery rates for all the firearms. Lastly, in the ammunition category, neither the cartridge, nor the cartridge case was a substrate conducive to the development of latent prints. Anecdotally, in the tens of thousands of cartridges and cartridge cases processed at the BCA in the last 30 years, only a handful of identifiable latent prints have been recovered. These tended to be on large caliber rifle or shotgun ammunition. Therefore, these results for ammunition processing are not surprising to us.

TABLE 3 Latent Print Recovery Rates for Plastic Bag Exhibits in the 2003/2004 Data Set

	Plastic bags		
	Ziploc bag	Sandwich bag	Garbage bag
Number of cases with selected exhibit type	30	20	3
Number of exhibits processed	133	65	3
Number of identifiable latent prints	22	2	2
Identifiable latent print recovery rate	17%	3%	67%
Number of "identification" decision reported	12	0	2

TABLE 4 Latent Print Recovery Rates for Firearms Exhibits in the 2003/2004 Data Set

	Firearms			
	Revolver	Pistol	Shotgun	Rifle
Number of cases with selected exhibit type	11	42	16	24
Number of exhibits processed	14	50	14	32
Number of identifiable latent prints	5	3	1	5
Identifiable latent print recovery rate	36%	6%	7%	16%
Number of "identification" decision reported	2	2	0	1

Conclusion Rate Data (2009–2010 cases)

The results in the following sections originate from the 2009/2010 data set. In these data, we primarily explored the distribution of conclusions reported by the analysts. We also explored the potential impact of case information and interaction with investigators. Lastly we identified and explored a subset of cases where a single latent print was recovered and associated with a suspect in the case. These issues were not explored in, and therefore not comparable to, the 2003/2004 data set.

Finger and Palm Print Distribution

As previously noted in Table 1, there were 480 cases (out of 885) cases with identifiable latent prints (see Table 1). In these 480 cases, there were a total of 1,446 identifiable latent prints that were recovered. Table 6 shows the distribution of whether they came from a finger, a palm, or a finger joint (including cases where the anatomical origin cannot be determined). We also investigated if these distributions were dependent on case type, i.e. was the analyst more likely to recover palm prints in a homicide than in a burglary. There was no significant change in the distribution per case type category (crimes against people, property crimes, drugs, weapons, other).

Approximately 1 in 7 recovered identifiable latent prints was a latent palm print. With respect to the rate of identification, latent fingerprints and latent palm prints were being identified at fairly similar rates (41% and 32%, respectively). This is in sharp contrast to the 11% rate for latent finger joints or "unknown" (when the analyst could not state with any certainty if the latent print was from a finger or palm due to the lack of anatomical or orientation focal points to associate with a finger or palm). There are two reasons (not mutually exclusive) for this. The first is that an analyst may have a better chance of finding the latent print "match" if he or she knows where to look. Since many comparisons are still being done manually by the analyst and without the aid of computers, the analyst must have a good idea where to look for the latent print, or search every conceivable area of friction ridge skin in each suspect or victim. The second reason is that these latent prints tend to be from areas of the skin not routinely captured during standard booking procedures and therefore the proper comparable area was not recorded. The exemplars are incomplete and a "match" is impossible.

An important point from these data is that a significant amount of latent print evidence originates from

TABLE 5 Latent Print Recovery Rates for Ammunition Exhibits in the 2003/2004 Data Set. A Cartridge is Unfired Ammunition; A Cartridge Case is the Case from a Fired Cartridge

	Ammunition	
	Cartridge	Cartridge case
Number of cases with selected exhibit type	31	22
Number of exhibits processed	253	88
Number of identifiable latent prints	0	0
Identifiable latent print recovery rate	0%	0%
Number of "identification" decision reported	0	0

FIGURE 5 **Examples of various types of plastic bag exhibits. These examples illustrate the style of bags categorized in the 2003/2004 data set as "sandwich bags" (left), "Ziploc bags" (center), and "garbage bags" (right).**

palm prints. There are still a number of agencies without the capabilities of searching palm print databases or recording palm prints during booking. There is no doubt that they are missing opportunities to identify suspects in cases. The need for specific palm print comparison training and the need for technology which capitalizes on palm print recording and databases is an absolute necessity.

Rates of Identification

We examined the number of latent prints that were deemed "identifiable" in four broad categories of case types: crimes against people, property crimes, weapons cases, and drugs. Table 7 shows the distribution for these case type categories. From Table 7 it can be seen that about half (54%) of the submitted cases to BCA in the 2009/2010 data set resulted in at least 1 "identifiable" latent print found on the evidence. These "identifiable" latent prints were predominantly found in property crimes and crimes against people (59% and 51% of those cases respectively). In drugs and weapons cases the chance of finding an "identifiable" latent print was significantly lower. This is consistent and explainable with the previously considered 2003/2004 latent print recovery data from drugs and weapons exhibits (see Table 1). It is interesting to also note that crimes against people and drug cases produced the most number of "identifiable" latent prints per case, although these cases represent a smaller fraction of all the cases submitted to BCA. This trend can be observed in Table 7.

FIGURE 6 **Examples of various types of firearms exhibits. These examples illustrate the style of firearms categorized in the 2003/2004 data set as "revolver (A)," "pistol (B)," "shotgun (C)," and "rifle (D)."**

TABLE 6 The Distribution of Identifiable Latent Prints that Originated from Fingers, Palms, or Finger Joints/Unknown

	Identifiable Latent Prints (N = 1446)		
	Fingers	Palms	Joints/Unknown
Number of identifiable latent prints (% of total)	1124 (78%)	221 (15%)	101 (7%)
Number that were identified (% of identifiable latent prints)	461 (41%)	72 (32%)	11 (11%)

In crimes against people cases, there may be several reasons for observing more "identifiable" latent prints per case. One reason is that a full battery of possible examinations are typically done in these types of cases, and only a single routine process may be employed in property crimes. Thus using all available sequential processes may result in more recovered latent prints. It may also be influenced by the relevance of the evidence collected by specialized and trained crime scene technicians. Another consideration, as suggested by some commentators, is motivation (Charlton, Fraser-Mackenzie, and Dror 2010). This is the notion that forensic analysts motivated by the severity of the crime will be more inclined to include marginal latent prints (thus "pushing the envelope") in a conscious or subconscious drive to aid investigators in serious and violent crimes. This issue will be explored in a later section of the present paper.

One argument against the notion of motivated analysts pushing the envelope to find marginal latent prints in more serious or violent crimes is the fact that Table 7 shows that in 51% of crimes against people identifiable latent prints were recovered. This can be compared to the 59% of property crimes where identifiable latent prints were recovered. One would expect a much higher percentage of identifiable latent prints claimed in crimes against people if the seriousness of the crime was influencing the analysts' decisions for value determinations. This is not to say that it is not occurring in some isolated incidents, but clearly there is not a trend here of rampant bias to include marginal latent prints in the "identifiable" category in these cases.

It is important to also consider, that although there is a small percentage of total cases (6%) with 6 or more identifiable latent prints, these cases tend to be very time consuming. When these cases have multiple suspects and victims against which to compare, a substantial amount of comparison time will be spent by the initial analyst and possibly, a second analyst who will have to verify the conclusions in a case. A more detailed analysis showed that these cases with more than 6 identifiable latent prints were predominantly homicide cases or stalking/harassment cases. Anecdotally, homicide cases tend to produce more exhibits and have more processing, and stalking cases tend to produce large amounts of identifiable latent prints often on a series of letters sent to the victim over time—many of which are handled by several people before finally involving the police.

Investigating further, we explored the rate of identification and exclusion decisions. Table 8 shows the

TABLE 7 Distribution of Identifiable Latent Prints Per Case Type Category

	Number of cases considered	Number of cases with at least 1 identifiable latent print (% of cases considered)	Total number of identifiable latent prints	Average number of identifiable latent prints per case (where there was at least 1 latent print)
Property crimes	655	384 (59%)	1014	2.6
Crimes against people	129	66 (51%)	307	4.6
Drugs	79	26 (33%)	114	4.4
Weapons	22	4 (18%)	11	2.8
Total	885	480 (54%)	1446	3.0

TABLE 8 Rate of "Identification" and "Exclusion" Decisions Sorted by Case Type Category

	Number of identifiable latent prints considered	Number of "identification" decisions	Number of "exclusion" decisions
Property Crimes	1014	370 (36%)	657
Crimes Against People	307	126 (41%)	511
Drugs	114	47 (41%)	69
Weapons	11	1 (9%)	17
Total	1446	544	1254

distribution of these rates for the four case type categories. We see that the rate of identification for property crimes, crimes against people and drugs cases are all approximately the same rates (36%, 41%, and 41%, respectively). This is evidence against the notion that analysts are more "motivated" to make (unwarranted) "identification" decisions in crimes against people because of their need to aid police (Charlton, Fraser-Mackenzie, and Dror 2010). Again, this does not preclude the possibility of occurrence in isolated incidents. Interestingly, the rate of "identification" decisions is exceptionally low and the rate of "exclusion" decisions is quite high in weapons cases, compared to the other case types in Table 8. One possible explanation for this is that police officers who are recovering these weapons (especially from a vehicle or off of the suspect) may not be wearing gloves since the primary purpose of the search may be to render their environment safe. In Minnesota, unfortunately, peace officers, fire and emergency personnel do not have their fingerprints in a non-criminal database (by Minnesota statute). Therefore, a number of these exhibits may have police officer prints on them, without any way of identifying the officer in the case. In Table 8, for all case types, we see that the total number of "exclusion" decisions are significantly greater than (by about 2.5 times) the total number of "identification" decisions. It is important to remember that a latent print can only be "identified" once, but a single latent print in the case can result in one "exclusion" decision ***per considered individual***. Therefore, this imbalance of "identification" versus "exclusion" decisions is not surprising, especially in crimes against people where there are significantly more individuals against which to compare.

Number of Suspects, Victims, and Effectiveness of AFIS

Table 9 shows the number of suspect names provided in each case by the submitting officer in the 2009/2010 data set. It is not surprising that weapons and drugs cases almost always (86% and 96% of the time, respectively) have at least one suspect named. It is also not surprising that these cases commonly have multiple suspects named. Often these cases are requested for latent print analysis when a raid or search of a dwelling or vehicle is performed by law enforcement. When they recover the contraband in the dwelling or vehicle, the parties deny knowledge or ownership of the items. Latent prints are usually requested for the government to prove "ownership" or

TABLE 9 Distribution of the Number of Suspects Provided Per Case Type Category

	Number of cases with no suspect provided (% case type total)	Number of cases with 1 suspect provided (% case type total)	Number of cases with 2 to 5 suspects provided (% case type total)	Number of cases with 6 or more suspects provided (% case type total)	Total number of cases
Crimes against people	28 (22%)	63 (49%)	37 (29%)	1 (<1%)	129
Property crimes	380 (58%)	158 (24%)	113 (17%)	4 (1%)	655
Weapons	3 (14%)	6 (27%)	13 (59%)	0	22
Drugs	3 (4%)	39 (49%)	37 (47%)	0	79
Totals	414 (47%)	266 (30%)	200 (23%)	5 (<1%)	885

at least "knowledge of" through contact established by a latent print identification. In crimes against people, it can be seen that most crimes against people (78%) have at least one suspect named in the case. This may be because the nature of these crimes requires contact between two people. The victims may know the perpetrator or perhaps there is a more intense investigation in these cases because of the severity of these crimes. Just over half (58%) of the property crimes submitted to BCA do not have a suspect named. These crimes are often committed when there are no victims present or witnesses to the crime. These cases will require AFIS searches to generate potential suspects.

AFIS was used in 323 out of the 885 reviewed cases (36%). In the BCA-LPU, AFIS is typically utilized for any unidentified latent prints in a case, but only after they have been compared and possibly identified to the victim/elimination prints or a suspect proffered by the case investigator. Furthermore, the unidentified latent prints must be suitable for an AFIS search. Certain types of latent prints (e.g. finger joints, extreme fingertips, etc.) may be identifiable, but not appropriate for a search in AFIS because these areas of the friction ridge skin are not recorded during a standard booking in Minnesota. In the 323 cases where AFIS was utilized, 99 cases generated new suspects. Eighty-two of the 99 cases (83%) where a new suspect was developed were property crimes; 11 cases (11%) were crimes against people, and 6 cases (6%) were drug cases. No new suspects were developed with AFIS in weapons cases.

In the 323 cases where AFIS was used, a total of 658 latent prints were searched in AFIS. This averages to 2 latent prints per case that were entered into AFIS (median = 1). Eleven cases had AFIS entry of 6 or more latent prints; one homicide case had 56 entries and generated 7 new suspects. The AFIS searches led to the development of 111 new suspects based on identifications made from AFIS resulting in an AFIS hit rate of 17%. This also means that AFIS provided a new suspect in approximately 1 in 3 cases (99 of 323 cases) where it was used, and that BCA generated new suspects using AFIS in approximately 1 in 10 of all cases submitted to BCA (99 of 885 cases).

In Table 10, it was reported that there were 544 "identification" decisions in the 2009/2010 data set. These 544 "identification" decisions are sorted into the number of "identification" decisions reported to a suspect in the case versus a victim/elimination source. It should be noted that in drugs and weapons cases there is rarely a "victim" listed, and officer elimination prints are rarely submitted. The proportion of "identification" decisions to suspect versus victim is nearly equal in crimes against people. In property crimes, a suspect was three times more likely to be identified than a victim/elimination source. This is likely due to the reasons as discussed previously: there tends to be contact between the perpetrator and victim in crimes against people, whereas in property crimes the victim(s) are not present during the commission of the crime.

TABLE 10 Distribution of "Identification" Decisions Attributed to Suspects or Victims/Elimination Sources for the 2009/2010 Data Set

	Number of "identification" decisions (N = 544)	
	Suspects	Victim/Elimination
All cases	396 (73%)	148 (27%)
By case type:		
Crimes against people	70	56
Property crimes	278	92
Drugs	47	0
Weapons	1	0

However, it seems plausible that in property crimes, since these are typically burglary or auto theft cases at BCA, we could be equally (or more) likely to find victim prints on surfaces that the victim routinely touches. Since this was not the case in the 2009/2010 data set, does it have something to do with smart choices made at a crime scene? Is this because information exchanged between the victim and the investigator leads to better choices of the most relevant evidence?

Contextual Information and Interaction with Investigators

Table 11 shows the distribution of cases where the level of interaction between the case analyst and the submitting officer(s) or prosecutor was categorized as "high," "moderate," or "none/minimal" based on criteria previously discussed in *Materials and Methods*. Table 11 also shows the distribution of cases where the level of context information available to the case analyst was categorized as "high," "moderate," or "none/minimal" based on the previously discussed criteria.

TABLE 11 **Distribution of Case Type, Level of Contextual Case Information Supplied to the Analyst, and Level of Interaction Between the Analyst and Investigators**

Level of Interaction	High				Moderate				None/minimal			
Amount of Context Information	High	Mod	None-Minimal	Total	High	Mod	None-Minimal	Total	High	Mod	None-Minimal	Total
Crimes against people	13	0	4	17	10	7	7	24	34	16	38	88
Property Crimes	4	2	8	14	14	11	10	35	141	95	370	606
Drugs	1	1	7	9	0	2	10	12	4	7	47	58
Weapons	0	1	2	3	1	0	0	1	6	1	11	18
Totals for "Level of Interaction"				43				72				770
Totals for "Amount of Context Information"		Number of Cases with High Context Information		228		Number of Cases with Moderate Context Information		143		Number of Cases with No/minimal Context Information		514

Table 11 shows that in 87% of the cases (770 out of 885), there was minimal interaction between the analyst and the investigation. In these cases, evidence was received with a request to process, the analyst performed the examinations, and issued a report, with no communications between the requester and the analyst. It should be recognized, that other agencies may have a routinely different level of interaction with investigators. A smaller police department may have investigators directly handing evidence and interacting with analysts, or possibly the crime scene investigator *is also* the latent print analyst in the case. These data show that most examinations are routine tests with minimal or no interaction between the BCA analysts and investigators.

Table 11 also shows that 58% of cases submitted at BCA have no context/case information provided (514 out of 885), while 26% have a high amount of context/case information provided (228 out of 885) and 16% have a moderate amount of context/case information provided (143 out of 885). Further analysis showed that it was predominantly smaller/rural agencies which were providing more context information and longer, more detailed reports (38% of the time from rural agencies versus 24% from large metropolitan cities).

Two subsets of those data were compared: the cases where there was high context information and high interaction (high context/high interaction; N = 18) versus the cases where there was no context information and no interaction (no context/no interaction; N = 466). The reason for doing so is that it has been asserted that the high context/high interaction cases are essentially where there is the most danger of bias—that the analyst is receiving significant non-domain information and cues from investigators. This is actually a very limited number of cases in the sample (2%). This is in contrast to the 53% of cases with no context information and interaction with investigators.

It can be seen in Table 12 that when comparing no context/no interaction cases against high context/high interaction cases, the most obvious difference is that the high context/high interaction cases produced a disproportionately larger number of "identifiable" latent prints (an average of 6.7 per case versus 1.4 in cases of no context/no interaction). This is explainable given that many of the high context/high interaction are disproportionately crimes against people (and specifically 11 out of 13 are homicides). Homicides, as previously discussed, tend to generate significantly more evidence, and have the highest level of context information and interaction between investigators, prosecutors, and analysts.

Only 25 of the 121 identifiable latent prints resulted in an "identification" decision (a 21% identification rate) in the high context/high interaction cases. It is striking to note that in the no context/no interaction cases, 142 of the 650 identifiable latent prints resulted

TABLE 12 Distribution of "Identification" and "Exclusion" Decisions Sorted by Case Type, Level of Contextual Case Information Supplied to the Analyst, and Level of Interaction Between the Analyst and Investigators

	N	Number of identifiable latent prints	Number of latent prints identified to suspect	Number of exclusion decisions to suspect(s)
No context -No interaction				
All case types	466	650	142	172
Crimes against people	38	44	8	4
Property crimes	370	579	128	151
Drugs	47	22	6	12
Weapons	11	5	0	5
High context – High interaction				
All case types	18	121	25	334
Crimes against people	13	106	20	301
Property crimes	4	7	0	26
Drugs	1	8	5	7
Weapons	0	0	0	0

in an "identification" decision (a 22% identification rate). Essentially there was no difference in the rate of identification between these two subgroups. This is not compelling evidence that analysts are highly motivated to find only evidence to support the police theory and are being influenced by interactions with police and prosecutors (Koppl and Sacks 2013). This is not to say that it could not have happened in any one of these cases, but rather, there is no compelling evidence of such a trend or routine practice.

There was a difference in the rates of exclusions to suspects: there were nearly 3 "exclusion" decisions per latent print for high context/high interaction cases, whereas there was only 1 "exclusion" decision for every 4 latent prints in no context/no interaction cases. Proportionately, there were 12 times as many exclusions of suspects in high context/high interaction cases as there were in no context/no interaction cases. This is likely due to the higher number of suspects against which to compare in homicide cases in the high context/high interaction cases compared to the large number of property crimes, where there is usually no suspect provided about half the time, dominating the no context/no interaction cases.

Another way to look at the above data is that if an agency *was* to decide to shield an analyst from all context information and interaction with the investigators it would not necessarily have a deleterious effect on the number of "identification" decisions. A fair question however is whether the necessary sequential unmasking steps are worth the effort. At an agency like BCA, it would certainly require hiring additional technical staff and changing workflow procedures, writing computer code and creating permissions on who has access to information and how it will be disseminated. If this were done for all sections at the BCA, it would feasibly require at least 5 technically trained staff to manage case information and coordinate cases among bench analysts. This is likely a minimum salary cost (not including benefits) of $250,000 per year. Given current backlogs and a need for faster turn-around time, is this really the highest priority? Those that call for sequential unmasking procedures in all cases have not offered a realistic analysis on the impact on work flow and cost to implement full blinding procedures (Kassin, Dror, and Kukucka 2013). More importantly, no pilot studies have been published showing that testing errors will be decreased with such procedures in place. Before widespread implementation of such procedures, the authors call for research demonstrating that in a complex, high through-put crime laboratory these procedures will have any serious reduction of error. A cost-benefit analysis, with actual data from those who understand the workings of a crime lab, has yet to be offered (5; Kassin, Dror, and Kukucka 2013). Perhaps this money might be better spent on the back-end, limiting which evidence is presented in court and how it may be presented to a jury (for example, using "hot-tubbing" approaches) (Champod and Vuille 2010). Or perhaps this money could be used for expert fees and independent testing to review cases for defense, when there is a dispute of the crime lab's findings. In this vein, Saks,

et al. proposed a *forensic voucher* system (Saks et al. 2001). These *select* cases could then be subjected to a sequential unmasking procedure during an independent review, rather than subject all cases *a priori* to such a labor intensive approach.

Single Latent Print Associations and the Potential for Bias Effects

Given the attention the Brandon Mayfield case has been given, the authors felt it important to investigate the realistic possibilities of the frequency of cases in the 2009/2010 data set that could have "Mayfield-case-like" factors. In the Mayfield case, latent prints recovered from evidence at the scene of a commuter train bombing in Madrid, Spain, on March 11, 2004, were sent to federal agencies around the world. The FBI in the U.S. received these images and searched them in their AFIS database. A single, complex latent print, was erroneously identified to an American named Brandon Mayfield (Stacey 2004). It was the only physical evidence associating Mayfield to the case. The case analysts were exposed to significant contextual information and there were significant interactions between the fingerprint examiners and Spanish officers. However, these interactions between U.S. and Spanish officials and the exposure to extraneous contextual information came *after* the "identification" decision was declared, but the analysts continued to maintain the decision, even in the face of contradictory information. Nonetheless, this case is treated as a poster child for high bias and context effects (National Research Council 2009).

The authors explored how many of the cases with "identification" decisions in the 2009/2010 data set reported a single "identification" decision to a suspect in the case. To be clear, there could have been multiple identifiable latent prints and multiple suspects proffered, but only one of the latent prints in the case was identified to a suspect. Thus the latent print evidence in the case is a single link. This choice has been made because in all of the reported cases of erroneous identifications, it has always been a single erroneous identification decision to an individual. The prevailing theory is that these errors are relatively rare events. The likelihood of one erroneous identification decision being made to a single suspect, and then verified by a second examiner is estimated to be exceedingly low—much less than 0.1% (Ulery et al. 2012). The chance of it happening twice to the same individual with two different latent prints would be significantly smaller. It is the primary reason that SWGFAST and the FBI have both chosen to focus their attention during "blind verification" on single conclusion decisions (Cole 2013, Scientific Working Group on Friction Ridge Analysis Study and Technology [SWGFAST] 2012).

In the 2009/2010 data set, 89 of the 396 "identification" decisions to suspects (see Table 10) were single "identification" decisions to a suspect. When we further examined the assignment of the level of context information and interaction in these 89 cases, we found that only 1 of these cases was "high context/high interaction." In the 885 cases reviewed, a single case had a single identification to a suspect with the analyst being exposed to high level of context information and having a high level of interaction with investigators. It was a homicide case. Figure 7 shows the latent print in this case. Forty-two (42) of the 89 "single ID cases" (47%) had no context information/interaction. The remaining 46 "single ID cases" (52%) had some combination of context information and interaction other than "high/high" or "none/none."

The latent print in Figure 7 shows a relatively non-complex latent palm print. The latent print has a large amount of clear ridge detail, with intermittent areas of distortion. The latent print is in blood and was processed with a dye stain; it exhibits areas of classic blood

FIGURE 7 **The latent print from the only case in the 2009/2010 data set with a single "identification" decision to a suspect and where there is high context/high interaction.**

matrix distortion effects (Langenburg 2008). The authors provided this blood print to five latent print experts, certified by the International Association for Identification (IAI) and asked them to rate the difficulty. All five experts indicated the blood print to be "easy" for comparisons purposes.

It is intriguing that only one case for 12 months of randomly sampled case data met the conditions of "high context/high interaction/single identification to a suspect." Furthermore, the palm print examination in the case is "easy" from the perspective of a fingerprint expert. In fact, based on previous research at BCA (Neumann et al. 2011), the percentage of cases with difficult, marginal latent print examinations is relatively small (<5%). It would appear to be uncommon for a case to have high context, high interaction, a single identification to a suspect, *and also* be of marginal value or a difficult examination. Research to date has shown relatively little error from bias effects for experts and novices when the latent print comparisons are deemed "easy." Errors from bias effects were much more pronounced when the examinations were deemed "difficult" and/or dealt with "exclusion" decisions (Langenburg, Champod, and Wertheim 2009). Again, we make the point, is it necessary to blind all cases when such a small fraction pose any real risk of error? As a more resource friendly option, we could utilize sequential unmasking techniques on this small subset of cases and instances. These cases could be further vetted by identifying them as ideal for review by defense experts, who could then utilize a process of sequential unmasking when reviewing the conclusions of the laboratory.

CONCLUSIONS

The present study and accumulated data sets sampled from four different years at the BCA (2003, 2004, 2009, and 2010) revealed a number of interesting trends. The data are useful for managers to compare laboratory output. They are useful for researchers needing accurate estimates of latent print results from actual casework. They are useful for policy and decision makers to understand the impact that external factors can have on latent print results (e.g., an increase in DNA property crime cases, submission of known or potential suspects, etc.).

We have summarized the major trends observed in the present study as follows:

- From 2003/2004 to 2009/2010, there was an increase in the number (and proportion) of property crime case submissions for latent prints. We theorize this increase may be a "trickle down" effect from increased DNA submissions. Unfortunately for the latent print unit, the personnel and resources have not adjusted accordingly to the increase, thus contributing to the problems of a growing backlog and decreasing morale.
- Just over half of the cases submitted to the BCA-LPU revealed at least one identifiable latent print. Approximately 1 in 4 cases submitted to the BCA-LPU resulted in at least one "identification" result. Approximately 3 in 4 "identification" decisions were to a suspect in the case versus a victim/elimination source. However, approximately half the cases submitted to the BCA-LPU do not have a suspect named by investigators. These cases with no named suspects were predominantly (over 90% of the time) property crimes.
- AFIS was used in about 1 in 3 cases submitted to the BCA-LPU and was used predominantly in property crimes (as noted, due to the lack of provided suspects). A new suspect was generated in about 1 in 3 of the searched cases and had a latent print "hit" rate of 17%. Most cases where a search was required had 1 to 2 latent prints to search in AFIS.
- Approximately 1 in 7 latent prints appeared to originate from a palm (as opposed to a finger or finger joint). This demonstrates the need for palm print databases/exemplars and training on palm prints.
- While only about half of the cases submitted to BCA-LPU had any processing (powder and lift, cyanoacrylate fuming, etc.) done prior to submission, these cases nearly doubled the chance of finding an identifiable latent print. This is an important message to crime scene technicians weighing the risk/benefit of processing the exhibit in the field versus submitting the exhibit to a lab where it may take several months before being processed.
- Non-porous processes (cyanoacrylate fuming followed by dye stain or powder) were the most commonly employed process by the BCA-LPU.
- The recovery rate for identifiable latent prints from plastic bags (submitted mostly in drugs cases) was 13%. The recovery rate for identifiable latent prints from firearms was 13%. No identifiable latent prints were recovered from fired or unfired ammunition in the study.

- The rate of identifiable latent prints that were subsequently identified was approximately the same for property crimes, crimes against people, and drug cases (all around 40%).
- Most cases (87%) submitted to the BCA-LPU have no interaction between the analyst and the investigator in the case. Just over half (58%) of the cases submitted to the BCA-LPU have no case information/context information submitted other than the requested forms that include suspect/victim names, dates of birth, exhibits to be examined, etc. This resulted in half (53%) of the cases having no interaction/no context information, and only 2% of the cases having a high level of interaction/high level of context information exchanged between the forensic analyst and the police investigators.
- The rate of latent print "identification" decisions was the same for identifiable latent prints recovered in cases of no context/no interaction versus cases of high context/high interaction (21% and 22%, respectively). This is not compelling evidence of a trend that forensic analysts are being motivated and influenced by context information or interaction with law enforcement to produce more "identification" decisions to "aid the police."
- Approximately 10% of the BCA-LPU cases resulted in a single latent print "identification" decision to one of the suspects in the case. Half of these cases were classified as no context/no interaction, while only one case in the entire set had a single identification to a suspect in the case under the high context/high interaction condition.

From these findings we draw three conclusions. The first conclusion is that these data were valuable to the BCA-LPU in understanding the basic effectiveness and rate of success for current processes. It is less effective to go to management and say "we need palm print training because we see a lot of palm prints in our cases," versus "palm prints are an integral part of my duties—1 in 7 of the latent prints I examine are palm prints; without training or a database to search, I am not utilizing a large portion of my evidence." Managers and policy makers tend to react to data and dollars versus vague assertions. The data also give the BCA-LPU a baseline performance statistic, so that if we make changes to policy or processes, we will have data against which to compare the effectiveness of the change.

The second conclusion is that, unlike our fictional *CSI* counterparts on television, most case submissions are actually unsuccessful. In half the cases submitted we find no identifiable latent prints, and in the half that we do, only half of ***those*** cases result in an "identification" decision reported by the analyst. Of ***those*** "identification" decisions, three-fourths of the time they are to a suspect in the case. So in effect, only about 1 in 6 cases submitted to the BCA-LPU are returned with what is likely to be a "helpful" result to law enforcement (i.e., a suspect was identified with latent print evidence).

We are hopeful that data in the present paper, and some other similar papers, can be presented by other individuals during testimony, and thus not require an analyst to testify to why latent prints were not found in this case and the absence of identifiable latent prints is common. Especially for a state or federal agency, the travel time and costs can be a resource drain. When waiting time, delays, and continuances are factored in, this can be a serious waste of analyst time and tax dollars. Data such as that reported in this paper can be relayed by the local crime scene technician or an investigator (provided they have some basic forensic experience), thus precluding the need for a lab analyst to appear and give testimony.

The last conclusion is that there was little evidence of a trend for forensic analysts at the BCA-LPU to be biased toward aiding law enforcement from interactions or information exchanged between the fingerprint examiner and police investigators. The fact that the rate of "identification" decisions was identical in the subsets of no context/no interaction and high context/high interaction does not show a tendency for the analyst in those high context/high interaction cases (which tended to be crimes against people, and specifically homicide cases) to push the envelope and either claim more identifiable latent prints or claim more "identification" decisions.

This does not mean that we dispute the inherent dangers of error from bias, nor do we ignore the research that has demonstrated bias effects. We believe that there is usefulness in sequential unmasking or blind verification procedures, but to date, there are no studies that demonstrate such procedures applied to all cases will in effect reduce the number of errors in casework or be cost effective and worth the resources dedicated to instituting a masked workflow. In Langenburg 2012, it was proposed that instituting blind verification

in all cases would lead to more erroneous *exclusions*, and these errors would become a constant drain on resources by constantly performing quality reviews and dealing with corrective action issues.

Based on the data in this study, and still recognizing the obvious concern of error from bias effects, it makes the most sense from a resource standpoint to recommend that if a sequential unmasking approach is to be instituted, then it should be used in the small subset of cases where the effect of bias is most likely to have an impact. For the BCA-LPU, this would represent from as many as 10% of the submitted cases (for *all* cases where there is a single "identification" decision") to as low as 1% (for cases where there is a single *complex* "identification" to a *suspect*). In this way, a much more resource friendly approach could be adopted. BCA-LPU currently has a standard operating procedure (SOP) for "Blind Verification," and this standard captures the essence of the similarly titled SWGFAST standard. The BCA-LPU "Blind Verification" SOP is applied currently exactly as described above, judiciously, when the perceived risk or benefit is sufficient to justify its use.

Limitations and Further Research

The sampled cases were a cross-section of cases for the specified years 2003–2004 and 2009–2010. Personnel, policies, and procedures have changed significantly in the last decade or so in the BCA-LPU. Those cases sampled represented the attitudes and procedures of the day. The cases selected were representative samples of a specific time window in the BCA-LPU. Each case, however, has its own unique set of circumstances, and so while we looked at overall trends in the present study, this does not mean to imply that in one singular case an analyst could have done something different, or been influenced by context information, or made an error, etc. The focus was on general and distinctive trends.

Another limitation is how the cases were categorized and assessed. For example if a case was a burglary where the perpetrator left behind a note bearing racial epithets and threats, is this a property crime or a crime against people? If it was clear, we used the higher potential criminal charge in the case, but often, this information may not be available upon submission so the case is classified as best as possible. When assigning the level of interaction between analyst and investigators or the level of contextual information available in the case, we again, had to make some judgment calls. Typically, we opted for a higher level of interaction/context information if there was any doubt. For example, if the case only had a short note such as "we are looking for subject's prints on the gun," this was designated as a "minimal to no context" case. If the officer wrote (and this would be extraordinary and did not occur in these samples) "we are looking for subject's prints on the gun—we know he did it and he's a bad person who needs to come off the streets," then this would be categorized as a "high context" case even though it is a single, short statement made to the laboratory. While length was a consideration, content of the information was also considered as well. This is where reasonable judgment was exercised, but was subjective nonetheless.

With respect to the effects of context information and interactions with law enforcement, we only compared the two conditions of high context/high interaction versus no context/no interaction. There may be bias effects present in cases that have some combination of context/interaction other than "high context" *and* "high interaction." Given the attention that has been placed on high context and high interaction with law enforcement, this seemed to be a reasonable starting point. Other combinations, can, and should, be explored. Furthermore, we don't know if bias effects from context information are weaker or stronger influences than those influences from interaction with police investigators. Perhaps a "moderate context" case may produce bias effects equivalent to a "high interaction" case. We simply do not know enough about the frequency and impact of bias effects leading to error in forensic casework.

Finally, the data obviously represent the casework, policies, procedures, and personnel of the BCA-LPU. These data may not be representative for agencies of a different size or with different workflows. For some agencies, what constitutes 'a case' may differ dramatically than BCA-LPU. How AFIS searches/cases, technical reviews, cases are documented, exhibits processed, etc. will significantly affect the counts. It is important to carefully consider the BCA-LPU policies, demographics, and workflow that is described in the introduction of this paper before making comparisons to another agency.

The authors envision a few follow-up studies from this. Firstly, one of the authors works in Geneva,

Switzerland. It would be interesting to make some comparisons between the two laboratories. The perception is that the U.S. has so much more crime (specifically violent crime) compared to European countries. It would be interesting to compare data sets between the two laboratories. Secondly, during the study, we identified a list of cases that bore single "identification" and "exclusion" decisions. We could measure the repeatability and reproducibility of those decisions under different context. A set-up similar to Dror, et al (2006) could be employed. Furthermore, we could assign a level of difficulty in advance for each of the decisions and use minutiae counts to predict outcomes based on recent studies (Ulery et al. 2013, Neumann et al. 2013). Lastly, we are interested in more recovery rate data. The 2003/2004 dataset was ten years old and we would prefer to re-examine recovery rates under newer policies and procedures, including the use of Indanedione for porous exhibit processing and the use of digital capture and enhancement.

ACKNOWLEDGMENTS

The authors wish to thank Brenda Hummel, Hamline University. She was the intern who kindly extracted the data from the case information and LIMS. We are thankful for and appreciate the collaborative spirit of the Forensic Laboratory of the Canton Police Geneva, Switzerland to loan the author, Flore Bochet, for a time to perform this research. We wish to also thank the BCA management and BCA-LPU latent print examiners for their willingness to share these data for the benefit of the community. Thank you to the anonymous reviewers and their helpful comments.

REFERENCES

Barnum, C. A. and D. R. Klasey. 1997. Factors Affecting the recovery of latent prints on firearms. *J. Forensic Identification* 47(2): 141–147.

Champod, C. and J. Vuille. 2010. Preuve scientifique en Europe – Admissibilité, appréciation et égalité des armes. Strasbourg, Germany: Conseil de l'Europe - Bureau du Comité européen pour les problèmes criminels (CDPC).

Charlton, D., P. A. F. Fraser-Mackenzie, and I. E. Dror. 2010. Emotional experiences and motivating factors associated with fingerprint analysis. *J. Forensic Sci.* 55(2): 385–393.

Cole, S. A. 2006. The prevalence and potential causes of wrongful conviction by fingerprint evidence. *Golden Gate Univ. Law Rev.* 37(1): 39–105.

Cole, S. A. 2013. Implementing counter-measures against confirmation bias in forensic science. *J. Appl. Res. Memory Cognit.* 2(1): 61–62.

Cook, R., I. W. Evett, G. Jackson, P. J. Jones, and J. A. Lambert. 1998. A model for case assessment and interpretation. *Sci. Justice* 38(3): 151–156.

Cook, R., I. W. Evett, G. Jackson, P. J. Jones, and J. A. Lambert. 1999. Case pre-assessment and review in a two-way transfer case. *Sci. Justice* 39(2): 103–111.

Dhingsa, R., A. Qayyum, F. V. Coakley, Y. Lu, K. D. Jones, M. G. Swanson, P. R. Carroll, H. Hricak, and J. Kurhanewicz. 2004. Prostate cancer localization with endorectal MR imaging and MR spectroscopic imaging: effect of clinical data on reader accuracy. *Radiology* 230(1): 215–220.

Dieltjes, P., R. Mieremet, S. Zuniga, T. Kraaijenbrink, J. Pijpe, and P. de Knijff. 2011. A sensitive method to extract DNA from biological traces present on ammunition for the purpose of genetic profiling. *Int. J. Legal Med.* 125(4): 597–602.

Dror, I. E. 2013. The ambition to be scientific: Human expert performance and objectivity. *Sci. Justice* 53(2):81–82.

Dror, I. E., and D. Charlton. 2006. Why experts make errors. *J. Forensic Identif.* 56(4): 600–616.

Dror, I. E., D. Charlton, and A. E. Péron. 2005. Contextual information renders fingerprint experts vulnerable to making erroneous identifications. *Appl. Cognit. Psychol.* 19(6): 799–809.

Dror, I. E., and G. Hampikian. 2011. Subjectivity and bias in forensic DNA mixture interpretation. *Sci. Justice* 51(4): 204–208.

Durose, M. R. 2008. *Census of publicly funded forensic crime laboratories, 2005*. Washington, DC: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Haber, Lyn, and R. N. Haber. 2008. Scientific validation of fingerprint evidence under Daubert. *Law, Probability and Risk* 7(2): *Sci. Justice* 87–109.

Horsman–Hall, K. M., Y. Orihuela, S. L. Karczynski, A. L. Davis, J. D. Ban, and S. A. Greenspoon. 2009. Development of STR profiles from firearms and fired cartridge cases. *Forensic Sci. Int. Genet.* 3(4): 242–250.

Johnson, S. 2010. Development of latent prints on firearms evidence. *J. Forensic Identif.* 60(2):148–151.

Kassin, S. M., I. E. Dror, and J. Kukucka. 2013. The forensic confirmation bias: Problems, perspectives, and proposed solutions. *J. Appl. Res. Memory Cognit.* 2(1): 42–52.

Kelty, S. F., R. Julian, and A. Ross. 2013. Dismantling the justice silos: Avoiding the pitfalls and reaping the benefits of information-sharing between forensic science, medicine and law. *Forensic Sci. Int.* 230(1–3): 8–15.

Koppl, R., and M. Sacks. 2013. The criminal justice system creates incentives for false convictions. *Criminal Justice Ethics* 32(2): 126–162.

Koppl, R. 2005. How to improve forensic science. *Eur. J. Law Econ.* 20(3): 255–286.

Krane, D. E., S. Ford, J. R. Gilder, K. Inman, A. Jamieson, R. Koppl, I. L. Kornfield, D. M. Risinger, N. Rudin, M. S. Taylor, and W. C. Thompson. 2008. Sequential unmasking: A Means of minimizing observer effects in forensic dna interpretation. *J. Forensic Sci.* 53(4): 1006–1007.

Langenburg, G. 2008. Deposition of bloody friction ridge impressions. *J. Forensic Identif.* 58(3): 355–389.

Langenburg, G. 2012. *A critical analysis and review of the ACE-V process*. Lausanne, France: Ecole des Sciences Criminelles (ESC)-Institut de Police Scientifique (IPS), University of Lausanne.

Langenburg, G., and C. Champod. 2011. The GYRO System: A recommended approach to more transparent documentation. *J. Forensic Identif.* 61(4): 373–384.

Langenburg, G., C. Champod, and P. Wertheim. 2009. Testing for Potential contextual bias effects during the verification stage of the ACE-V methodology when conducting fingerprint comparisons. *J. Forensic Sci.* 54(3): 571–582.

Loy, C.T., and L. Irwig. 2004. Accuracy of diagnostic tests read with and without clinical information: A systematic review. *J. Am. Med. Assoc.* 292(13): 1602–1609.

Maldonado, B. 2012. Study on developoing latent fingerprints on firearm evidence. *J. Forensic Identif.* 62(5): 425–429.

Margot, P. 2011. Forensic science on trial–What is the law of the land? *Aust. J. Forensic Sci.* 43(2–3): 89–103.

Mnookin, J. L. 2010. The courts, the NAS, and the future of forensic science. *Brooklyn Law Rev.* 75(4):1209–1276.

National Research Council. 2009. *Strengthening forensic science in the United States: A path forward*. Washington, D.C.: The National Academies Press.

Neumann, C., C. Champod, M. Yoo, T. Genessay, and G. Langenburg. 2013. *Improving the understanding and the reliability of the concept of "sufficiency" in friction ridge examination*. Washington, DC: U.S. Department of Justice.

Neumann, C., I. Mateos-Garcia, G. Langenburg, M. Schwartz, M. Koolen, and J. Kostroski. 2011. Operational benefits and challenges of the use of fingerprint statistical models: A field study. *Forensic Sci. Int.* 212(1–3):32–46.

Office of the Inspector General (OIG). 2006. *A review of the FBI's handling of the Brandon Mayfield case*. Washington, DC: U.S. Department of Justice.

Peterson, J. L., M. J. Hickman, K. J. Strom, and D. J. Johnson. 2013. Effect of forensic evidence on criminal justice case processing. *J. Forensic Sci.* 58(Suppl 1): S78–90.

Peterson, J. L., and P. Markham. 1995. Crime laboratory proficiency testing results, 1978–1991, II: Resolving questions of common origin. *J. Forensic Sci.* 40(6): 1009–1029.

Potchen, E. J., T. G. Cooper, A. E. Sierra, G. R. Aben, M. J. Potchen, M. G. Potter, and J. E. Siebert. 2000. Measuring performance in chest radiography. *Radiology* 217:456–459.

Potchen, E. J., J. W. Gard, P. Lazar, P. Lahaie, and M. Andary. 1979. The effect of clinical history data on chest film interpretation: direction or distraction. *Invest. Radiol.* 14:404.

Pratt, A. 2012. Fingerprints and firearms. *J. Forensic Identi.* 62(3): 234–242.

Roberts, P., and C. Willmore. 1993. *The role of forensic evidence in criminal proceedings, Royal Commission on Criminal Justice Research Study No. 11*. London, UK: HMSO.

Saks, M., L. Constantine, M. Dolezal, J. Garcia, G. Horton, T. Leavell, M. Levin, J. Muntz, R. Pastor, L. Rivera, J. Stewart, F. Strumpf, C. Titus, and H. VanderHaar. 2001. Model prevention and remedy of Erroneous Convictions Act. *Arizona State Law J.* 33:665–718.

Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST). *Document #14 Standard for the application of blind verification of friction ridge examinations* (11/14/12 ver. 2.0) 2012. http://www.swgfast.org/Documents.html, accessed June 16, 2014.

Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST). *Document #10 Standards for examining friction ridge impressions and resulting conclusions* (04/27/13 ver. 2.0) 2013. http://www.swgfast.org/Documents.html, accessed June 16, 2014.

Stacey, R. B. 2004. A report on the erroneous fingerprint individualization in the Madrid train bombing case. *J. Forensic Identi.* 54(6): 706–718.

Thornton, J. I. 2010. Letter to the editor—-a rejection of "working blind" as a cure for contextual bias. *J. Forensic Sci.* 55(6):1663.

Ulery, B. T., R. A. Hicklin, J. Buscaglia, and M. A. Roberts. 2012. Repeatability and reproducibility of decisions by latent fingerprint examiners. *PLoS One* 7(3): e32800. doi: 10.1371/journal.pone.0032800.

Ulery, B. T., R. A. Hicklin, G. I. Kiebuzinski, M. A. Roberts, and J. Buscaglia. 2013. Understanding the sufficiency of information for latent fingerprint value determinations. *Forensic Sci. Int.* 230(1–3): 99–106.

U.S. Census Bureau. *Annual estimates of the population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2012*. Population Division 2012. http://www.census.gov/popest/data/national/totals/2012/index.html (accessed July 25, 2013).

Wells, Gary L., M. M. Wilford, and L. Smalarz. 2013. Forensic science testing: The forensic filler-control method for controlling contextual bias, estimating error rates, and calibrating analysts' reports. *J. Appl. Res. Memory Cog.* 2(1): 53–55.

WILLIAM C. THOMPSON

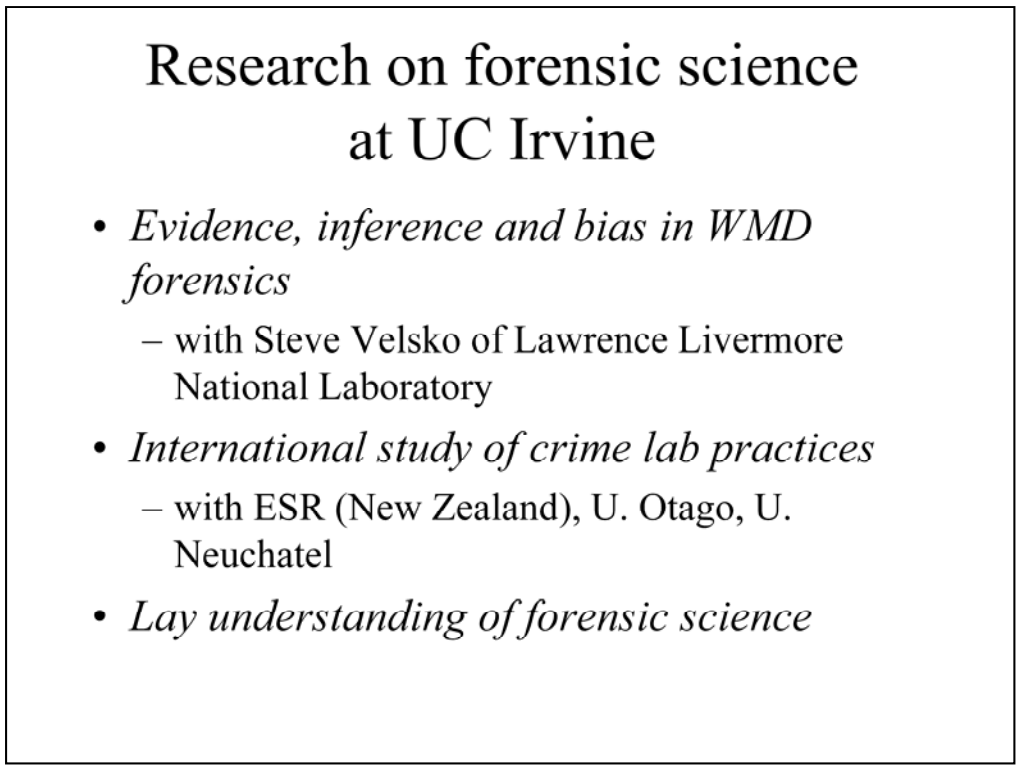
William C. Thompson is a professor in the Department of Criminology, Law & Society at the University of California, Irvine (UCI); he has joint appointments in Psychology and in UCI's School of Law, where he has taught Evidence. He received a Ph.D. in Psychology from Stanford University and a J.D. from the University of California, Berkeley. He has published extensively on the use and misuse of scientific and statistical evidence in the courtroom and on jurors' reactions to such evidence, focusing particularly on forensic DNA analysis. Although primarily an academic, Thompson occasionally practices law. He has litigated *Frye* and *Daubert* issues in trial and appellate courts and has represented clients in jury trials involving novel scientific and statistical issues.

With Steven Velsko of Lawrence Livermore National Laboratory, Thompson has led a three-year research project titled "Evidence, inference and bias in WMD forensics" that examines communication and inference in national security investigations, including the efforts of forensic experts to manage contextual information in order to avoid bias. Thompson will also lead an NIJ-funded research project on methods for context management in forensic laboratories.

Thompson was a member of the Task Force that drafted the ABA's Standards on DNA Evidence, he served on the California Crime Laboratory Review Task Force, and he has been a member of SWG-Speaker—the scientific working group on speaker identification. Thompson is currently vice-Chair of the Human Factors Committee that will provide guidance to the Organization of Scientific Advisory Committees (OSAC) on ways to improve human performance, reduce errors, and minimize cognitive bias in forensic science.

Minimizing Contextual Bias in Forensic Science

William C. Thompson
University of California, Irvine
National Commission on Forensic Science—August 26, 2014



I am an academic psychologist. I study human judgment and decision making and I have been particularly interested in the production and use of forensic science.

I am also a lawyer. I have litigated a number of cases involving contested forensic evidence.

My research group at UC Irvine is currently engaged in three lines of research on forensic science. First, we are collaborating with researchers from Lawrence Livermore National Lab on a study of problems of inference and bias in national security investigations involving forensic science, particularly those involving weapons of mass destruction. (The project is funded by the UC Lab Fees Research Fund). We are conducting interviews and reviewing historic episodes in order to trace the roots of investigative errors. Contextual bias is emerging as an important theme in this research. I think there is much to be learned from a comparison of how National Laboratories and crime laboratories view and address this issue.

Second, I am collaborating with researchers from several countries on an international study of how crime laboratories view and are addressing the issue of contextual bias. We are seeking NIJ funding for this research. We think a close examination of actual laboratory practices will help address questions about the practicality of various methods for addressing contextual bias.

Third, we have an active program of research that looks at how lay people (such as jurors) respond to forensic science evidence as a function of how it is presented and explained.

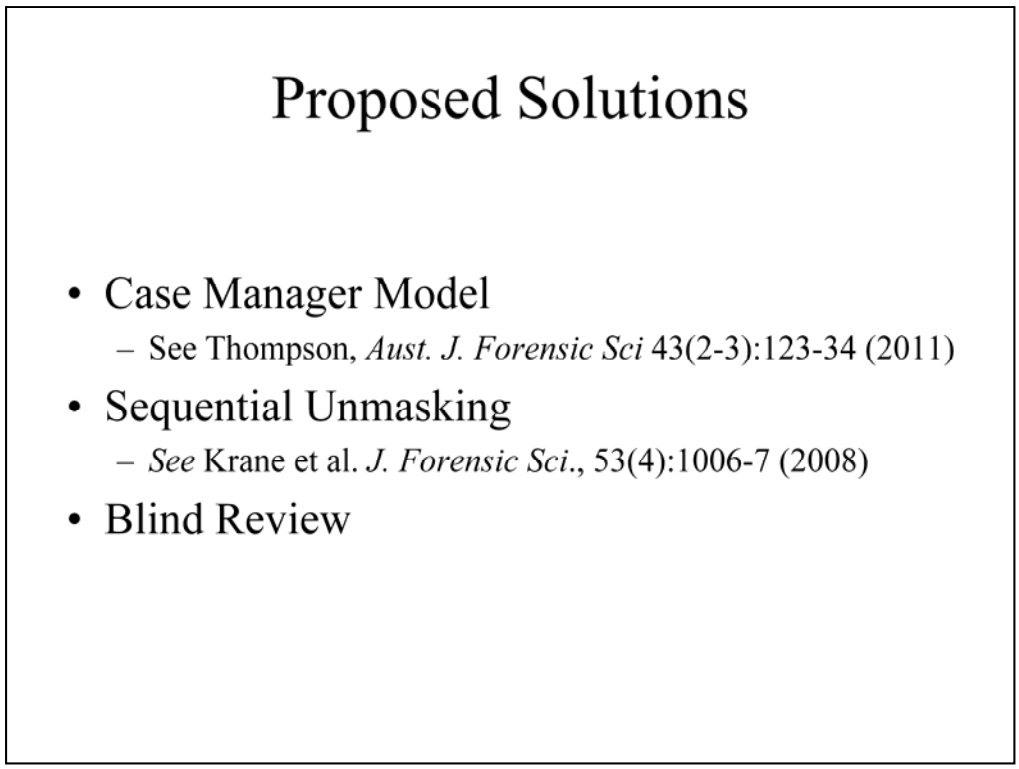
Contextual Bias

- *Bias* is said to occur when an analyst's judgment is influenced by *information irrelevant to the task*
- The influence may be:
  - Motivational—affecting *disposition or motive* to reach a particular result
  - Cognitive—affecting *interpretation and assessment* of data
- Most powerful when analysts rely on:
  - *Subjective judgment*
  - To interpret *potentially ambiguous data*

In order to talk about contextual bias, we need to discuss which aspects of the surrounding context a forensic scientist should and should not consider when making a forensic assessment. *Bias*, as I use that term here, arises when the forensic scientist is influenced by contextual information that *should not be considered* because it is irrelevant to the scientific task.

Bias can occur without conscious awareness and may arise from both motivational and cognitive mechanisms. It is a well-known human tendency to interpret data in a manner consistent with one's expectations and desires.

Contextual bias is less likely to be a factor when the data being examined are clear-cut or where standards exist that allow a single possible interpretation in each instance. It is more likely to be important when the data to be interpreted are potentially ambiguous or subject to more than one possible interpretation, and where analysts must rely more heavily on subjective judgment based on general knowledge, training and experience.



Case manager (a *trained forensic scientist*)

- Communicates with police
- Participates in decisions about collection, testing
- Manages work flow to Analyst

Analyst (*another* trained forensic scientist)

- Performs analytic tests and comparisons
- While blind to any information *unnecessary to the analysis*
- Prepares a written report

The same individual can perform both roles, but *not in the same case.*

Sequential Unmasking

See Krane et al. *J. Forensic Sci*., 53(4):1006-7 (2008), and subsequent commentary

Analysis/interpretations of evidentiary samples is performed *and documented*, as far as possible, before analyst is made aware of characteristics of reference samples

Information about reference samples is *unmasked* only when needed to complete the comparison

Blind Case Review

Critical judgments are replicated by a second analyst

Who is blind to unnecessary contextual information

Who has no expectations regarding outcome

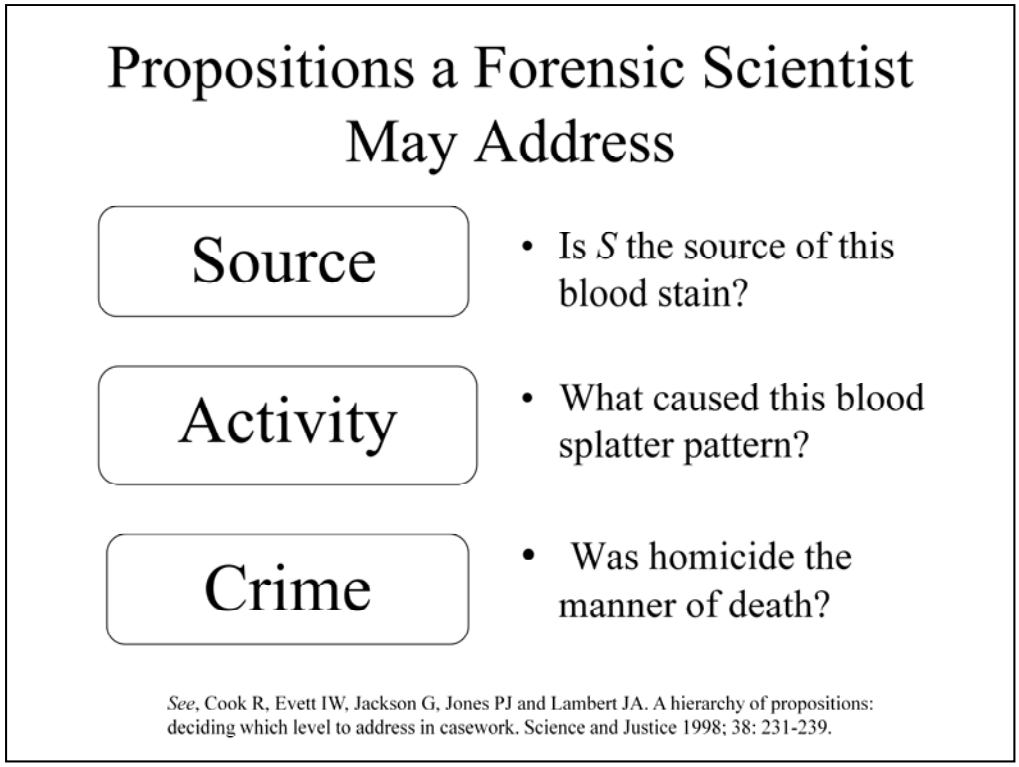
"I called this a match, what do you think Joe?" is probably not good enough

But how do we decide what is *task-relevant*?

Will efforts to shield analysts from potentially biasing contextual information deprive them of information they need to do their jobs?

Is it possible to draw a sharp analytic distinction between information that is *task-relevant* and *task-irrelevant* in forensic science?

If we cannot draw a sharp analytic distinction between task-relevant and task-irrelevant information, then efforts to reduce the influence of task-irrelevant information are likely to founder.



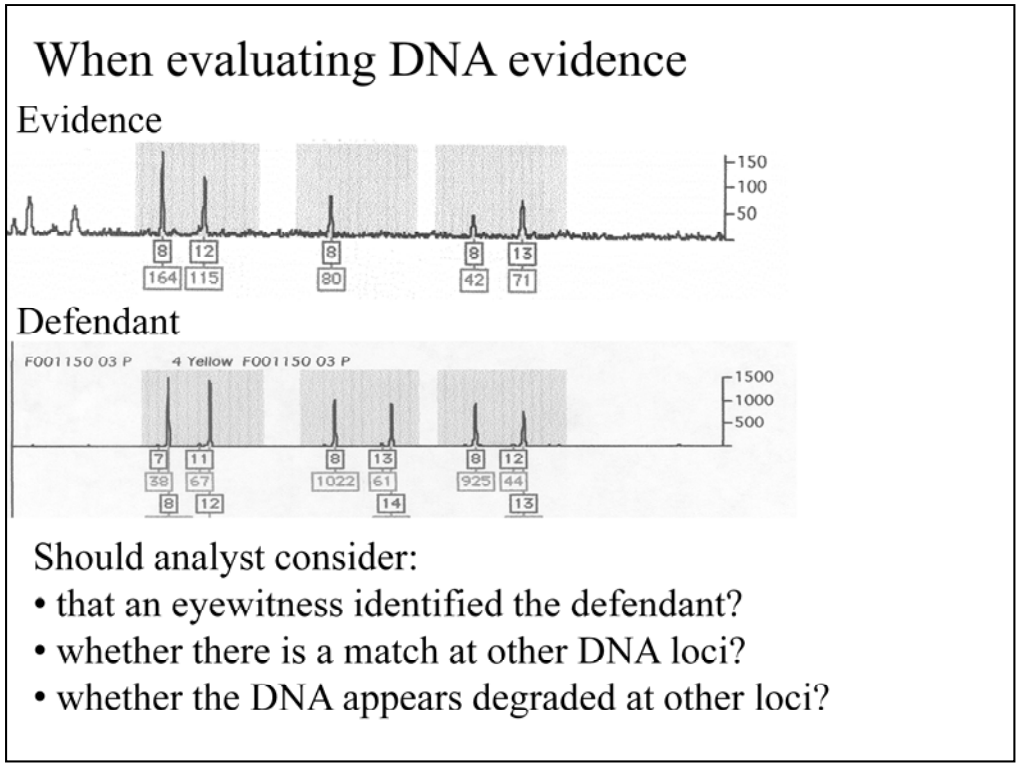
Most forensic scientists confine themselves to opining on source level propositions. The issue of whether a crime occurred, and what crime it was, is a matter for the legal system (judge or jury) rather than a forensic scientist. An exception is the medical examiner who is sometimes asked to make an independent determination of both cause and manner of death.

Criterion for Task-Relevance

Information is *task-relevant* if (and only if) it affects the conditional probability (under the relevant propositions) of the data the expert will evaluate.

Information that affects the probability that a relevant proposition is true, but not the conditional probability of the data under that proposition, is *task-irrelevant*.

I believe this definition of task-relevance is vitally important, but it is a bit technical an abstract. So I will explain it through some examples.



Here are DNA profiles from an evidentiary sample in a sexual assault case and from a criminal defendant. Could the defendant be the source of the evidentiary sample? Notice that one of the defendant's alleles was not detected in the evidentiary profile. Is this a true genetic difference (indicating the defendant was not the source)? Or did the discrepancy arise from "allelic dropout" (which can occur when the underlying DNA is degraded or insufficient in quantity)?

The analyst must make a subjective judgment based on data that are somewhat ambiguous (in that reasonable experts have differed in their interpretations).

A DNA analyst from a major laboratory recently told me that disagreements among analysts about issues of interpretation arise in about 10 percent of their cases (typically in cases involving mixed samples or samples with limited or degraded DNA). Thus, even with the best validated form of forensic science evidence, there can be ambiguities that analysts must resolve through the use of subjective judgment. This is the very situation in which we expect the effects of contextual bias to be most influential.

But what types of information are task-irrelevant and therefore potentially biasing? And which types constitute task-relevant information that the analyst may properly consider?

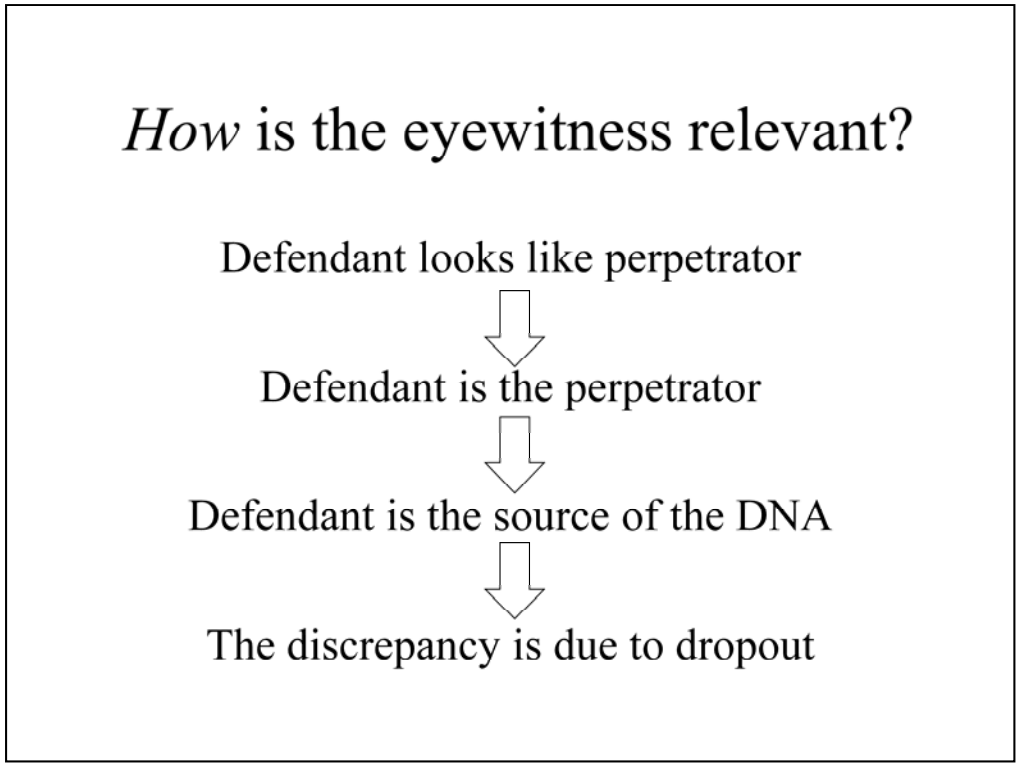
Rule 401 (Federal Rules of Evidence)

Evidence is relevant if:

(a) it has any tendency to make a fact more or less probable than it would be without the evidence; and

(b) the fact is of consequence in determining the action.

All of the information mention would be relevant to a juror under the Federal Rules. We must distinguish what is relevant for the jury from what is task-relevant for the analyst. One might think of this as distinguishing legal relevance from scientific relevance.

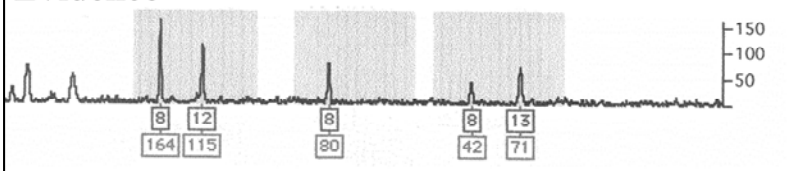


Consider the line of reasoning that links the eyewitness evidence to the assessment of the DNA evidence. Notice that it requires the DNA analyst to reason "backward" from an assessment of the defendant's guilt to an assessment of the DNA evidence. This kind of reasoning might well be reasonable for a juror who is trying to make sense of the entire case. But I will argue that it is entirely inappropriate for a forensic scientist who purports to perform an independent scientific assessment of the evidence.

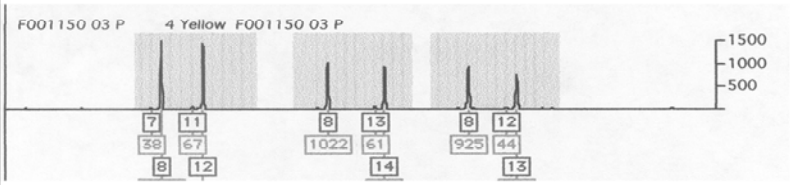
The forensic scientist is not in a good position to assess the other evidence in the case and has no business doing so. Moreover, the legal system expects that the forensic scientist's conclusions will stem from an assessment of the scientific evidence, not from consideration of other evidence in the case. The jurors may not realize that the expert is basing his or her conclusions in part on evidence the jury has already considered, which creates the potential for double-counting. More importantly, it allows the forensic assessment to be influenced (tainted) by other evidence, undermining its independence.

When evaluating consistency of DNA profiles

Evidence

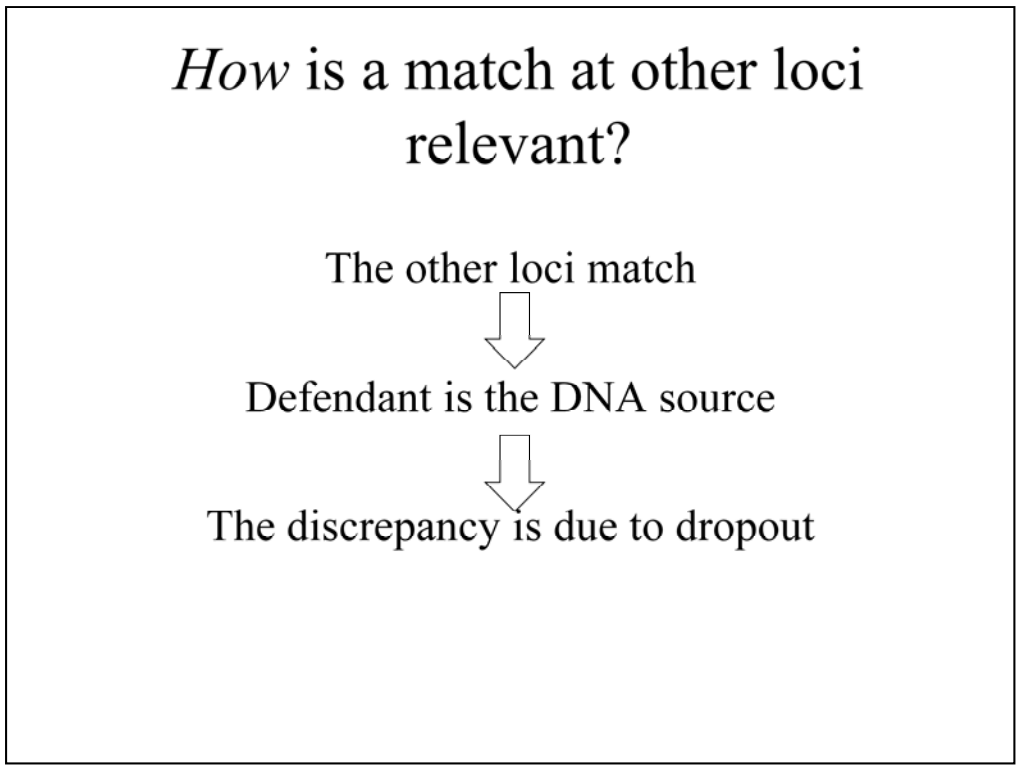


Defendant



Should analyst consider:

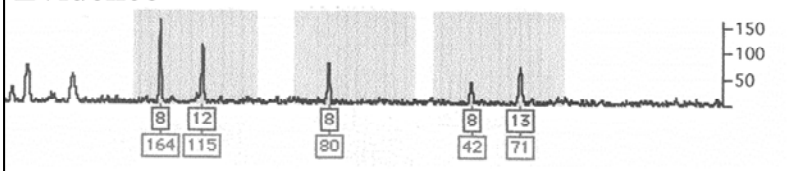
- that an eyewitness identified the defendant?
- whether there is a match at other DNA loci?
- whether the DNA appears degraded at other loci?



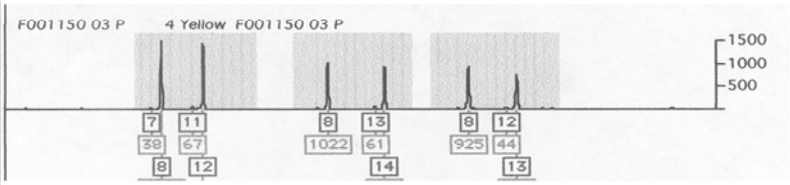
The same kind of backward reasoning is invoked when the analyst's judgments about the discrepancy between the profiles is influenced by whether the defendant matches at the other loci.

When evaluating consistency of DNA profiles

Evidence

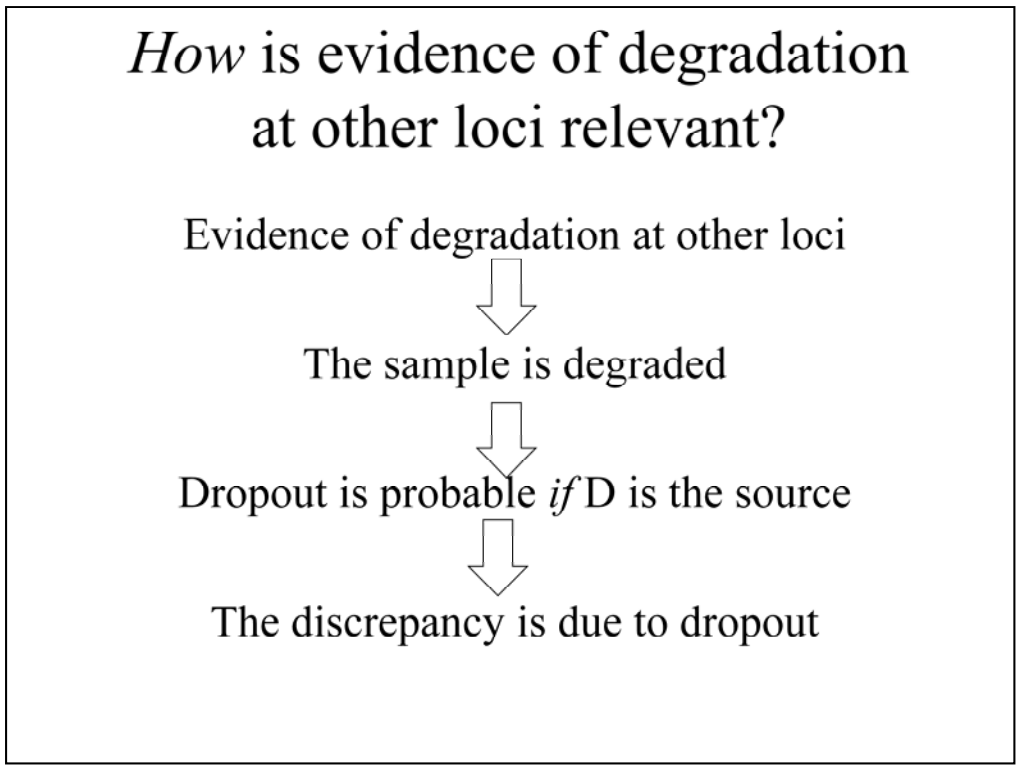


Defendant

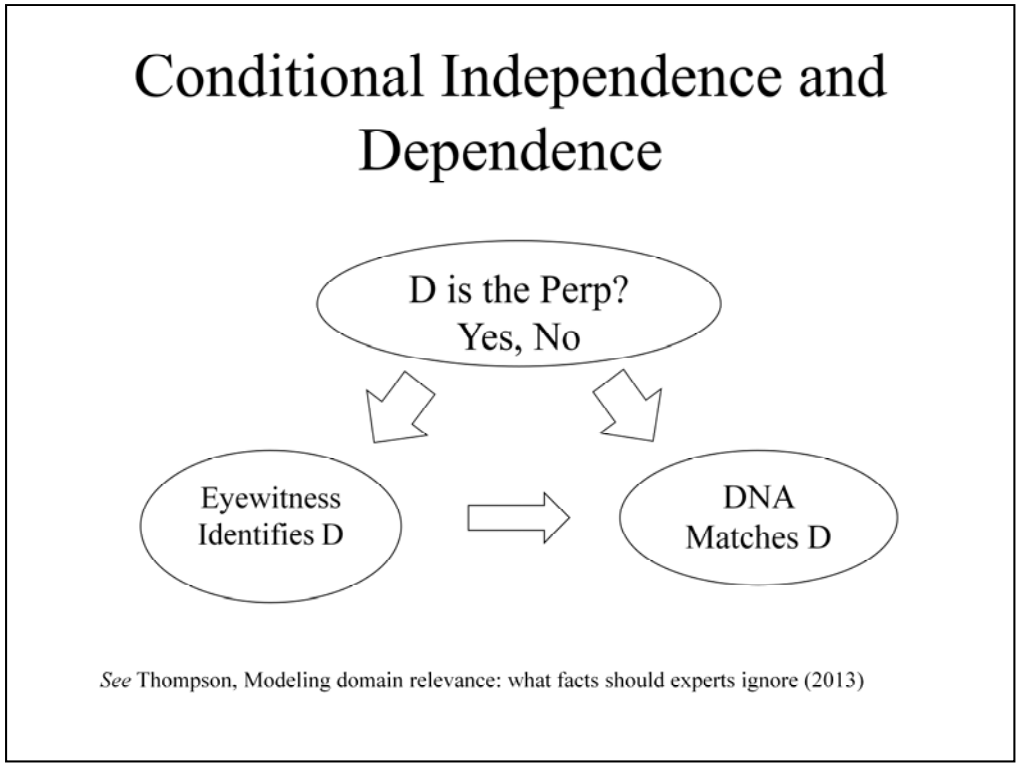


Should analyst consider:

- that an eyewitness identified the defendant?
- whether there is a match at other DNA loci?
- whether the DNA appears degraded at other loci?



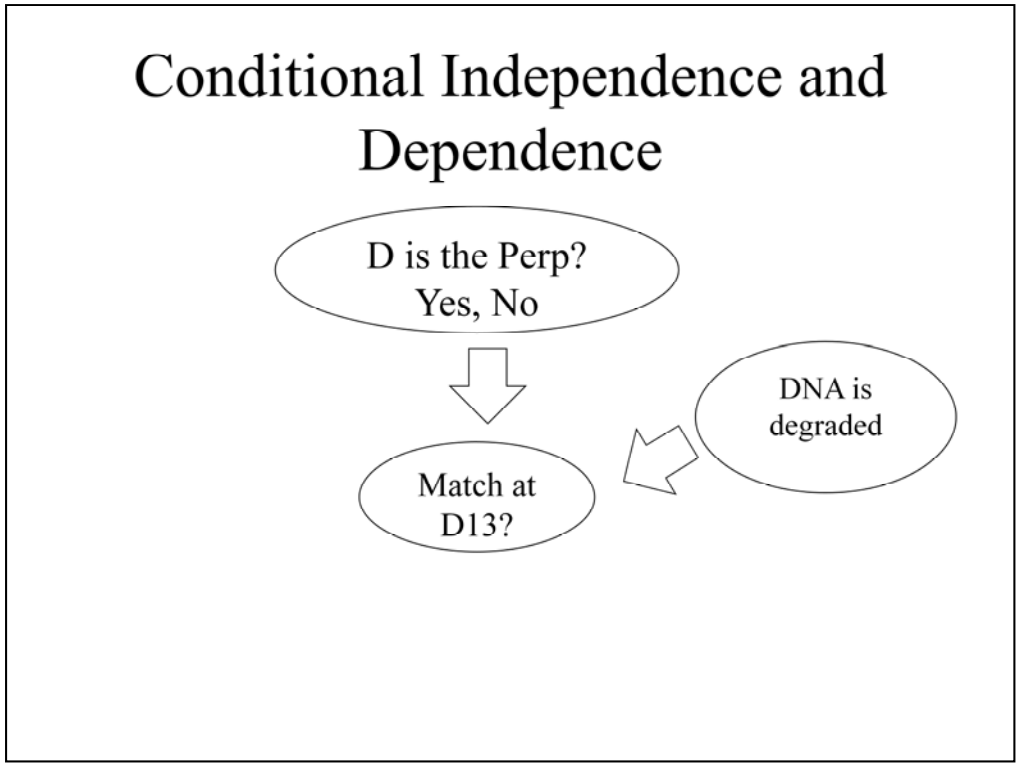
But this line of inference is different. It does not require the analyst to draw conclusions about the probability the defendant is the source. The analyst's judgment rests solely on information within the scientific domain (DNA degradation) and does not depend on the analyst's assessment of the overall likelihood the defendant is the source.



People who study human inference often use diagrams called Bayes nets to illustrate the logical connections between various propositions under consideration. The basic proposition under consideration by the jury is whether the defendant is the perpetrator of the crime. The arrow from this proposition to the eyewitness identification indicates that the eyewitness evidence is probative—we expect an eyewitness identification to be more likely if the defendant is the perpetrator. Similarly, we expect a DNA match to be more likely if the defendant is the perpetrator. But notice there is no arrow from the eyewitness to the DNA match. The two pieced of evidence are said to be conditionally independent.

But when the analyst takes the eyewitness identification into account when evaluating the DNA, that independence is destroyed. The two pieces of evidence are now conditionally dependent.

In a conference paper in the background readings, I use Bayes nets to model the effects of a DNA analyst taking account of eyewitness evidence in a case like this one. The models paint a compelling picture of what happens to the probative value of the forensic evidence when the analyst is influenced by information that would otherwise be conditionally independent. Under all reasonable assumptions about how the influence would work, the probative value of the forensic evidence is reduced, lessening its value for the jury, when the analyst is influenced by the eyewitness.



By contrast, the value of the forensic evidence for the jury is always enhanced, never diminished, when the analyst considers domain-relevant information like the degradation of the DNA.

The Criminalist's Paradox

- By considering "task-irrelevant" information the analyst becomes more likely to reach the correct conclusion
- But undermines the probative value of the conclusion reached
- By helping themselves be "right," analysts may increase chances the justice system will go wrong.

Possible Solution

- How can analysts consider whether the DNA is degraded at other loci (task-relevant) without knowing whether the DNA matches at other loci (task-irrelevant)?
  - See, Krane et al., (2008) Sequential unmasking: a means of minimizing observer effects in forensic DNA interpretation. J Forensic Sci;53(4):1006–7.

The analyst must evaluate the evidentiary DNA sample, assess its level of degradation and the probability of allelic dropout, before knowing the defendant's DNA profile. That way the critical scientific determinations cannot be influenced by backward reasoning.

Blinding Regimes

WMD Forensics

- Efforts to insulate technical analysts from "task-irrelevant" information, including results of other forms of technical analysis
- Technical and investigative information integrated by "all source analysts."

Crime Laboratories

- Analysts often exposes to "task-irrelevant" information
  - Participation in investigations
  - Communications with investigators

Exposure to Task-Irrelevant Information

From crime lab notes:

- *"D. Aboto [prosecutor] left msg. stating this S. is suspected in other rapes but they cant find the V. Need this case to put S away."*
- *"Suspect-known crip gang member--keeps 'skating' on charges-never serves time. This robbery he gets hit in head with bar stool--left blood trail. Miller [deputy DA] wants to connect this guy to scene w/DNA …"*
- *"We need you to match [this latent print] to our crook right away because he is about to leave the country"*

Emotional involvement with cases

DNA Lab Notes (*Commonwealth v. Davis*)

- *"I asked how they got their suspect. He is a convicted rapist and the MO matches the former rape…The suspect was recently released from prison and works in the same building as the victim…She was afraid of him. Also his demeanor was suspicious when they brought him in for questioning…He also fits the general description of the man witnesses saw leaving the area on the night they think she died…So, I said, you basically have nothing to connect him directly with the murder (unless we find his DNA). He said yes."*

Use of task-irrelevant information

Testimony of David Senn in *NY v. Dean*, 2012, RT p. 87:

[After examining a bite mark] …[i]f I then found that DNA [evidence] came back as not excluding that same person, my confidence level would increase. I might be willing to upgrade my opinion from cannot exclude to probable….Now, many odontologists say you shouldn't have any awareness of the DNA results compared to the bite mark…but if I subsequently get them, then I reserve the right to write a revised opinion. And I have done that.

Can forensic scientists ignore task-irrelevant information (if they try)?

- I reject the insinuation that we do not have the wit or the intellectual capacity to deal with bias, of whatever sort. If we are unable to acknowledge and compensate for bias, we have no business in our profession to begin with, and certainly no legitimate plea to the indulgence of the legal system .
  – John Thornton, *J.Forensic Sciences* (2010).

Is the solution just to tell forensic scientists to ignore task-irrelevant information and trust that they are capable of doing so because they are professionals?

Response to Thornton

"Let us be clear. We are not ''insinuating'' that forensic scientists lack this intellectual capacity; we are asserting that it is a proven and well-accepted scientific fact that all human beings, including forensic scientists, lack this capacity."

(Thompson et al. Response to Thornton, *JFS* 2010)

The “bias blind spot” and “introspection illusion”

- People often believe they were influenced by factors that did not affect their judgments,
- and believe they were *not* influenced by factors that *did* affect their judgments.
- So they cannot reliably compensate for their own biases

Contextual bias is recognized and addressed in most areas of science

- Prevalence of blind and double-blind procedures *whenever an important determination rests on subjective judgment*
- Examples from Astronomy to Zoology
- Failure to address observer effects called a *hallmark of junk science*
  - Peter Huber, *Galileo's Revenge: Junk Science in the Courtroom* (1991)

Recommendations for the Commission

1. Issue a statement of principles on the issue of contextual bias

- Three suggested principles:
  - When drawing scientific conclusions, forensic scientists should rely solely on *task-relevant* information.
  - Forensic scientists should shield themselves from *task-irrelevant* information when making judgments that require subjective assessment of potentially ambiguous evidence.
  - When task-relevant information is potentially biasing, it should be disclosed to the analyst at a time and manner designed to minimize its biasing potential.

Recommendations for the Commission

2. Ask the Forensic Science Standards Board (of OSAC) to work with its Scientific Area Committees (with guidance form the Human Factors Committee) to implement the principles articulated by the Commission.

Tasks for OSAC

The Scientific Area Committees to (with advice of the Human Factors Committee) should:

- Determine what information is task-relevant and task-irrelevant for common tasks in each domain of forensic science
- Determine the best ways to shield analysts from task-irrelevant information when making critical judgments in each domain
- Develop standards and model protocols for addressing contextual bias

Thank you

Email:
william.thompson@uci.edu

Australian Journal of Forensic Sciences
Vol. 43, Nos. 2–3, June–September 2011, 123–134

Taylor & Francis
Taylor & Francis Group

What role should investigative facts play in the evaluation of scientific evidence?

William C. Thompson*

University of California, Irvine, Dept. Criminology, Law & Society, USA

Concern about contextual bias has led some authorities to recommend that forensic scientists know as little as possible about the facts of the underlying case when interpreting physical evidence; but concern about contextual ignorance has led other authorities to recommend, to the contrary, that forensic scientists know as much as possible in order to frame questions properly. This article recommends a case manager model that addresses both concerns. This article also responds to standard objections to the use of blind procedures in forensic science, explaining why contextual bias cannot be conquered through willpower; why use of domain-irrelevant contextual facts undermines the value of forensic evidence; how a well-known cognitive illusion (the 'introspection illusion') can mislead forensic scientists into thinking they can control their biases, when they cannot; and how a paradoxical feature of forensic inference (the 'criminalist's paradox') can mislead analysts into thinking they should rely on contextual facts, when they should not.

Keywords: context; bias; blind procedures; observer effect; domain-relevant; case manager

Introduction

When called upon to analyze and interpret physical evidence, what should a forensic scientist know about the facts of the underlying case? Although this is a fundamental question for the forensic sciences, it has received minimal attention in the forensic science literature. What little commentary exists is sharply divided between commentators who have focused on two very different concerns.

Concern about contextual bias has led some commentators to recommend that forensic scientists know as little as possible about the facts of the case. Consider, for example, this passage from an early treatise on document examination by William E. Hagan[1]:

> . . .the examiner must depend wholly upon what is seen [in the forensic examination], leaving out of consideration all suggestions or hints from interested parties; and if possible it best subserves the conditions of fair examination that the expert should not know the interest which the party employing him to make the examination has in the result. Where the expert has no knowledge of the moral evidence or aspects of the case

*Email: william.thompson@uci.edu

ISSN 0045-0618 print/ISSN 1834-562X online

DOI: 10.1080/00450618.2010.541499
http://www.informaworld.com

> in which signatures are a matter of context, there is nothing to mislead him, or to influence the forming of an opinion; and while knowing of the case as presented by one side of the context might or might not shade the opinion formulated, yet it is better that the latter be based entirely on what the writing itself shows, and nothing else. (Ref. 1, p. 82)

Although published in 1894, this statement is entirely consistent with more recent commentary[2,3] calling for greater use of blind procedures in forensic analysis. (I thank Charles Berger and Reinoud Stoel for bringing the Hagen passage to my attention).

Other commentators have focused on a different concern: that ignorance or misunderstanding of the facts of a case may cause forensic scientists to ask and answer the wrong questions. For example, Inman and Rudin[4] described cases in which forensic laboratories performed analyses that were useless and even harmful to a criminal investigation because the analysts misunderstood the factual context of the case. To remedy such problems, they urged that forensic scientists should know as much as possible about the fact of the case. Although they acknowledged a risk that investigative facts might 'subconsciously bias' the examination and interpretation of evidence, they argued that adequate 'checks and balances' exist to minimize that problem. By their account, contextual ignorance is a greater evil than contextual bias. (Inman and Rudin have taken a more nuanced position, however, in recent writings[5,6] and now endorse the need for blind procedures in some circumstances).

Controversy over what forensic scientists should know has grown more heated recently as a result of two developments. On one hand, forensic scientists are becoming involved earlier, and more deeply, in criminal investigations. In order to bring scientific expertise to the crime scene, and avoid the kinds of problems discussed by Inman and Rudin, police in many jurisdictions have been integrating forensic scientists into investigative teams, particularly those assigned to investigate homicides and other major crimes. As a result of this trend, forensic scientists in many jurisdictions tend to have more knowledge of investigative facts than they did in the past.

On the other hand, concerns about contextual bias are growing. Academic commentary suggesting that forensic scientists are subject to contextual bias[2,3] has been supported by empirical studies showing startling evidence of such bias[7–9] and illustrating its consequences[10], and by the discovery of high profile errors that have been attributed, in part, to contextual bias[11–13]. In its 2009 report on forensic science[14], the United States National Research Council acknowledged these concerns and agreed that they are a problem for the field, declaring unequivocally that 'forensic science experts are vulnerable to cognitive and contextual bias' and that this bias 'renders experts vulnerable to making erroneous identifications.' The NRC report called for research on methods to address this problem, and major funding agencies have begun to invest resources in this project (for example, the US National Science Foundation funded an important conference[15]).

The situation might appear, at first glance, to pose an insoluble dilemma. Do forensic scientists have too much contextual knowledge, or too little? Should they institute blind procedures for interpretation, and risk asking the wrong questions; or should they learn as much as they can about the case, and risk contextual bias? Practitioners might be forgiven for feeling that they will be criticized no matter which course they take. But there are ways out of this dilemma. It is possible to address the problem of contextual bias in a scientifically rigorous manner while still

maintaining a useful and appropriate involvement of forensic scientists in the investigative process. In order to achieve this goal, however, it is necessary to make an honest and thoughtful assessment of the appropriate role of forensic scientists in criminal investigations, and the appropriate role of investigative facts in the analysis and interpretation of scientific evidence. This article comments on these important issues.

The role of forensic scientists in criminal investigations

Let us begin by considering the role (or roles) that forensic scientists might play in a criminal investigation. An array of possibilities exist that range from deep involvement in the investigation to little or no involvement.

At one extreme is what I will call the 'CSI model' (based on the television series) in which forensic scientists are integral parts of the investigative team. They work directly with detectives, help determine the direction of the investigation, help evaluate the culpability of suspects, and sometimes even participate in the interrogations. The same individuals perform and interpret tests back at the laboratory.

At the other extreme is what I will call the 'blind service lab model'. In this model, the forensic scientists work in the crime lab and have no direct involvement in the investigation. Their job is limited to analyzing and comparing evidence samples submitted by investigators. The investigators specify what tests and comparisons they want (e.g., determine the nature and chemical composition of this white powder; compare the DNA profile of the blood on this garment to the DNA profile of these reference samples), but provide little or no information about the case. The analysts in the lab are 'blind' in that they do not know the identity of the samples and do not know what is at stake when they make their determinations. They can conclude that the blood stain on 'Garment A' has the same DNA profile as 'Reference Sample #6' without knowing how that determination will affect the case, or even what the case is about.

As far as I know, there are no actual forensic laboratories that fit exactly either the CSI model or the blind service lab model. These two models are prototypical and are used here merely to illustrate the range of variation that is possible in the degree to which forensic scientists are involved in investigations. Most jurisdictions fall somewhere between the two models.

In light of the discussion above, the primary advantages and disadvantages of the two prototypical models should be clear. The CSI model addresses the concerns about contextual ignorance, but leaves analysts vulnerable to contextual bias; the blind service lab model avoids contextual bias during the analysis of evidence, but may result in contextual ignorance. An obvious question that arises when the models are contrasted in this manner is whether there might be some hybrid of the two approaches that would achieve the benefits of each, while suffering the disadvantages of neither. Two such hybrid approaches have been proposed.

One hybrid approach has been called the 'case manager model'. This approach seeks to minimize both contextual ignorance and contextual bias through a separation of functions. Forensic scientists serve either as case managers or analysts. The role of case manager is to communicate with police officers and detectives, participate in decisions about what specimens to collect at crime scenes and how to test those specimens, and manage the flow of work to the laboratory. The role of the

analyst is to perform analytic tests and comparisons on specimens submitted to the laboratory in accordance with the instructions of the case managers.

This separation of function allows case managers to be fully informed of the investigative context (like forensic scientists in the CSI Model), while analyst remain blind to context and are thereby protected from contextual bias (like analysts in a blind service lab). The case managers convey to the analysts only those investigative facts that are directly pertinent to the scientific assessment. For example, if analysts are examining latent prints, they might be told the nature of the surface from which a latent print was collected; if they are analyzing a biological stain, they might be told about the substrate and the environmental conditions to which the stain was exposed. The analysts record the results of their 'blind' analyses in written reports, which are conveyed to the case managers. The case managers then present these reports to the investigative team and provide any advice the investigators need to understand and draw appropriate conclusions from the reports.

A second hybrid approach, which was proposed specifically for forensic DNA analysis, is called sequential unmasking[5]. This approach controls the sequence in which various analyses are performed in order to minimize the potential for contextual bias. A key concern, addressed by this approach, is that knowledge of a suspect's DNA profile might influence an analysts' interpretation of evidentiary samples. To avoid this, analysts make an initial examination of evidentiary samples before learning the profiles of any known or suspected contributors. Based solely on examination of the evidentiary profile, the analyst determines and records the possible genotypes of all possible contributors. At that point, information about known or expected contributors is 'unmasked'. In a sexual assault case, for example, the analyst learns the profile of the victim and any other expected contributors, such as the victim's husband. Then, while still ignorant of the profiles of any suspects, the analyst again examines the evidentiary profile and, in light of the information about known contributors, determines and records the possible genotypes of all unknown contributors. Only at that point, after the analyst's interpretation of the evidentiary sample has been 'fixed' and recorded, is information about the profile of the suspect 'unmasked' so that the analyst can compare it to the evidentiary profile.

Sequential unmasking does not purport to be a complete solution to the problem of contextual bias. It may be feasible only for tests such as DNA analysis for which analysts, after examining evidentiary samples, can determine and list the characteristics of possible contributors. And it will not eliminate all possible forms of contextual bias. For example, it will not prevent contextual bias if and when an analyst who is aware of investigative facts must compare a suspect's profile to an evidentiary profile in order to estimate the probability of allelic dropout under the hypothesis that the suspect was a contributor. But it minimizes one important type of contextual bias, and does so with relatively little extra effort by the analyst and without the need for a second person to act as case manager.

Although these hybrid approaches are available, few laboratories have adopted them. In the forensic laboratories with which I am familiar, blind or anonymous testing is rare. Forensic scientists are almost always informed of the nature of cases; they usually are fully cognizant of the consequences, for the investigation, of their determinations.

Even DNA analysts, who typically spend most of their time at the bench processing samples, stay informed about what the samples are and how the results of their work will affect the case. There often are entries in DNA analysts' lab notes that

show a deep knowledge, if not a personal involvement, in the investigation. Because I am interested in the psychological dynamics of experts' interpretation of evidence, these notes have long been intriguing to me and I have compiled an extensive collection.

For example, a DNA analyst in Virginia wrote:

> Matt told me D. Abato left message stating this S. is suspected in other rapes but they can't find the V. Need this case to put S. away... [D stands for Detective; S for Suspect and V for Victims].

In a California case, the DNA analyst wrote:

> Suspect-known crip gang member–keeps 'skating' on charges-never serves time. This robbery he gets hit in head with bar stool–left blood trail. Miller wants to connect this guy to scene w/DNA ... [Miller was the Deputy District Attorney who was prosecuting the case].

The authors of these notes are not blind testers of anonymous samples. They clearly know and care about the course of criminal investigations. They are in touch with the detectives. As they perform and interpret their tests, they know what is at stake. They are involved. These are the very conditions under which contextual bias is likely to be a problem[2,3].

Why not adopt blind procedures?

It is well established that human beings are vulnerable to contextual bias. The existence of contextual bias (also known as observer effects) has been called 'one of the most venerable ideas of traditional epistemology'[16] as well as 'one of the better demonstrated findings of twentieth century psychology'[16]. Because the problem is widely recognized, scientists in most fields assiduously guard against it[17]. Particularly when scientists must rely on subjective judgment to interpret the results of an experiment, they routinely take careful steps to mask or shield the person interpreting data from extraneous information that might improperly influence the interpretation[17]. Blind procedures are also widely used for peer-review of scientific articles, for grading of written examinations, and for other functions for which it is important to minimize contextual bias[17].

One of my academic colleagues is an evolutionary biologist who has made a lifelong study of the Australian finch. In recent years she has made extensive use of DNA testing to determine the lineage of the birds in her aviaries. It is important for theoretical purposes to know, for example, whether male birds with bright plumage have more 'mating opportunities' and whether the male bird in a bonded pair is actually the father of his partner's offspring. When I asked this academic scientist whether she employed blind procedures when interpreting DNA tests in her laboratory, she was adamant that such procedures are essential. She pointed to the well-known danger that a scientist's pet theories can influence interpretation of data, and stated that she would neither be able to obtain support from major funding agencies nor publish her findings in peer-reviewed scientific journals if she failed to use blind methods. Even if others did not insist on such procedures, she would still use them, she said, to satisfy her own standards of scientific rigor. 'You must understand that this work is extremely important,' she declared, 'it affects our understanding of the entire evolutionary history of the finch!'

I present this anecdote here because it raises an important question. If blind methods are considered essential for studies of bird mating, why do we fail to use such procedures when interpreting forensic tests that may have consequences of the most serious nature for human beings?

Can bias be eliminated through willpower?

I have wondered about the absence of blind procedures in forensic science for many years and, as a result, have made a point of asking forensic scientists why they fail to use such procedures[18]. One common response is that blind procedures are unnecessary for individuals who have proper values and standards of personal integrity. Those who give this response often claim to be insulted at the very suggestion that they might be biased.

This response construes contextual bias as a personal moral failure. According to this view, contextual bias arises when analysts *allow* their scientific judgments to be influenced by extraneous facts; bias is only a problem for analysts who are poorly trained (because they do not realize they should ignore contextual facts) or analysts who are unethical (and therefore are unwilling to ignore contextual facts). Hence, contextual bias, if it exists at all, should be addressed by better training and by weeding out 'bad apples.'

The problem with this response is that it rests on a faulty understanding of human judgment and decision making. Psychologists who study the operation of the human mind in judgmental tasks have shown repeatedly that people lack conscious awareness of factors that influence them[19]. This research has a clear implication for the present discussion: contextual bias cannot be conquered by force of will because people are not consciously aware of the extent to which they are influenced by contextual factors.

One of the most famous and frequently cited articles in twentieth-century psychology[19] reviews a plethora of studies showing that people are often unaware of factors that influence them. When asked, people confidently claim to know whether a particular factor influenced them or not, but these verbal reports are often wrong. People often believe they were influenced by factors that did not affect their judgments; and believe they were *not* influenced by factors that *did* affect their judgments. In one consumer study, for example, researchers discovered they could manipulate people's judgments about the relative quality of four pairs of socks by changing the position of the socks in an array. Whichever pair of socks occupied the right-most position tended to be judged highest in quality. When asked whether they had been influenced by this contextual factor (the position of the socks), people denied it and instead attributed their judgments to inherent properties of the socks. But the results of the study showed their verbal reports were wrong. The quality of the sock was not what was affecting the judgments – whichever pair occupied the right-most position was strongly preferred[19].

Because similar results have been found in hundreds of studies, there is a consensus among cognitive psychologists that people have 'an intellectual blind spot' when it comes to recognizing their own biases[20–26]. The blind spot arises from a fundamental property of the human mind: we 'have no direct access to higher order mental process such as those involved in evaluation, judgment, problem-solving and the initiation of behavior.'[19] In other words, we are not able, through introspection, to directly observe and monitor the mental processes we use to make judgments.

When people are asked to explain their judgments, they cite factors that according to their a priori expectations should have influenced them, but these reports are sometimes wrong because, as studies have repeatedly shown, people can be influenced by factors that they did not know or expect would influence them[19]. People who claim to know whether they were influence by a particular factor are falling victim to what psychologists call the introspection illusion[25]. An article by three of the world's leading researchers on cognitive bias explained the illusion as follows:

> We tend to treat our own introspections as something of a gold standard in assessing why we have responded in a particular way and whether our judgments have been tainted by bias ... [but] the faith people have in the validity of their own introspections is misplaced[25].

The inevitability of contextual bias is recognized and accepted in most scientific fields[17]. One can imagine the reaction if a medical researcher claimed that he need not use blind procedures in his clinical trials because he is a person of integrity who will not *allow* himself to be biased. The claim would not only be rejected, it would invoke derision and ridicule. In my view, forensic scientists who claim to be able to avoid contextual bias through force of will deserve a similar reaction. These claims are unsupportable both because they are wrong and because they display a dangerous ignorance of scientific fact concerning human judgment.

Do forensic scientists make better judgments when they consider context?

A second argument sometimes offered against the use of blind procedures is that contextual knowledge is helpful because it leads to better and more accurate judgments. According to this argument, analysts are more likely to reach correct conclusions – that is, conclusions that coincide with the ground truth – when they consider the big picture and take all evidence into account. Some claim that the term 'contextual bias' is a misnomer. As one put it, 'if this so-called bias leads toward the truth, is it really a bias?'

When making this point, forensic scientists sometimes draw analogies between themselves and other professionals. We would not want physicians to be 'blind' to context when diagnosing illness, they argue, because they will make the best judgments by considering the entire context of the case, not by focusing narrowly on the results of a physical examination or diagnostic tests. I have also heard forensic scientists compare themselves to medical examiners and coroners, who often take into account contextual factors when determining cause of death. When deciding whether a questioned death was due to homicide, suicide or natural causes, for example, medical examiners do not confine themselves to the scientific evidence derived from examinations of the decedent, they also consider what William Hagen called the moral evidence, such as the life circumstances of the decedent, the decedent's writings, and witnesses' statements regarding the apparent physical and mental state of the decedent prior to death. If medical examiners can and do consider contextual factors when determining cause of death, why should a forensic scientist avoid consideration of such factors when deciding whether a suspect was the source of a trace left at a crime scene?

To answer this question one must consider the respective roles of the medical examiner/coroner and the forensic scientist in the legal system. We generally expect

the person or body charged with making an ultimate determination on some legal issue to take into account all relevant evidence. The medical examiner is charged by law with making an ultimate determination about cause of death and is therefore expected to consider all relevant evidence, including contextual factors. Analogously, physicians have the final say on medical diagnoses and are therefore expected to consider all relevant facts, including context. But the forensic scientist occupies a different position.

While forensic science evidence may address a variety of propositions related to crime, including propositions about the source of traces, the activity that led to the deposit of traces, and whether those activities constitute a crime[27,28], it is not the forensic scientist's job to make final determinations about the truthfulness of any of these propositions. The role of the forensic scientist is to provide input to the judicial process, the final judgment about the truthfulness of various propositions in the case is made by the trier-of-fact (typically a judge or jury). While we expect the trier-of-fact to consider all relevant evidence, including contextual factors, it does not logically follow that forensic scientists should consider such factors.

With regard to the medical analogy, the role of the forensic scientist is closer to the role of supporting medical personnel, such as medical lab technicians, radiologists, and experts who interpret imaging tests, than to the role of the physician making the ultimate diagnosis. Accordingly, a question worth asking is whether the judgments of ancillary medical personnel are improved if they consider the full context of a case when interpreting laboratory tests. This question has been examined for a number of medical procedures 29–34, and the answer is a resounding no. For example, one study looked at whether experts interpreting echocardiographs made better judgments with or without being informed about other clinical information in the case, such as the patient's medical history and symptoms, results of blood cultures, and results of a physical examination (including whether a heart murmur was detected)[29]. The study found that exposure to the clinical information greatly increased the false positive rate of these medical experts, and this 'clinical information bias' thereby undermined the diagnostic value of the electrocardiogram. Similar findings have been reported in studies on other medical procedures 30–34. In these medical situations it is clearly better if those running and interpreting diagnostic tests remain unaware of contextual factors and focus their attention solely on their own findings.

The criminalist's paradox

The conclusion that less information is better may be difficult for some forensic scientists to accept, however, due to what I will call 'the criminalist's paradox'. By considering contextual information, analysts may well become more likely to interpret their evidence correctly – that is, to reach conclusions that correspond to what actually happened. Yet by doing so, they also (paradoxically) undermine the ability of the trier-of-fact to determine the truth, and thereby reduce the likelihood the legal system will reach a just outcome. This is the paradox: by helping themselves be 'right' such analysts make it more likely that the justice system will go wrong. By trying to give the 'right' answer, they prevent themselves from providing the best evidence.

To illustrate, let us consider the situation of an expert asked to examine a latent print in order to determine if it was made by a particular suspect. Let us imagine that

it is a difficult case, a close call. When comparing the smudged, partial latent print to the suspect's fingerprint, the analyst sees a number of common features, but some discrepancies. He must think hard and long about whether the similarities are sufficient to conclude that the prints were made by the same finger, or whether the discrepancies are sufficient to conclude that they were not. Suppose that at that point the analyst learns an important contextual fact – the suspect has confessed to police that he touched the item from which the latent print was lifted – and after learning this fact the analyst decides to report that the prints match. Has the analyst done something wrong?

Although we know that false confessions are possible, confessions are generally considered strong and reliable evidence. Consequently, the analyst, if he is thinking logically, should be far more confident that the prints match after hearing about the confession than before. Because the confession is strong evidence that the prints have a common source, the analyst's determination (match or not) is more likely to be consistent with what actually happened if the analyst considers the confession than if he does not. If the goal of the analyst is to be right – that is, to make the determination that corresponds to the truth – he is better off relying on the confession.

If the analyst is influenced by the confession, however, it creates a serious potential problem for the legal system. Part of the problem is that (as with the echocardiographs) the false positive probability (FPP) is likely to be higher (perhaps much higher) if the analyst considers this contextual fact. The FPP is the conditional probability that the analyst will report a match if the two prints in fact come from different people. It can be demonstrated mathematically that the probative value of forensic evidence, as measured by the likelihood ratio, decreases as the FPP increases, other factors being equal[35]. When the FPP is low, even small increases can drastically decrease the value of forensic evidence[36]. Suppose, for example, that the FPP in our latent print case is 1 in 10,000 if the analyst is blind to contextual facts. If learning about the confession increases the false positive probability to 2 in 10,000 then, other factors being equal, the probative value of this forensic evidence will decrease by half; if the false positive rate increases to 1 in 1000, the probative value of the forensic evidence will be reduced by 90%. These effects might be moderated somewhat in practice by simultaneous changes in the probability of a true positive and a false negative, but under any plausible assumptions about the values of these variables, the probative value of the latent print match will decrease to the extent the analyst is influenced by the confession.

One way to look at the matter is that the analyst undermines the independence of the forensic evidence (vis-à-vis other evidence in the case) when he considers contextual facts. If the analyst is unaware of the confession, then the evidence of the confession and the evidence of the latent print exist in a relationship that Bayesian theorists call conditional independence[37]. The only connection between the two pieces of evidence is that both are linked to a source-level proposition about the case – i.e., that the suspect is the source of the latent print. If knowledge of the confession makes the analyst more likely to declare a match, however, the two items of evidence are no longer conditionally independent. The value of the latent print comparison now depends, in part, on the accuracy of the confession. This dependency reduces the incremental probative value of the confession – that is, the ability of the confession to add new, independent insight to the case. One might say the forensic evidence becomes less valuable in its own right because it has been colored (some

might say tainted) by the other evidence in the case. As legal scholar Michael Risinger explained, 'results [of forensic tests] are never made epistemically better, and are often made worse' when analysts are exposed to 'domain-irrelevant information'[38].

Another way to look at the matter is that the analyst's use of contextual facts creates 'double-counting' of evidence. The evidence of the confession is effectively counted twice – once by the analyst, who uses it to resolve his uncertainty about whether the prints match, and again by the jury. The jurors are unlikely to understand or appreciate that the latent print identification was colored by evidence of the confession. They think they are receiving two independent pieces of evidence, and therefore give the evidence as a whole more weight than they should.

If one could trust that the other evidence in the case always pointed in the right direction, then allowing a forensic scientist to be influenced by this information would be less problematic. The rub, of course, is that other evidence sometimes points in the wrong direction. Even confessions are sometimes false, misunderstood, or misreported[39]. When the other evidence points in the wrong direction, the ability of forensic science to correct matters, to put the investigation back on track, is reduced to the extent the forensic evidence is colored by other investigative facts.

Indeed, a lack of independence among different pieces of evidence is a prominent feature of many erroneous convictions. An interesting example is the case of Josiah Sutton, a young Texan who was falsely convicted of rape after being identified by two seemingly independent pieces of evidence[40]. The victim identified him and a DNA analyst reported that he could not be excluded from a mixed DNA sample taken from the victim[40]. But Sutton was exonerated when subsequent post-conviction DNA testing definitively excluded him. Examination of the case showed that the seemingly independent pieces of evidence were actually linked. The victim knew about the DNA evidence when she testified, and this knowledge appears to have bolstered an otherwise shaky identification[40]. And the DNA analyst knew about the victim's identification of Sutton before she conducted the DNA test, which may explain why she misinterpreted the DNA test results, failing to notice that various items of evidence (and particularly information about the profiles of other mixture contributors), taken together, indicated that Sutton should be excluded, rather than included, in the DNA mixture[40]. In other words, each faulty piece of evidence managed to prop up the other. Although the case looked powerfully persuasive to the jury, it rested on an inferential house of cards.

Conclusion

Contextual bias is a serious problem that demands careful consideration by the forensic science community. Forensic scientists will only embarrass themselves if they insist, against the weight of scientific evidence, that they are able to avoid contextual bias by force of will. And they will embarrass themselves further if they take the epistemologically bankrupt position that contextual bias isn't really a bias. The field should instead focus its attention on how best to deal with the problem.

This article offered a number of suggestions for managing contextual bias. It showed that forensic scientists need not choose between knowing too much and knowing too little about the factual context of a case. By using the case manager model, forensic scientists can provide effective advice to police and investigators

while also interpreting evidence in a rigorously blind manner, although these separate functions cannot be performed by the same person. Procedures such as sequential unmasking will also help, particularly in situations where a case manager is not feasible.

The problem of contextual bias will not be solved with excuses and half-measures. Forensic scientists need to join the rest of the scientific community in using more rigorous procedures for interpreting evidence.

References

1. Hagan WE. A treatise on disputed handwriting and the determination of genuine from forged signatures. New York: Banks & Brothers; 1894 (available online at http://books.google.com).
2. Risinger DM, Saks MJ, Thompson WC, Rosenthal R. The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. Cal L Rev. 2002;90:1–56.
3. Saks MJ, Risinger DM, Rosenthal R, Thompson WC. Context effects in forensic science. Sci & Just. 2003;43(2): 77–90.
4. Inman K, Rudin N. Q&A: how much should the analyst know. CAC News. 1997;Fall:18–19. (available online at (http://www.cacnews.org/news/fall97.pdf).
5. Krane DE, Ford S, Gilder J, Inman K, Jamieson A, Koppl R, Kornfield I, Risinger DM, Rudin N, Taylor MS, Thompson WC. Sequential unmasking: a means of minimizing observer effects in forensic DNA interpretation. J Forens Sci. 2008;53(4): 1006–1007.
6. Krane DE, Ford S, Gilder J, Inman K, Jamieson A, Koppl R, Kornfield I, Risinger DM, Rudin N, Taylor MS, Thompson WC. Commentary on Budowle, et al. A perspective on errors, bias and interpretation in the forensic sciences and directions for continuing advancement. J Forens Sci. 2010;55(1): 273–274.
7. Dror IE, Charlton D, Peron A. Contextual information renders experts vulnerable to making erroneous identifications. Forens Sci Intl. 2006;156:74–78.
8. Dror IE, Charlton D. Why experts make errors. J Forens Ident 2006;56(4): 600–616.
9. Dror IE, Rosenthal R. Meta-analytically quantifying the reliability and biasability of forensic experts. J Forens Sci 2008;53(4): 900–903.
10. Thompson WC. Painting the target around the matching profile: the Texas sharpshooter fallacy in forensic DNA interpretation. Law, Prob & Risk. 2009;8:257–276.
11. Stacey RB. Report on the erroneous fingerprint individualization in the Madrid train bombing case. J Forens Ident. 2004;54(6): 706–718.
12. Office of the Inspector General. A review of the FBI's handling of the Brandon Mayfield case. Washington, DC: US Department of Justice; 2006.
13. Thompson WC, Cole SA. Lessons from the Brandon Mayfield case. The Champion. 2005;March:32–34.
14. Committee on Identifying the Needs of the Forensic Sciences Community. National Research Council, Strengthening forensic science in the United States: A path forward. Washington, DC: National Academies Press; 2009.
15. Cognitive Bias and Forensic Science Workshop [Internet]. Illinois. . Northwestern University School of Law [cited 2010 Oct 10]. Available from http://www.law.northwestern.edu/faculty/conferences/workshops/cognitivebias/
16. Nisbett R, Ross L. Human inference. Englewood Cliffs, N.J.: Prentice Hall, Inc, 1980. p. 67.
17. Thompson WC. Interpretation: observer effects. In: Wiley encyclopedia of forensic science. Chichester, UK: Wiley, 2009.1575–1579.
18. The need for blind procedures in forensic science [Internet]. Scientific Testimony: An Online Journal [cited 2010 Oct 10]. . Available from http://scientific.org/open-forum/articles/blind.html
19. Nisbett RE, Wilson TD. Telling more than we can know: verbal reports on mental processes. Psych Rev. 1977;84:231–259.
20. Dawson E, Gilovich T, Regan DT. Motivated reasoning and the Wason selection task. Pers Soc Psych Bull. 2002;28:1379–1387.

21. Ditto P, Lopez DF. Motivated skepticism: use of differential decision criteria for preferred and nonpreferred conclusions. J Pers Soc Psych. 1992;63:568–584.
22. Dunning D, Meyerowitz JA, Holzberg AD. Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving appraisal of ability. J Pers Soc Psych. 1989;57:1082–1090.
23. Kunda Z. Motivated inference: self-serving generation and evaluation of causal theories. J Pers Soc Psych. 1987;53:636–647.
24. Wilson TD, Brekke N. Mental contamination and mental correction: unwanted influences on judgments and evaluations. Psych Bull. 1994;116:117–142.
25. Pronin E, Gilovich T, Ross L. Objectivity in the eye of the beholder: divergent perceptions of bias in self versus others. Psych Rev. 2004;111:781–799.
26. Pronin E. Perception and misperception of bias in human judgment. Trends in Cog Sci. 2006;11:37–43.
27. Cook R, Evett IW, Jackson G, Jones PJ, Lambert JA. A hierarchy of propositions: deciding which level to address in casework. Sci & Justice. 1998;38:231–239.
28. Evett IW, Jackson G, Lambert JA. More on the hierarchy of propositions: exploring the distinction between explanations and propositions. Sci & Justice. 2000;40:3–10.
29. Tape TG, Panzer RJ. Echocardiography, endocarditis, and clinical information bias. J Gen Int Med. 1986;1:300–304.
30. Schreiber MH. The clinical history as a factor in roentgenogram interpretation. J Amer Med Assn. 1963;185:137–139.
31. Doubilet P, Herman PG. Interpretation of radiographs: effect of clinical history. Am J. Radiol. 1981;137:1055–1058.
32. Potchen EJ, Gard JW, Lazar P, Lahaie P, Andary M. The effect of clinical history data on chest film interpretation: direction or distraction? Invest Radiol. 1979;14:404.
33. Swensson RG, Hessel SJ, Herman PG. Omissions in radiology: faulty search or stringent reporting criteria? Radiology. 1977;123:563–567.
34. Elstein AS, Shulman LS, Sprafka SA. Medical problem solving: an analysis of clinical reasoning. Cambridge, MA: Harvard University Press, 1978.
35. Aitken CGG. Statistics and the evaluation of evidence for forensic scientists. Chichester, UK: Wiley; 1995.
36. Thompson WC, Taroni F, Aitken CGG. How the probability of a false positive affects the value of DNA evidence. J Forens Sci. 2003;48(1): 47–54.
37. Schum DA. Evidential foundations of probabilistic reasoning. New York: Wiley, 1994.
38. Risinger DM. The NAS/NRC report on forensic science: a glass nine-tenths full (this is about the other tenth). Jurimetrics. 2009;50:21–34.
39. Drizen S, Leo R. The problem of false confessions in the post-DNA world. No Carolina L Rev. 2004;82:891–1007.
40. Thompson WC. Beyond bad apples: analyzing the role of forensic science in wrongful convictions. Southwestern L Rev. 2008;37(4): 1027–1050.

Modeling Domain-Relevance:
What Facts Should Experts Ignore

William C. Thompson*
University of California, Irvine
November 1, 2013

This is paper was presented at the conference *Blinding as a Solution to Institutional Corruption: When Does Less Information Result in Better Decisions* at the Edmond J. Safra Center for Ethics, Harvard University, Nov 1-2, 2013.

Introduction

People who make important decisions often rely on the assessments of experts. For example, jurors often rely on evidence collected and interpreted by forensic scientists when finding facts and reaching a verdict; physicians often rely on the results of clinical tests conducted and interpreted by medical technicians when diagnosing illness and determining a course of treatment; national security officials often rely on experts' assessments of intelligence data when evaluating the seriousness of potential threats to national security and deciding how to respond. In each of these examples the expert is embedded in an organization that has, as a goal, making good decisions about important matters. But the role of the expert is limited. The expert provides valuable input but the ultimate decision is made by someone else. I will call experts in this position "embedded experts."

In this paper I will discuss an issue that inevitably arises when evaluating the performance of embedded experts: what information should they consider when making their assessments? In particular, what should they know and take into account regarding the broader investigative facts of the case? Consider the following questions:

* Department of Criminology, Law & Society and School of Law, University of California, Irvine, CA. 92697. Email: william.thompson@uci.edu. I thank Michael Risinger for helpful comments on this manuscript.

- Should the forensic scientists who evaluate latent print or DNA evidence that may link a suspect to a crime scene know and consider other evidence (e.g., the suspect's statements; eyewitness accounts) that suggest the suspect was present at the scene?

- Should medical technicians who conduct and evaluate clinical tests know and take into account the patient's symptoms or medical history?

- Should a seismologist who is asked by an intelligence agency to determine the location and nature of a seismic disturbance based on seismographic data know and take into account other information that may relate to the event, such as satellite imagery suggesting there may have been an underground explosion in a certain area?

When deciding which investigative facts embedded experts should know and consider, it is necessary to balance competing concerns. If experts know too much about investigative facts, it may lead to contextual bias[1] and to double-counting of evidence. For example, Michael Risinger (2013) has written about hyptothetical forensic bite mark experts who refuse to issue an opinion about whether a bite mark matches the teeth of a suspect in cases where the bite mark has been swabbed for DNA analysis until they know whether or not the DNA analysis implicates the suspect. Risinger makes the thoroughly plausible suggestion that the bite mark experts will be influenced by the DNA evidence when evaluating whether the bite marks match (contextual bias) and, indeed, that they are unlikely to issue an opinion that contradicts the DNA evidence. This means that the bite mark evidence is redundant with the DNA evidence—it has little or no probative value for evaluating whether the suspect is the biter beyond the value already provided by the DNA evidence. But this redundancy may not be apparent to the jury. Not realizing that the bite mark evidence is completely determined by the DNA evidence, the jurors may think that it provides additional probative value beyond what is provided by the DNA test, even though it does not. As a result, they are likely to give the forensic evidence more weight collectively than

[1] Contextual bias occurs when an expert's judgment on a matter within his expertise is influenced contextual factors that should not have influenced the judgment. The primary goal of this paper is to specify in an analytically rigorous manner what contextual information an embedded expert should and should not consider when assessing evidence.

it deserves—a situation that decision theorists call "double-counting" (Schum, 1994; Schum & Martin, 1983).

But problems can also arise if embedded experts know too little about the broader investigative facts of a case. To evaluate the data they collect, scientific experts often need to estimate the conditional probability of the observed results under various hypotheses about underlying events. Their analyses can go awry if they misunderstand which underlying events the ultimate decision maker needs to assess[2] or if they fail to consider important variables that moderate or mediate the connection between the relevant underlying events and their data. Thus, if they know too little about the investigative context, embedded experts may produce evaluations that are irrelevant or misleading.

There is also a danger of under-counting of evidence. Important factors may be ignored or given too little weight because the decision maker thinks they were taken into account by the embedded experts, and incorporated into their expert opinions, when they were not. Suppose, for example, that national security officials receive scientific estimates of the probability that a seismic disturbance was caused by an underground nuclear detonation rather than an earthquake. If they mistakenly think that these estimates take into account the full range of intelligence available to the agency, when they are in fact based solely on seismic analysis, then other

[2] For example, in a previous publication I discussed a criminal case in which a DNA expert opined that the observed data were a billion times more likely if the defendant was a contributor to a mixed DNA sample found at a murder scene than if the defendant was not a contributor (Thompson, 2009). Unfortunately, the expert's computations rested on the assumption that the other contributors to the mixed sample were unknown individuals. They were in fact known and when the genetic characteristics of these individuals was taken into account, and other computational errors were corrected, then the observed data were only two times as likely if the defendant was a contributor than if he was not.

intelligence data inconsistent with the seismic analysis may never figure (or be given too little weight) in the ultimate decision the officials make about this event.[3]

The question of what embedded experts should know about contextual information has been examined most extensively in the field of clinical medicine. Experts on medical decision making have identified a "clinical information bias" that can arise when medical technicians and ancillary medical personnel interpret laboratory tests while knowing too much about the underlying facts of a case. For example, one study looked at whether experts who interpret echocardiographs made better judgments with or without being informed about other clinical information, such as the patient's medical history and symptoms, the physician's observations, and the results of other clinical tests (Tape & Panzer, 1986). The study found that exposure to this contextual information increased the false positive rate of the echocardiograph procedure by causing experts to report higher rates of abnormality and thereby undermined the diagnostic value of the procedure. The information that the echocardiograph was able to provide to the physician was more probative, and therefore more valuable, when the experts interpreting the test were not informed of other clinical information and focused solely on their own findings. Similar results have been reported in other medical studies (Schreiber, 1963; Doubilet & Herman, 1981; Potchen et al. 1979; Swensson, Hessel & Herman, 1977).

[3] In recent interviews with forensic scientists and security officials at a national security laboratory who are involved in WMD investigations, my colleague Stephan Velsko and I observed some fascinating differences of opinion among experts about which factors are taken into account at various stages of investigations. For example, one official told us that the statistics provided by a biological laboratory regarding the probability that a questioned substance contains various pathogens were posterior probabilities computed by taking into account both the prior probability of observing the pathogen in question and the conditional probabilities of obtaining the observed results of bio-assays if the pathogen in question was and was not present. Further examination of the laboratory's methods suggested, to the contrary, that the statistics provided were likelihood ratios that did not take account of the prior probability of observing particular organisms. This observation raises an interesting question—if the official thinks the experts' statistics take account of prior probabilities, but they do not, then who is considering the priors? In such a case, the answer may be no one.

In forensic science an active debate has arisen about the desirability of blinding criminalists to contextual information. A number of academic commentators, including the author of this paper, have argued that contextual bias is a serious problem in forensic science and have urged greater use of blinding and masking procedures to shield forensic experts from potentially biasing contextual information (Risinger et al., 2002; Dror & Charton, 2006; Krane et al. 2008; Thompson, 2011). A number of forensic practitioners have opposed these proposals on grounds that criminalists need contextual information to perform their work intelligently, and that they have the wisdom and ability to avoid being influenced by information they should not consider (Budowle, et al. 2009; Ostrum, 2009; Thornton, 2010; Wells, 2009). But the debate has been muddied by the failure of those on either side to specify clearly what contextual information a forensic scientist should and should not consider when making judgments. Although the term "domain-relevant" is often used to describe those factors that an expert should consider, it is not particularly helpful to say that an expert should confine himself to considering domain-relevant information without careful specification of what falls in and out of any particular expert's domain. The definitional problems are further complicated by assertions in the literature that certain information may be relevant at one stage of analysis and not others (Krane et al., 2008), and that some facts, although domain-relevant, may be so prejudicial that they should be excluded from consideration nonetheless (Dror, 2012).

This paper will propose a formal method for assessing whether a particular contextual fact is domain-relevant or domain-irrelevant—i.e., whether the fact is one that an embedded expert should or should not consider. The method involves modeling the inferential logic that the expert will use to draw conclusions from the contextual fact. It uses formal models of cascaded inference that were developed by David Schum and his colleagues (Schum, 1994;

Kadane & Schum, 1996; Schum & Martin, 1982) and implements those models using Bayes' nets (Taroni, Aitken, Garbolino & Biedermann, 2006). The method focuses on the probative value of the evidence that the expert provides to the ultimate decision maker.[4] It allows an assessment of whether the probative value of the expert's evidence will be enhanced or degraded if the expert considers a particular contextual fact. The criterion for domain-relevance is then simple: the expert *should* consider the contextual fact if doing so will enhance the probative value of the expert's evidence; the expert *should not* consider a contextual fact if doing so will detract from the probative value of the expert's evidence.

In years past, the method that I am proposing here would have been impractical due to the difficulty of assessing how contextual factors might affect the probative value of expert evidence. As Schum and his colleagues have shown (Schum, 1994; Kadane & Schum, 1996), the probative value of evidence may depend on a variety of contextual factors that can connect in innumerable ways. Equations of daunting complexity are often required to model these connections in order to explore the implications of one items of evidence for the value of another item of evidence (see, e.g., the analysis of the Sacco and Vanzetti case presented by Kadane and Schum, 1996). But this task has become considerably easier with the development of Bayes' net software that incorporates the probability models described by Schum and his colleagues into an easy-to-use graphical user interface (Taroni et al., 2006). In this paper I will show how this software can be used to assess the effect of contextual facts on the probative value of expert evidence.[5]

---

[4] As I will explain more fully below, the probative value of the expert's evidence (as I use that term here) depends on the relative probability of the evidence being produced under propositions that are relevant to the factfinder; it does not necessarily depend on whether the expert's conclusion is correct.

[5] I have used the free demonstration version of a software program called Hugin Lite 7.8 to prepare the figures and perform all the calculations reported in this article. I have verified the accuracy of most of the calculations using the equations developed by Schum and a hand calculator.

Modeling Evidence in a Burglary Case

To illustrate the method, let's begin with a simple model of a common situation in criminal justice. A house is burglarized. An eyewitness identifies a suspect as the man he saw leaving the house after the burglary. A latent print examiner is asked to compare the suspect's fingerprints to latent prints found inside the house on surfaces that may have been touched by the burglar. The examiner finds a partial latent print that looks quite similar to the suspect's fingerprint, but has to think long and hard about whether the latent print contains sufficient detail to justify declaring it a match to the suspect. When deciding whether to report that the prints match, should the examiner consider and give any weight to the eyewitness identification?

Forensic scientists are sometimes told to consider only evidence that is *relevant* to a scientific assessment. Is the eyewitness evidence relevant? According to the Federal Rules of Evidence (Rule 401), evidence is relevant if it has any tendency to make a fact that is of consequence to the matter being determined more or less probable than it would be without the evidence. By this definition, the eyewitness evidence is clearly relevant to the question of whether the latent print came from the suspect. Any logical person would think it more likely that the prints match after hearing that an eyewitness had seen the suspect leaving the scene than before. Hence, the examiner will be sorely tempted to rely on this evidence. It is fair to assume that most forensic scientists are highly motivated by a desire to be right—to reach the correct conclusion. And our uncertain latent examiner may well be more likely to reach the right conclusion—that is, the conclusion that coincides with the ground truth—if she considers the eyewitness identification than if she does not. This fact has contributed greatly to confusion among forensic scientists about the dangers of contextual bias—I heard one forensic scientist

comment on the matter as follow: "if this so-called bias leads toward the truth, it is really a bias?"

To see why it is undesirable for the latent print examiner to rely on the eyewitness evidence when deciding whether to declare the prints to match, we must consider not just *whether* the eyewitness evidence is relevant to the expert's assessment but *how* it is relevant. Bayes' nets are an elegant way to describe the inferential connections between items of evidence.

Bayes' nets describe the inferential connections between events and underlying propositions about the world. In our hypothetical burglary case, those in the legal system must evaluate two alternative propositions: G—the suspect committed the burglary; and NG—someone else committed the burglary. The first event we will consider is E—that the eyewitness identified the suspect as the man he saw leaving the scene of the crime. The probative value of E for distinguishing G,NG depends on the relative probability of E occurring under G and NG, which can be described with a likelihood ratio: $L_E = p(E|G)/p(E|NG)$ (Lempert, 1977). Figure 1 shows a very simple Bayes' net illustrating the connection between E and G,NG. An arrow is drawn from G,NG to E to indicate that the probability of E depends on whether G or NG is true, and hence that E has probative value for distinguishing G,NG.

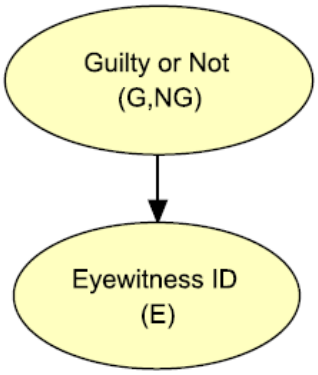


Figure 1: Bayes' Net for Eyewitness Evidence

The second event we will consider is the fingerprint examiner's report on whether the prints match. In order for our model to take account of the possibility of an examiner error, we will distinguish the event that the examiner reports a match, F*, or does not report a match, NF*, from two underlying propositions about the latent print: F—that it was made by the suspect; and NF—that it was made by someone else. We will assume that the probability of F,NF varies depending on G,NG and that the probability of F*,NF* varies depending on F,NF. Figure 2 shows how these propositions and events are displayed in a Bayes' net.

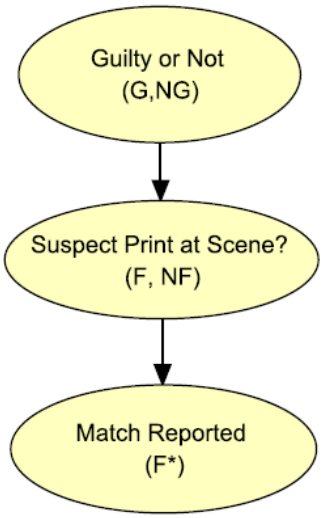


Figure 2: Bayes' Net for Fingerprint Evidence

Once a model like that shown in Figure 2 is created using Bayes' net software, the software can assist the user in assessing how variations in relevant conditional probabilities affect the value of evidence. The user can input various estimates of the key conditionals to see how variations in these probabilities affect the probative value of an item of evidence, as described by its likelihood ratio. For example, the model shown in Figure 2 makes it easy to determine how small changes in the probability that the expert will falsely report a match—

p(F*|NF)—affect the probative value of the fingerprint evidence in the case we have been discussing.[6]

This brings us to the key question—how the probative value of the examiner's report of a fingerprint match, F*, is affected if the examiner is influenced by the eyewitness evidence, E. To answer this question we must consider the collective or combined probative value of E and F*.

Schum and his colleagues have shown that the combined value of two items of evidence for proving some underlying proposition depends, in part, on how the two items are connected. The simplest relationship occurs when the two items are connected solely by their links to the underlying proposition, as illustrated by Figure 3. In this model, the fingerprint examiner is not influenced by the eyewitness evidence. The conditional probability that the examiner will report a match depends entirely on whether the prints are actually from the suspect, which in turn depends on whether the suspect is guilty. Similarly, the conditional probability of the eyewitness identifying the suspect is not affected by the fingerprint evidence. It depends entirely on whether the suspect is guilty. In this circumstance, the two items of evidence are said to be *conditionally independent*. They are not fully independent because they both are affected by whether the suspect is guilty.[7] The examiner is more likely to report that the prints match if the eyewitness

[6] To illustrate, let's assume that the latent print at the crime scene is very likely to be the suspect's if the suspect is guilty—let's say p(F|G)=0.9—but is extremely unlikely to be the suspect's if he is not guilty—let's say p(F|NG)=0.000001. If the finger print examiner was infallible, then p(F*|F)=1.00 and p(F*|NF)=0. When those values are entered into the Bayes' net software, and the software is asked to indicate how the examiner's report of a fingerprint match, F*, affects the probability of G,NG, the software provides a response that indicates the value of the likelihood ratio for F*. In this instance $LR_{F*} \cong 900{,}000$. To see how the value of the examiner's report is affected by the possibility of a false report of a match, we can vary our estimates of p(F*|F) and p(F*|NF) to see how that affects LR_{F*}. This exercise yields the important insight that even seemingly small increases in the probability a match will falsely be reported can dramatically undermine the value of forensic evidence. If the value of p(F*|F) were increased from zero to 0.01 (1 chance in 100) for example, then $LR_{F*} \cong 90$. The small chance of a false match report decreased the value of the reported match 10,000-fold. For detailed explanations and mathematical demonstration of why this is so, *see* Schum & DuCharme (1971) or Thompson, Taroni & Aitken (2003).

[7] If there is no connection at all between two items of evidence, decision theorists say they are *unconditionally independent*.

identifies the suspect, but that correlation arises solely because the suspect is more likely to be guilty if the eyewitness identifies him.

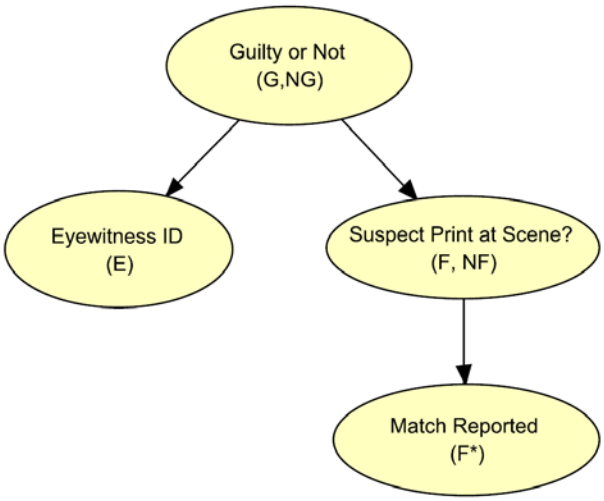


Figure 3: Bayes' Net Showing Conditional Independence of Eyewitness and Fingerprint Evidence

When two items of evidence are conditionally independent, then their joint value for proving the underlying proposition to which they both relate can be determined by multiplying the likelihood ratios of the two items (Schum & Martin, 1983). If, for example, the likelihood ratio for the eyewitness is 10 and the likelihood ratio for the fingerprint evidence is 1000, then the likelihood ratio for the two items 10,000.

When two items of evidence are not conditionally independent, their joint value for proving an underlying proposition can be much lower than the product of their individual likelihood ratios. To illustrate how this can happen, let's consider the relationship shown in Figure 4, in which the examiner's decision to report a fingerprint match is affected by the eyewitness evidence as well as whether the prints are from the same person. Let us assume that learning about the eyewitness identification makes the fingerprint examiner more likely to report a match. The effect need not be large to have a significant impact on the collective value of the evidence. To illustrate, I will instantiate the Bayes' nets shown in Figures 3 and 4 with

conditional probabilities. I will then use Bayes' net software to explore how the examiner's consideration of the eyewitness evidence may affect probative value of the fingerprint evidence.

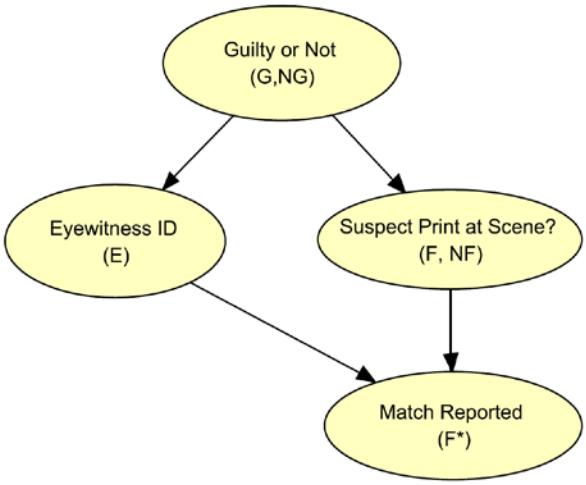


Figure 4: Bayes' Net Showing Conditional Dependence of Eyewitness and Fingerprint Evidence

Let's begin by examining the value of the reported fingerprint match in the model shown in Figure 3, where the fingerprint examiner's judgment is not affected by the eyewitness evidence. Let's focus first on the value of the eyewitness identification, E. Let's assume that p(E|G)=0.9 and p(E|NG)=0.1, which means that the eyewitness evidence, considered by itself, has a likelihood ratio of 9.[8]

For the fingerprint evidence I will assume p(F|G)=0.9 and p(F|NG)=0.000001, which means that I judge the presence of the suspect's fingerprint at the crime scene is highly probative of guilt.[9] But I do not believe the examiner is perfectly reliable. I will assume that p(F*|F)=0.9—in other words, there is a 90% chance the examiner will report the latent print matches the suspect if the latent print was actually made by the suspect; I will also assume that p(F*|NF)=0.001, which means that there is 1 chance in 1,000 that the expert will report the latent

[8] LR_E= 0.9/0.1 = 9. The conditional probabilities used here are arbitrary estimates that I will use solely for purposes of illustration. Readers who are curious about the effect of making different assumptions about the conditionals can download the free software and explore the matter themselves.

[9] LR_F = 0.9/0.000001 = 900,000.

print matches if it was not made by the suspect. Entering these estimates into Bayes' net software makes it easy to compute the likelihood ratio for the reported fingerprint match. Given these conditional probabilities, then $LR_{F*} \cong 809$.

Using the Bayes' net shown in Figure 3 we can also estimate the joint or combined value of the eyewitness and fingerprint evidence, $LR_{E,F*}$. According to the software, $LR_{E,F*} \cong 7281$. In other words, the joint value of the two items of evidence is the product of the two likelihood ratios—7281 = 809 x 9. This means that the items each have as much probative value when the two are considered together as when each is considered separately, and this will always be true when the items are conditionally independent.

We now come to the crux of the matter. What happens to the probative value of the two items of evidence when they are no longer conditionally independent because the fingerprint examiner has been influenced by the eyewitness evidence when deciding whether to report that the prints match? Let's assume that knowledge of the eyewitness identification increases the probability that the examiner will report a match. Suppose, for example, that p(F*|F) increases from 0.9 to 0.95 and that p(F*|NF) increases from 0.001 to 0.002—in other words, knowing about the eyewitness cuts the false negative rate by half and causes the false positive rate to double. Entering these probability estimates into the Bayes' net software allows us rapidly to determine that the likelihood ratio describing the value of the combined value of the evidence under these conditional probabilities is approximately 3846, which is slightly more than half the size of the likelihood ratio for these two items when they were conditionally independent. In other words, the collective value of the eyewitness and fingerprint evidence is substantially diminished when the fingerprint examiner is influenced by the eyewitness evidence.

The decrease in probative value arises almost entirely from the increase in the probability of a false report of a match. As noted earlier, even small increases in the false positive probability of a forensic test—in this example, an increase from 1-in-1000 to 2-in-1000—can dramatically undermine the probative value of the test results. The value of the eyewitness identification is the same in the model shown in Figure 4 as in the model shown in Figure 3. What changes is the incremental value provided by the fingerprint examiner's report. When the fingerprint examiner's report is colored by the eyewitness identification, its incremental probative value is diminished. In our example, the evidence of the fingerprint match increases the likelihood ratio by a factor of 809 when it is conditionally independent of the eyewitness evidence, but by a factor of only 427 when it is influenced by the eyewitness evidence. By taking into account the eyewitness evidence, the fingerprint examiner diminishes substantially the incremental value of the evidence that she can provide to the trier-of-fact.

Thus we have an answer to the question of whether the fingerprint examiner should take into account the eyewitness evidence. If the eyewitness evidence is likely to influence her in the way I have suggested, by being more likely to report a match, then clearly she should not take it into account. Of course we must consider whether the conditional probabilities that I have used in these illustrations are realistic and whether the results of the analysis might change with other plausible estimates. I have tried a variety of different estimates and have found that the incremental probative value of the fingerprint evidence is diminished under any estimates that I consider even remotely plausible. While it is possible to imagine hypothetical circumstances in which the probative value of fingerprint evidence would be unaffected or even enhanced if the

examiner considered the eyewitness, these circumstances seem fanciful and would only occur if there were serious deficiencies in the expert's analytic method or reporting.[10]

In order to use this method to draw conclusions about which contextual facts an expert should and should not consider, one must make judgments about how exposure to those likely to influence expert judgment. The usefulness of this method will depend on the accuracy of those judgments.

The Criminalist's Paradox

The analysis just described provides a formal illustration of what I have elsewhere called "the criminalist's paradox":

> By considering contextual information, analysts may well become more likely to interpret their evidence correctly – that is, to reach conclusions that correspond to what actually happened. Yet by doing so, they also (paradoxically) undermine the ability of the trier-of-fact to determine the truth, and thereby reduce the likelihood the legal system will reach a just outcome. This is the paradox: by helping themselves be 'right' such analysts make it more likely that the justice system will go wrong. By trying to give the 'right' answer, they prevent themselves from providing the best evidence (Thompson, 2011).

I believe this paradox underlies much of the resistance in the field of forensic science to more widespread use of blinding and masking procedures. The paradox makes it difficult for embedded experts to recognize the advantages of being blind to other evidence. That is because the advantages accrue to others—the ultimate decision makers.

[10] Suppose, for example, that an examiner was unwilling to report a latent print match unless she knew that there was some other evidence to indicate that the person who matched had touched the item from which the latent print was lifted. In that case, the expert's evidence would be worthless unless the expert was exposed to other evidence, in which case it might have some value. If the expert's reporting standards were flawed in this particular way, then the justice system might arguably be better off if such an expert was exposed to other evidence than if the expert was blind. But such a situation seems extremely unlikely. If such a situation arose, it would clearly be better to address it by training the examiner to adopt more appropriate reporting standards than by using exposure to contextual facts to overcome the effects of an inappropriate reporting threshold.

References

Budowle B, Bottrell MC, Bunch SG, Fram R, Harrison D, Meagher S, et al. (2009). A perspective on errors, bias, and interpretation in the forensic sciences and direction for continuing advancement. J Forensic Sci 54(4):798–809.

Doubilet P, Herman PG. Interpretation of radiographs: effect of clinical history. Am J Radiol. 1981;137:1055–58.

Dror IE. Combating bias: the next step in fighting cognitive and psychological contamination. J Forensic Sci 2012;57(1):276-7.

Dror IE, Charlton D. Why experts make errors. J Foren Ident 2006;56(4):600–12.

Dror IE, Rosenthal R. Meta-analytically quantifying the reliability and biasibility of forensic experts. J Forensic Sci 2008;53:900–3.

Kadane, Joseph B. & Schum, David A. (1996). A Probabilistic Analysis of the Sacco and Vanzetti Evidence. New York: John Wiley and Sons.

Krane DE, Ford S, Gilder J, Inman K, Jamieson A, Koppl R, et al. Sequential unmasking: a means of minimizing observer effects in forensic DNA interpretation. J Forensic Sci 2008;53(4):1006–7.

Lempert, Richard (1977). Modeling relevance. Michigan Law Review, 75: 1021-1057.

Ostrum B. Commentary on: sequential unmasking: a means of minimizing observer effects in forensic DNA interpretation. J Forensic Sci 2009;54(6).

Potchen EJ, Gard JW, Lazar P, Lahaie P, Andary M. The effect of clinical history data on chest film interpretation: direction or distraction? Invest Radiol. 1979;14:404.

Risinger, R. Michael (2013). "Costs?" of Reducing Information Available for Decision. Paper to be presented at the conference *Blinding as a Solution to Institutional Corruption: When Does Less Information Result in Better Decisions* at the Edmond J. Safra Center for Ethics, Harvard University, Nov 1-2, 2013.

Risinger DM, Saks MJ, Thompson WC, Rosenthal R. (2002). The Daubert/Kumho implications of observer effects in forensic science: hidden problems of expectation and suggestion. Calif Law Rev 90:1–55.

Schreiber MH. The clinical history as a factor in roentgenogram interpretation. J Amer Med Assn. 1963;185:137–139.

Schum, David A. (1994). Evidential Foundations of Probabilistic Reasoning. New York: John Wiley and Sons.

Schum, David A. & DuCharme, Wesley M. (1971). Comments on the relationship between the impact and the reliability of evidence. Organizational Behavior and Human Performance, 6, 111-131 (1971).

Schum, David A. & Martin, Anne (1982). Formal and empirical research on cascaded inference in jurisprudence. *Law & Society Review 17*: 105-151.

Swensson RG, Hessel SJ, Herman PG. Omissions in radiology: faulty search or stringent reporting criteria? Radiology. 1977;123:563–567.

Tape TG, Panzer RJ. Echocardiography, endocarditis, and clinical information bias. J Gen Int Med. 1986;1:300–4.

Taroni, Franco, Aitken, Colin, Garbolino, Paolo, and Biedermann, Alex (2006) Bayesian Networks and Probabilistic Inference in Forensic Science. Chichester, UK: John Wiley and Sons.

Thompson William C. What role should investigative facts play in the evaluation of scientific evidence. Aust J Forensic Sci 2011;43(2-3):123-34.

Thompson William C. (2009) Painting the target around the matching profile: the Texas sharpshooter fallacy in forensic DNA interpretation. *Law, Probability & Risk 8*:257–76.

Thompson, William C., Taroni, Franco & Aitken, Colin (2003). How the probability of a false positive affects the value of DNA evidence. Journal of Forensic Sciences, 48: 47-54.

Thornton JI. Letter to the editor—a rejection of ‘‘working blind’’ as a cure for contextual bias. J Forensic Sci 2010;55(6):1663

Wells JD. Commentary on: sequential unmasking: a means of minimizing observer effects in forensic DNA interpretation. J Forensic Sci 2009;54(2):500.

Note: This is an excerpt of a grant proposal that was recently funded by the National Institute of Justice, NIJ award 2014-DN-BX-K032

Response to: US-DOJ Solicitation: (CFDA No. 16.560)
Research and Development in Forensic Science for Criminal Justice Purposes

Entitled:

Developing Effective Methods for Addressing Contextual Bias in Forensic Science

Principal Investigators:

Prof. William C. Thompson Department of Criminology, Law & Society University of California, Irvine Irvine, CA. 92697 USA Ph. 949-824-6156, Fax 949-824-3001 william.thompson@uci.edu	Dr. Michael C. Taylor Institute of Environmental Science and Research 27 Creyke Rd, Christchurch 8031 New Zealand Ph. (64) 3 351-0013, Fax (64) 3 351-0046 michael.taylor@esr.cri.nz

Submission Date: April 21, 2014

I. Statement of the Problem

In 2009, the National Research Council (NRC) identified a number of weaknesses in the scientific foundations of forensic science (NRC, 2009). One concern was that "forensic science experts are vulnerable to cognitive and contextual bias" that "renders experts vulnerable to making erroneous identifications." (p. 4, note 8).

Several empirical studies have found evidence that supports these concerns (*see* Kassin et al., 2013 for a review). For example, Dror and his colleagues have presented evidence that fingerprint and DNA examiners can be influenced by task-irrelevant information when making forensic comparisons (Dror & Hampikian, 2011; Dror, Charlton & Peron, 2006; Dror & Rosenthal, 2008). One of the authors of this proposal (Taylor) recently completed an experimental study that showed that bloodstain pattern analysts were influenced by case-specific contextual information (Taylor, Laber, Kish & Owens, 2014).

These findings are consistent with a large psychological literature showing that human beings are susceptible to contextual bias, that people can be biased without being aware of it, and that even well-trained experts are susceptible to bias (for reviews of this literature, *see* Risinger et al., 2002; Saks et al., 2003; Thompson, 2009a; Kassin et al., 2013). In fact, given the nature of expertise, there are good reasons to suggest that experts might be *particularly* vulnerable to bias (Dror, 2011).

To address these concerns, academic commentators have encouraged forensic scientists to make greater use of "blinding" or "masking" procedures designed to shield analysts from exposure to task-irrelevant information that might inadvertently influence them (Risinger et al. 2002; Krane et al. 2008). But these proposals face a number of conceptual and practical difficulties (see e.g., Butt, 2013; Charlton, 2013; Budowle et.al., 2009; Ostrum, 2009; Wells, 2009). Forensic laboratories are complex organizations, in which the task of evaluating evidence is distributed across a number of individuals. Some of these individuals need to be informed about the context of a case in order to do their jobs properly (see e.g. Butt, 2013). For example, to decide what samples to collect at a crime scene and what examinations or analyses are needed, laboratory personnel must communicate with the police about the nature of the case and the information that the police need to solve it. To combine findings from multiple examinations (e.g., bloodstain patterns and DNA analysis of the blood) into an integrated interpretation of what happened at the crime scene, laboratory personnel may also need information about the context. Even those forensic scientists who confine themselves to addressing "source-level" hypotheses—i.e., whether two items have a common source—may need contextual information about the history of the samples (e.g., the environment in which they were collected, their likely age) in order to make an intelligent assessment.

Commentators have suggested three ways that these difficulties might be addressed. One approach is to separate functions in the laboratory, allowing some individuals to be aware of context while others are "blind." In the "case manager model" (Thompson, 2011; Stoel et al.,

2014; Found and Ganas, 2013), forensic scientists can serve either as case managers or analysts. The role of the case manager is to communicate with investigators, participate in decisions about what specimens to collect at crime scenes and what examinations are needed, and to manage the flow of work in the laboratory. In contrast, the role of the analyst is to perform analytic tests and comparisons on specimens submitted to the laboratory in accordance with the instructions of the case managers. In theory, this separation of functions allows case managers to be fully informed of the investigative context while analysts remain blind to context and are thereby protected from contextual bias. Whether such a separation is feasible and effective in practice is one of the issues that we propose to examine.

A second approach is to sequence the work flow in the laboratory in a manner designed to minimize the potential for contextual bias. The most widely discussed proposal of this type calls for DNA analysts to use a procedure known as "sequential unmasking" when comparing profiles (Krane et al. 2008; Stoel et al., 2014). Sequential unmasking requires analysts to make an initial examination and interpretation of the evidentiary profiles before learning the profiles of any known or suspected contributors. The DNA profiles of known contributors and possible suspects are then "unmasked" in a specific sequence to minimize the likelihood that information about the reference profiles will influence interpretation of the evidentiary samples. The FBI laboratory has adopted a similar procedure for latent print analysis. Called "linear ACE-V," the FBI's procedure involves temporary masking of reference prints while analysts make and record their initial assessments of the evidentiary prints (Office of the Inspector General, Department of Justice, 2011; *but see* Cole, 2013, who notes that details of the FBI's protocol are not yet public). The practicality and effectiveness of such procedures is another issue we propose to examine.

A third way to address the problem of contextual bias is through blind review procedures. Under this approach, the analyst who does the initial testing and interpretation has access to contextual information about the case, but colleagues who are assigned to review and verify the analysis do not. Proponents of this approach suggest that it may be easier and more practical to introduce blinding or sequential unmasking procedures during a review process than during the primary analysis. The feasibility and efficacy of blind review procedures is another issue that we propose to examine.

To summarize, commentators have proposed three possible approaches for dealing with contextual bias: utilization of a case manager; sequential unmasking, and blind review. But these proposals are relatively new—even the most avid proponents acknowledge that a number of issues will need to be examined and evaluated before they are adopted. There is uncertainty and disagreement about (1) the circumstances under which such procedures are necessary; (2) whether they would be practical and effective; and (3) how they might be implemented. The research we propose here is designed to cast new light on these important questions.

II. Project Design and Implementation

A. Overview

There are three components to the proposed research:

1) Interviews

The first component of the research involves interviews with directors and section heads of major forensic laboratories to assess their views of contextual bias and to gain insight into the feasibility of various proposals for dealing with it. During these interviews, we will seek answers to the following questions: (1) How much do these managers and leaders know about contextual bias, and what are their views of the issue? (2) What steps, if any, have they taken (or do they contemplate taking) to address the problem of contextual bias? (3) What issues or problems have they encountered (or do they anticipate) when taking steps to address contextual bias?

For reasons discussed below, we believe that forensic scientists' views on some of these issues—as well as their current practices—may vary depending on the institutional and legal environment in which the laboratory operates. For example, we anticipate that those who work in adversarial legal systems might view some of these issues differently than those in inquisitorial systems, and that those who work in laboratories operated by law enforcement agencies may differ from those who work in laboratories that are separated institutionally from law enforcement. To gain insight into this possibility, we will sample our forensic scientists from a diverse range of environments in which forensic science is practiced, including laboratories in inquisitorial legal systems (Switzerland and the Netherlands) as well as adversarial systems (United States, New Zealand, Australia). Within each type of legal system, we will sample from well-regarded laboratories that are operated by law enforcement agencies, as well as laboratories that are operated by agencies that are independent of law enforcement.

We will also include laboratories that are known to have taken positive steps to address the problem of contextual bias. At this point, it is clear that some laboratories have taken more steps than others (s*ee, e.g.*, Stoel et al. 2014, mentioning that the Netherlands Institute of Forensic Science has adopted sequential unmasking procedures for DNA testing and other context management procedures for firearms; *also* Found & Ganas, 2013, describing a "context management scheme" adopted by the Document Examination Unit of the Victoria (Australia) Police Forensic Services Department). Lessons learned by laboratories that have already taken steps to address the problem of contextual bias are likely to be of great value to other laboratories that are contemplating similar steps. A good way to assess the time and effort required by proposals such as sequential unmasking and blind review, for example, is to ask forensic scientists who are already using such procedures how well those methods are working and what

practical difficulties they entail. This first component of our research will address these questions.

2) Focused Research on Bloodstain Pattern Analysis (BPA) and Handwriting Analysis

The second component of the proposed research focuses in detail on the decision-making environment and decision processes of experts in two forensic domains: bloodstain pattern analysis and handwriting analysis. Both of these disciplines involve the analysis of patterns. Furthermore, in both disciplines experts rely almost exclusively on human perceptual and cognitive processes to form opinions regarding evidence. Hence these are disciplines in which the potential for contextual bias is likely to be an important issue (Found & Ganas, 2013; Miller, 1984). But the two disciplines differ in important ways as well. While handwriting analysis is focused primarily on source-level propositions (e.g., was a questioned document written by the same person who wrote the exemplar documents?), bloodstain pattern analysts are more likely to address activity-level propositions (e.g., was this bloodstain pattern caused by expiration or violent impact?). Consequently, the nature of the contextual information that analysts need to consider—and the measures needed to manage contextual bias—are likely to be very different in the two disciplines.

Our proposed studies will provide insight into the decision-making environment and decision processes in these two disciplines through: (1) detailed interviews with trained analysts; and (2) experimental studies in which analysts are asked to evaluate realistic case materials under controlled conditions.

The interviews will incorporate "think-aloud" studies, in which analysts will be asked to explain their thinking as they examine evidence from actual cases. Analysts will be offered various types of contextual information and asked to explain what contextual information they need (and why they need it) to do their jobs properly and, correspondingly, what contextual information they do not need (and why they consider it irrelevant to their task). The interviews will also explore the channels of communication via which analysts receive contextual information, allowing us to assess how difficult it might be to insulate them from task-irrelevant facts arriving through various channels.

The experimental studies will provide a controlled way of exploring how changes in contextual information and the nature of the case can affect analysts' decision-making processes.

The purpose of the proposed studies is not simply to demonstrate that contextual bias can occur. As discussed below, there is strong reason to believe that it can. Rather, the studies proposed here have two overarching goals: first, to gain additional insight into the process of decision-making by analysts in two different but equally important pattern matching disciplines; and second, to gain insight into the need for—and feasibility of—various procedures that might work to mitigate bias in each of these domains.

3) Protocol Development

The third component of this project is the development of a practical set of Contextual Information Management (CIM) procedural models that could realistically be implemented in forensic agencies. Relying on insights gained from our interviews and experimental studies, as well as a broad review of pertinent scholarly literature, we will develop written protocols that reflect best practices for mitigating contextual biases while also assuring that analysts are provided with all of the task-relevant information required to perform their analyses in a rigorous and proper manner. While our primary focus will be on protocols for handwriting analysis and bloodstain pattern analysis, we believe that model protocols for these two areas will be useful to forensic scientists in other pattern matching disciplines as they consider ways to improve their laboratory SOP's and protocols in an effort to address the problem of contextual bias.

B. Literature Review and Analysis

1) Observer effects, confirmation bias and contextual bias

Scientists have long recognized that the results of an observation can be affected by the state of the observer. As early as 1795, astronomers recognized that these observer effects can distort and undermine the accuracy of experts' observations (Risinger, et al., 2002). Where the observers' preconceptions or motives influence the interpretation of data, the phenomenon is sometimes called examiner bias or confirmation bias, although it is important to note that the "bias" entailed in the phenomenon may occur without the observer intending or even being aware of it (Thompson, 2009a).

The term context effect, sometimes used synonymously with observer effect, originated in psychology. It was initially used to describe circumstances in which the perception of a stimulus is affected by the surrounding context, as where a gray object looks lighter against a dark than a light background (Dresp-Langley & Reeves, 2012). In forensic science, however, the term context effect has been used more broadly to describe situation in which the results of a forensic analysis are affected by the circumstances in which it is performed, and particularly by the information available to the analyst. For example, a latent print examiner might become more likely to identify a latent print as that of the suspect when told that another analyst has already made the identification, or when told that other evidence indicates the suspect made the print (Dror & Charlton, 2006; Dror, Charlton & Peron, 2006). The other evidence might be said to provide a context that influences the analyst's interpretation of the data. A context effect becomes contextual bias when the influence of context is deemed improper or inappropriate, as when an analyst's scientific conclusions are affected by contextual information that should not have influenced them.

Concerns about contextual bias have been raised in a variety of scientific fields (Risinger, et al. 2002). The most common way to address such concerns is to adopt blind or double-blind methods that shield the person interpreting critical data from extraneous information that might improperly influence the interpretation (Thompson, 2011; 2009a; Risinger et al., 2002). Blind procedures are most common in fields where practitioners must rely on subjective judgment to

interpret data, such as medicine and psychology; they are seen less often in research in the physical sciences, perhaps because data in those areas is viewed as more objective and less subject to human interpretation (Sheldrake, 1999). But blind procedures are widely used in all fields for peer-review of scientific articles, for grading of written examinations, and for other functions for which contextual bias is a concern. These procedures avoid contextual bias by the straightforward expedient of preventing exposure to potentially biasing information.

There is growing evidence that forensic scientists are susceptible to contextual bias. Evidence of contextual bias has been found in latent print analysis (Dror, Charlton & Peron, 2006; Dror & Rosenthal, 2008), document examination (Miller, 1984), bite mark analysis (Osborne, Woods, Kieser & Zajac, 2013) and DNA analysis (Dror & Hampikian, 2011; Thompson, 2009b). The concerns have also been reinforced by the discovery of errors in latent print analysis (Stacey, 2004; Office of Inspector General, 2006), bite mark analysis (Pretty & Sweet, 2010) and DNA testing (Thompson, 2008; 2013) that have been attributed, at least in part, to contextual bias. In its 2009 report on forensic science, the National Research Council acknowledged these concerns and agreed that they are a problem for the field, declaring unequivocally that "forensic science experts are vulnerable to cognitive and contextual bias" and that this bias "renders experts vulnerable to making erroneous identifications." (NRC, 2009, p. 4, note 8).

Why have forensic scientists lagged behind other fields in addressing contextual bias? According to the NRC report, "[t]he forensic science disciplines are just beginning to become aware of contextual bias and the dangers it poses." (p. 185). A key issue we intend to explore in the interview component of this project is why forensic scientists have lagged behind other scientific fields in efforts to address this issue. Is the problem simply ignorance (arising from inadequate training) or are there other factors at play?

2) Do role conflicts inhibit reform?

One possibility that we intend to examine is that forensic scientists are reluctant to "blind" themselves to potentially biasing contextual facts because they are involved in aspects of the investigative process in which such facts are relevant. As noted earlier, forensic scientists may need detailed information about the nature of the case, the claims made by witnesses, and investigators' theories of the matter in order to decide which specimens to collect from the crime scene and what analytic examinations are needed. But this contextual information may be unnecessary and potentially biasing to a forensic scientist who performs analytic tasks, such as determining whether a specific individual could be the source of a latent prints or a contributor to a DNA sample. In other words, contextual information that is relevant to one task may be irrelevant and potentially biasing for another. The problem, then, may be that the same person is expected to perform (or wants to perform) both kinds of tasks. In the process of minimizing the potential for contextual bias on analytic tasks, these forensic scientists might fear that they would be compromising their effectiveness at other tasks, such as managing cases or advising investigators. The interview component of this project will explore forensic scientists' perceptions of the feasibility, practicality, and desirability of various proposals for resolving this dilemma.

As noted above, one proposal is to require forensic scientists who participate in a particular investigation to assume one of two possible roles—that of a case manager or that of an analyst. Case managers, who are fully aware of case context, would communicate with investigators, decide what examinations should be conducted, and assist investigators in understanding the results of examinations and evaluating various theories of the case. Analysts would conduct examinations in a manner specified by case managers and would be blind to task-irrelevant contextual information—at least until they completed the examinations and issued reports. But is it really feasible in a forensic laboratory to separate functions in this manner? What would be the advantages, disadvantages, and costs of this proposal? What issues have laboratories encountered (or do they anticipate they might encounter) with the adoption of such procedures? In our interviews, we will answer these questions. We will also explore forensic scientists' perceptions of the practicality and desirability of other proposals for addressing contextual bias, including sequential unmasking (Krane, 2008) and blind review.

3) Do adversarial pressures and incentives inhibit reform?

Some commentators have attributed problems in forensic science to pressures and incentives that arise from an adversarial system of justice. The claim is that forensic scientists may identify too closely with the side that they generally serve—whether the prosecution or defense—resulting in an orientation toward "winning" cases rather than doing the most careful science. For example, Mnookin, et al. (2011) attributed a variety of problems in forensic science to the immersion of many forensic scientists in a "culture of law-enforcement." The 2009 NRC report went so far as to recommend an institutional separation of crime laboratories from law enforcement as a key element of its proposed "path forward" for forensic science. Others have pointed out that defense experts may also be influenced by partisanship and adversarial pressures (Murrie et al., 2013).

How might adversarialism affect the way in which forensic scientists respond to the problem of contextual bias? One possibility is that a forensic scientist who is highly motivated to help "his side" win might resist blinding or masking procedures because he needs contextual information in order to slant his findings in a manner helpful to his side. While we doubt that there are many forensic scientists who are influenced by such crude motives, there may be more subtle incentives for resisting these procedures. For example, forensic scientists might find it more rewarding to be directly involved with investigators or lawyers in developing theories of a case than to operate in the contextual vacuum required for "blind" analysis.

Our interviews will explore the extent to which forensic scientists are involved with investigators and lawyers, how they feel about those contacts, and whether the nature of those contacts is associated with perceptions of the need for—and feasibility of—blinding procedures.

Another possible theory is that an adversarial environment may make it more difficult for forensic scientists to recognize contextual bias as a problem. That is, exposure to contextual information may come primarily from contacts with trusted colleagues who are investigators or

lawyers. Forensic scientists' respect for these colleagues may make the information that they provide seem benign, natural, helpful, and perhaps even necessary, rather than as potentially biasing.

One of the issues we will explore in our proposed interviews is whether forensic scientists who work in adversarial systems of justice view the issue of contextual bias differently than those who work in inquisitorial systems. In inquisitorial systems, forensic experts generally report to a court rather than to one of the parties in a contested case; they are therefore less likely to experience adversarial pressures. Whether forensic scientists in those setting have a different attitude toward contextual bias, and a different approach to dealing with it, is one of the questions this research will address. We will also compare forensic scientists who work for law enforcement agencies with those who work for agencies that are independent of law enforcement, to gain insight into whether a "culture of law enforcement" is linked to views and practices of forensic scientists on the issue of contextual bias.

4) Are forensic scientists blind to their own biases?

Psychologists have shown that people have a "blind spot" when it comes to recognizing their own biases (Pronin & Kugler, 2007; Pronin, Gilovich & Ross, 2004; Pronin, Yin & Ross, 2002). We all have difficulty identifying and correcting for bias because of a basic limitation of the human mental process: we have little insight into whether and how much our judgments have been influenced by particular facts or information to which we are exposed. We cannot rely on introspection to tell us whether—or how much—we have been influenced by any particular fact or factor. Hence, we cannot trust anyone's claim that a particular fact or factor had no influence on their judgment, at least not when the claim is based solely on introspection (Nisbett & Wilson, 1977; Wilson & Brekke, 1994)

But some forensic scientists have argued that members of their profession have the ability to avoid contextual bias simply through an act of will. One prominent forensic scientist who took this view was John Thornton, who declared:

I reject the insinuation that we do not have the wit or the intellectual capacity to deal with bias, of whatever sort. If we are unable to acknowledge and compensate for bias, we have no business in our profession to begin with, and certainly no legitimate plea to the indulgence of the legal system (Thornton, 2010).

Similar statements have been expressed by others (Mnookin et al., 2011). For example, in an open letter published in a professional journal, a Fellow and elected Chair of the Fingerprint Society said:

[A]ny fingerprint examiner who comes to a decision on identification and is swayed either way in that decision-making process under the influence of stories and gory images is either totally incapable of performing the noble tasks expected of him/her or is so immature that he/she should seek employment at Disneyland. . .(Leadbetter, 2007).

One goal of the research proposed here is to determine whether such attitudes are common. If forensic scientists have been slow to address the problem of contextual bias because they are overconfident in their own ability to avoid bias, then one promising way to encourage implementation of the blind or masking procedures is through educating forensic scientists about the psychological literature on contextual bias and on the introspection illusion. The research proposed here will assess the degree to which forensic scientists are familiar with literature on cognitive bias, and whether those more familiar with the literature are more likely to accept the need for corrective procedures.

5) Are forensic scientists confused about the task-relevance of contextual information?

Forensic scientists' slowness to address problems of contextual bias may also arise from uncertainty or confusion about what kinds of contextual information they should—and should not—consider when making their assessments. Commentators frequently propose that forensic scientists should avoid being influenced by information that is "domain-irrelevant" (e.g., Risinger et al. 2002) or "task-irrelevant," but they have not provided clear guidelines for distinguishing what is relevant and irrelevant. Nor has this issue been addressed in the forensic science literature.

Anecdotal evidence suggests that some forensic scientists are indeed confused or uncertain about what types of information they should rely on when forming scientific opinions. Consider, for example, the testimony of forensic odontologist David R. Senn in a June 12, 2012 deposition in the case of New York v. Dean. He explained that after linking a suspect's teeth to a bite mark:

If I then found that other evidence like the DNA swab that that was taken that had a positive amylase reaction came back as not excluding that same person, my confidence level would increase. I might be willing to upgrade my opinion from cannot exclude to probable….Now, many odontologists say you shouldn't have any awareness of the DNA results compared to the bite mark, and I agree that you shouldn't have them in advance, but if I subsequently get them, then I reserve the right to write a revised opinion. And I have done that. (p. 87)

The problem of assessing task relevance is complicated by the fact that forensic scientists might play multiple roles in an investigation. For example, a key task for a bloodstain pattern analyst is obviously to classify patterns—to determine, for example, whether a particular pattern of blood on a surface resulted from expiration or from impact. But the same analyst may also be involved with investigators in efforts to reconstruct events at the crime scene. This dual role complicates assessment of "task-relevance" because contextual information that is irrelevant to the pattern classification task (e.g., witness statements, medical reports) might be relevant and necessary to the reconstruction task. The dual role played by analyst may also complicate efforts to insulate analysts from potentially biasing contextual information. Through our interviews, we hope to explore and disentangle some of these issues, particularly as they arise in bloodstain pattern analysis.

Our interviews will explore forensic scientists' views about the task-relevance of contextual information. We will ask our informants to explain what types of contextual information they believe forensic scientists in various domains need to know (and why; and when) in order to do their jobs properly. We will seek to learn whether it is common or uncommon for forensic scientists to base interpretive conclusions on non-forensic evidence, or to base conclusions in one forensic domain on the results of tests in another domain (as Mr. Senn, quoted above, claimed to do). More generally, we will ascertain where analysts draw the line between *task-relevant* and *task-irrelevant* contextual information.

When developing CIM procedures, we will be guided in part by forensic scientists' assessments (as revealed by our interviews) of what they need to know to do a good job. But we will also be guided by normative assessments of the types of information that forensic scientists should and should not consider (e.g., Risinger, at al., 2002; Thompson, 2011; Page et al, 2011; Found & Ganas, 2013), including formal criteria for task-relevance recently developed by Thompson (2014). Using conditional probability models represented as Bayesian networks (see, Taroni, Aitken, Garbolino & Biedermann, 2006), Thompson (2014) offers mathematical criteria for assessing task-relevance and has demonstrated that these criteria maximize the usefulness of forensic science evidence to a trier-of-fact.

The normative models suggest that forensic scientists should base their conclusions solely on a scientific assessment of the evidence they are asked to evaluate. Forensic scientists who are influenced by evidence outside their specific scientific domain may undermine the accuracy of legal decision-making by usurping the function of the jury (Risinger, 2012). As Page et al. (2011, p. 108) explained, "if a forensic examiner reaches a conclusion that includes consideration of other factors other than the evidence before them, their conclusions should not carry the independent weight that the trier of fact has assumed is inherent in such testimony." Because the jury may not realize that the conclusion of the bite mark analyst are based partly on the DNA evidence, this evidence may be double-counted (Lempert, 1977; Schum & Martin, 1982; Thompson, 2011). Thus, forensic experts who stray beyond their domain may undermine the ability of the jury to make an accurate assessment of the case.

Through this research, we hope to provide to the forensic science community the most detailed and thoughtful guidelines yet proposed on how to draw the line between task-relevant and task-irrelevant contextual information, and on ways to ensure that analysts have full access to the former while avoiding the undue influence of the latter.

References

Budowle, B., Bottrell, M. C., Bunch, S. G., Fram, R., Harrison, D., Meagher, S., ... & Stacey, R. B. (2009). A Perspective on Errors, Bias, and Interpretation in the Forensic Sciences and Direction for Continuing Advancement. Journal of Forensic Sciences, 54(4), 798-809.

Butt, L. (2013). The forensic confirmation bias: problems, perspectives, and proposed solutions-Commentary by a forensic examiner. Journal of applied research in memory and cognition, 2(1), 59-60.

Charlton, D. (2013). Standards to avoid bias in fingerprint examination? Are such standards doomed to be based on fiscal expediency? Journal of applied research in memory and cognition, 2(1), 71-72.

Cole, S. A. (2013). Implementing counter-measures against confirmation bias in forensic science. Journal of Applied Research in Memory and Cognition, 2(1), 61-62.

Dresp-Langley, B., & Reeves, A. (2011). Simultaneous brightness and apparent depth from true colors on grey: Chevreul revisited. Seeing and perceiving, 25(6), 597-618.

Dror, I.E. (2012). Expectations, contextual information, and other cognitive influences in forensic laboratories. Sci.Justice 52(2), 132

Dror, I. E. (2011). The paradox of human expertise: Why experts get it wrong. In N. Kapur (Ed.) The Paradoxical Brain (pp. 177-188). Cambridge, UK: Cambridge University Press.

Dror, I. E., & Charlton, D. (2006). Why experts make errors. Journal of Forensic Iden- tification, 56, 600–616.

Dror, I. E., Charlton, D., & Peron, A. (2006). Contextual information renders experts vulnerable to making erroneous identifications. Forensic Science International, 156, 174–178. http://dx.doi.org/10.1016/j.forsciint.2005.10.017

Dror, I. E., & Hampikian, G. (2011). Subjectivity and bias in forensic DNA mixture interpretation. Science & Justice, 51(4), 204-208.

Dror, I., & Rosenthal, R. (2008). Meta-analytically Quantifying the Reliability and Biasability of Forensic Experts. Journal of Forensic Sciences, 53(4), 900-903.

Found, B., & Ganas, J. (2013). The management of domain irrelevant context information in forensic handwriting examination casework. Science & Justice, 53(2), 154-158.

Hagan, W. E. (1894). A Treatise on Disputed Handwriting and the Determination of Genuine from Forged Signatures: The Character and Composition of Inks, and Their Determination by Chemical Tests. The Effect of Age as Manifested in the Appearance of Written Instruments and Documents. Banks & Brothers.

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. Journal of Applied Research in Memory and Cognition, 2(1), 42-52.

Krane D.E., Ford S., Gilder J., Inman K., Jamieson A., Koppl R., et al. (2008). Sequential unmasking: a means of minimizing observer effects in forensic DNA interpretation. J Forensic Sci;53(4):1006–7.

Leadbetter M. (2007). Letter to the Editor. Fingerprint World. (Sept); 231.

Lempert, R. O. (1976). Modeling relevance. Mich. L. Rev., 75, 1021.

Miller, L. S. (1984). Bias among forensic document examiners: a need for procedural change. Journal of Police Science and Administration, 12(4), 407-411.
Mnookin JL, Cole SA, Dror IE, Fisher, BA, Houck, MM, et al. (2011). The need for a research culture in the forensic sciences. UCLA L Rev;58:725-79.
Murrie, D.C., Boccaccini, M.T., Guarnera, L.A., & Rufino, K.A. (2013). Are forensic experts biased by the side that retained them? Psychological science, *24*(10), 1889-1897.
National Research Council (2009) Strengthening Forensic Science in the United States: A Path Forward. Washington, D.C.: The National Academies Press.
Nisbett RE, Wilson TD. (1977). Telling more than we can know: verbal reports on mental processes. Psychol Rev;84:231–59.
Office of the Inspector General U.S. Department of Justice. (2006) *A Review of the FBI's Handling of the Brandon Mayfield Case*, Office of the Inspector General U.S. Department of Justice, Washington, DC, pp. 1–330.
Osborne, N. K., Woods, S., Kieser, J., & Zajac, R. (2014). Does contextual information bias bitemark comparisons?Science & Justice.
Ostrum, B. (2009). Commentary on: Authors' Response [J Forensic Sci 2009; 54 (2): 501] to Wells' comments [J Forensic Sci 2009; 54 (2): 500] regarding Krane DE, Ford S, Gilder JR, Inman K, Jamieson A, Koppl R, Kornfield IL, Risinger DM, Rudin N, Taylor MS, Thompson WC. Sequential unmasking: a means of minimizing observer effects in forensic DNA interpretation. Journal of forensic sciences, 54(6), 1498-1499.
Page, M, Taylor, J, Blenkin, M. (2011). Context effects and observer bias—implications for forensic odontology. J. Forensic Sci; doi: 10.1111/j.1556-4029.2011.01903.x Available online at: onlinelibrary.wiley.com
Pretty, I., & Sweet, D. (2010). A paradigm shift in the analysis of bitemarks. Forensic Science International, 201, 38-44.
Pronin E, Kugler MB.(2007). Valuing thoughts, ignoring behavior: the introspection illusion as a source of the bias blind spot. J Exp Soc Psychol; 43:565–78.
Pronin E, Gilovich T, Ross L. (2004). Objectivity in the eye of the beholder: divergent perceptions of bias in self versus others. Psychol Rev;111:781–99.
Pronin, E., Yin, D.Y., & Ross, L. (2002). The bias blind spot: perceptions of bias in self and others. Personality and Social Psychology Bulletin, 28, 369-381.
Pronin E. Perception and misperception of bias in human judgment. Trends Cogn Sci 2006;11:37–43.
Risinger DM, Saks MJ, Thompson WC, Rosenthal R (2002). The Daubert/Kumho implications of observer effects in forensic science: hidden problems of expectation and suggestion. Calif Law Rev; 90:1–55.
Saks MJ, Risinger DM, Rosenthal R, Thompson WC (2003). Context effects in forensic science: a review and application of the science of science to crime laboratory practice in the United States. Sci Justice; 43:77–90.
Schum, D. A., & Martin, A. W. (1982). Formal and empirical research on cascaded inference in jurisprudence. Law and Society Review, 105-151.
Sheldrake, R. (1999). How widely is blind assessment used in scientific research? Sciences (US), 93(22-3), 203.

Stacey RB. Report on the erroneous fingerprint individualization in the Madrid train bombing case, J Forensic Ident 2004;54(6):706–18.
Stoel, R. D., Dror, I. E., & Miller, L. S. (2014). Bias among forensic document examiners: Still a need for procedural changes. Australian Journal of Forensic Sciences, 46(1), 91-97.
Taroni, F., Aitken, C., Garbolino, P., & Biedermann, A. (2006). Bayesian Networks and Probabilistic Inference in Forensic Science. Front Matter (pp. i-xviii). John Wiley & Sons, Ltd.
Taylor, M., Laber, T., Kish, P. E., & Owens, G. (2014). Reliability Assessment of Current Methods in Bloodstain Pattern Analysis, Final Report for NIJ.
Thompson, W.C. (2008). Beyond bad apples: Analyzing the role of forensic science in wrongful convictions. Southwestern Law Review, 37(4), 1027-1050.
Thompson W.C., (2009a). Interpretation: Observer Effects, in *Wiley Encyclopedia of Forensic Science*, Jamieson, A., Moenssens, A. (eds). John Wiley & Sons Ltd., Chichester, UK, pp 1575-1579.
Thompson, W.C. (2009b). Painting the target around the matching profile: The Texas sharpshooter fallacy in forensic DNA interpretation. Law, Probability and Risk, 8, 257-276.
Thompson WC. (2011). What role should investigative facts play in the evaluation of scientific evidence. Aust J Forensic Sci;43 (2-3):123-34.
Thompson WC. Forensic DNA evidence: the myth of infallibility. In Krimsky S, Gruber J, editors. Genetic explanations: sense and nonsense. Cambridge, MA: Harvard University Press, 2013; 227-55.
Thompson, W.C. (2014). Modeling domain relevance: what facts should experts consider and ignore (Manuscript in preparation).
Thornton JI. (2010). Letter to the editor—a rejection of ‘‘working blind’’ as a cure for contextual bias. J Forensic Sci;55(6):1663
Wells, J. D. (2009). Commentary on: Krane DE, Ford S, Gilder JR, Inman K, Jamieson A, Koppl R, Kornfield IL, Risinger DM, Rudin N, Taylor MS, Thompson WC. Sequential unmasking: a means of minimizing observer effects in forensic DNA interpretation. J Forensic Sci 2008; 53 (4): 1006–7. Journal of forensic sciences, 54(2), 500-500.
Wilson TD, Brekke N. (1994). Mental contamination and mental correction: unwanted influences on judgments and evaluations. Psychol Bull;116:117–42.

frontiers in
GENETICS

OPINION ARTICLE
published: 28 October 2013
doi: 10.3389/fgene.2013.00220

The role of prior probability in forensic assessments

William C. Thompson[1], Joëlle Vuille[2]*, Alex Biedermann[3,4] and Franco Taroni[3]

[1] *Department of Criminology, Law and Society and School of Law, University of California, Irvine, Irvine, CA, USA*
[2] *Department of Criminology, Law and Society, University of California, Irvine, Irvine, CA, USA*
[3] *Faculty of Law and Criminal Justice, School of Criminal Justice, Institute of Forensic Science, University of Lausanne, Lausanne, Switzerland*
[4] *Department of Economics, Università Ca' Foscari Venezia, Venice, Italy*
**Correspondence: jvuille@uci.edu*

Edited by:
Qizhai Li, Chinese Academy of Sciences, China

Reviewed by:
Wenjun Xiong, Chinese Academy of Sciences, China

Keywords: DNA, Bayes Theorem, prior probability, expert testimony, forensic science

As the importance of forensic science in the legal system has grown, debate has arisen about the way forensic scientists should characterize their findings in order to communicate most effectively with legal fact-finders. This article will focus on one aspect of that debate: the framing of conclusions involving elements of probability. In particular, we will examine the contentious issue of whether forensic scientists, when asked to provide evidence that will be used to evaluate various competing propositions about physical evidence, should consider the prior probabilities that those propositions are true. Disputes about this issue have arisen in a number of contexts and recent examples suggest that opinions still diverge (e.g., Budowle et al., 2011; Biedermann et al., 2012). In this comment, we will argue that a reasoned approach to this issue depends on the role that forensic scientists are expected to play in the legal system.

To illustrate the underlying issues, let us begin with a generic example. A forensic scientist is asked to perform DNA profiling analyses of blood found at a crime scene and to compare the result to the DNA profile of a defendant who is charged with the crime. The defendant's guilt or innocence will be determined by a jury. The jurors' decision will depend in part on *their* assessment of two propositions of interest—H1: that the defendant was the source of the blood; and H2: that someone else was the source of the blood. What should the forensic scientist tell the jurors about the results of the DNA analysis?

The jurors might want the expert to tell them definitively which hypothesis is true, or to give them particular values for the so-called source probabilities—saying, for example, that there is a 0.998 probability the defendant is the source of the blood and only a probability of 0.002 that someone else was the source. But there is no way for the forensic scientist to reach such conclusions based on the forensic findings alone. To assess source probabilities, the forensic scientist must also consider other evidence in the case.

Suppose, for example, that the expert found that the defendant and the blood from the crime scene share a set of genetic markers found in one person in 1 million in the relevant population. Without considering other evidence in the case, the expert might make statements about the conditional probability of finding these results under the two hypotheses of interest. For example, the expert might conclude that the shared genetic markers were virtually certain to be found under H1 (defendant was the source), but had only 1 chance in 1 million of being found under H2 (someone else was the source). Based on this assessment the expert might also provide to the jury a so-called likelihood ratio—saying, for example, that the DNA profiling results are 1 million times more probable if the defendant rather than some other person was the source of the blood. But a likelihood ratio is not the same thing as a source probability. The likelihood ratio reflects the relative probability of the findings under the relevant propositions, not the probability that the propositions are true.

The only coherent way to draw conclusions about source probabilities on the basis of forensic evidence is to apply Bayes' rule, which requires that one begins with an assignment of prior probabilities to the propositions of interest (e.g., Robertson and Vignaux, 1995; Finkelstein and Fairley, 1970). Bayes' rule specifies how one ought to combine prior probabilities with the results of a DNA profiling analysis in order to find the so-called posterior probabilities that the defendant is the source of the blood. But the Bayesian approach will only work if the expert can begin with a prior probability.

This brings us to the crux of the debate: whether forensic scientists should even try to specify prior probabilities and, if so, how. It is occasionally suggested that forensic scientists should *assume* equal prior probabilities. This is sometimes described as a position of neutrality and is often justified with references to vague accessory "principles," such as the "Principle of Indifference" or the "Principle of Maximum Entropy," borrowed from other disciplines and contexts (Biedermann et al., 2007).

A prominent illustration can be found in paternity cases. When DNA analysts are asked to assist in the assessment of whether a particular man is the father of a child, they usually analyze the profiles of the mother, child, and the accused man, and assign conditional probabilities that the genetic characteristics found in the child (Ec) would be observed under two relevant hypotheses specifying that the accused is the father (H1) and that some other man (from a particular reference population) is the father (H2) conditioned on the alleged parents' DNA profiles (Em and Eam, for the mother and the accused man, respectively). In some cases, the analysts

limit themselves to reporting the ratio of these conditional probabilities—i.e., Pr(Ec|Em,Eam,H1)/Pr(Ec|Em,H2)—which is a likelihood ratio (although it is also referred to as the paternity index). But quite often, analysts go farther. They assume that the prior odds of H1 and H2 are equal and then, in accordance with Bayes' rule, they multiply the prior odds by the likelihood ratio (paternity index) to determine the posterior odds of paternity. Recall that odds are defined as a ratio between two probabilities; in this particular scenario, it is the ratio between Pr(H1) and Pr(H2). The posterior odds are typically restated as a probability. For example, if the DNA evidence supports paternity with a likelihood ratio of 1 million some analysts would report a probability of 0.999999 that the accused is the father.

While this approach is commonly used in civil paternity cases, courts in the United States have generally not allowed analysts to characterize their findings in this manner when paternity tests are offered as evidence in criminal cases—e.g., to prove the defendant committed rape or incest by showing he fathered a particular child. The assumption of equal prior odds appears to conflict with the presumption of innocence to which defendants in criminal trials have traditionally been entitled. In the view of most commentators, assuming that the accused starts with a probability of guilt of 0.5 falls far short of presuming him innocent. More fundamentally, making *any* default assumption about the prior probability is seen as violating the obligation of the legal system to deliver individualized justice based on the facts of each case (the attentive reader might have noted that circumstantial information *I* was omitted from the above mathematical notation). Consider that an assumption of equal priors is applied regardless of any other evidence in the case: an accused man who offers proof that he is infertile due to azoospermia and was not on the same planet as the mother at time of conception (i.e. an azoospermic cosmonaut) is treated the same as any other man. While the jury can take the other evidence into account they may have difficulty integrating it with the "probability of paternity" delivered by the forensic expert, or they may mistakenly assume that the "probability of paternity" is all they need consider.

Another suggested approach is that forensic scientists take upon themselves the responsibility for assessing the prior probability of the relevant hypotheses before updating them based on the scientific findings in accordance with Bayes' rule. For example, in the context of missing person identification, commentators declared that "[t]he forensic DNA community needs to develop guidelines for objectively computing prior odds" (Budowle et al., 2011, p. 15). The major objection to this approach, in the context of a criminal trial, is that it may result in forensic scientists going beyond their scientific expertise and usurping the role of the fact-finder. In order to assign prior contextually meaningful probabilities, the expert would need to take into account all of the evidence in the case. But experts are rarely in a good position to evaluate the non-scientific evidence and have no business doing so. The legal system places the responsibility for evaluating the evidence in a case on the fact-finder, whether judge or jury, not the expert witness. Jurors are carefully chosen for the task, are often shielded by evidentiary rules from information that the legal system determines that they should not consider, and are carefully instructed on the presumptions to make and standards to apply in reaching a verdict; experts are not. Allowing expert witnesses to take into account prior odds when considering the probative value of a scientific observation also raises the danger of double-counting certain pieces of evidence (Thompson, 2011).

Consequently, many commentators have suggested that forensic experts have no role in assessing prior probabilities. Because posterior probabilities can only be arrived at by assessing prior probabilities, they argue that experts cannot legitimately make statements about posterior probabilities either. As Redmayne explains (2001, p. 46): "(. . .) *the expert should not testify in terms such as (. . .) 'the blood probably came from the defendant', because one can only reach conclusions of this sort by making assumptions about the strength of other evidence against the defendant.*"

There may, however, be circumstances in which a forensic scientist could appropriately assign prior probabilities and use them as a basis for reaching other conclusions. One such circumstance arises when the expert is given the responsibility of making an overall evaluation of a case. For example, coroners are sometimes given full responsibility for determining the cause and manner of a death for legal purpose. (In jurisdictions of the Anglo-Saxon tradition, a coroner is a government official who investigates human deaths and makes independent determinations as to their time, manner, and cause. He should not be confused with the medical examiner, who merely provides information to a court in the course of criminal prosecution or civil litigation but has no judicial authority of his own). In such cases, the expert should certainly take account of all relevant evidence, including both scientific and non-scientific factors. There is no danger of the expert usurping the fact-finder when the expert *is* the fact-finder. The matter becomes more complicated, however, when an expert who has made a determination in the role of fact-finder is subsequently asked to present evidence to another fact-finder, as when a coroner who has determined that a death was due to homicide rather than suicide in an inquest is asked to testify in a subsequent criminal trial. In such cases, the dangers of usurpation and double-counting of evidence discussed above may still loom large.

Whether forensic scientists should take account of the prior probability of the hypotheses they are asked to help evaluate is a complicated question. The answer depends on the role the forensic scientist will be playing in the legal system. If forensic scientists will make the ultimate determination, for legal purposes, with regard to a particular proposition of interest, then they should, and indeed must, consider their prior probabilities that the hypotheses are true. If, however, the truth of the hypotheses will be addressed by someone else—e.g., a judge or jury—and the forensic scientists' role is limited to providing expert assistance, then forensic scientists should generally confine themselves to assign the conditional probability of the scientific findings under the given hypotheses of interest, and should leave to the legal decision maker the task of assessing prior and posterior probabilities.

ACKNOWLEDGMENTS

William C. Thompson was supported by the UC Lab Fees Research Program. Joëlle Vuille was supported by the Swiss National Science Foundation (grants PBLAP1-136958, PBLAP1-145850). Alex Biedermann was supported by a Research Mobility Grant of the Société Académique Vaudoise.

REFERENCES

Biedermann, A., Taroni, F., and Garbolino, P. (2007). Equal prior probabilities: can one do any better? *Forensic Sci. Int.* 172, 85–93. doi: 10.1016/j.forsciint.2006.12.008

Biedermann, A., Taroni, F., and Margot, P. (2012). Reply to Budowle, Ge, Chakraborty and Gill-King: use of prior odds for missing persons identifications. *Investig. Genet.* 3, 1–7. doi: 10.1186/2041-2223-3-2

Budowle, B., Ge, J., Chakraborty, R., and Gill-King, H. (2011). Use of prior odds for missing persons identifications. *Investig. Genet.* 2, 15. doi: 10.1186/2041-2223-2-15

Finkelstein, M. O., and Fairley, W. B. (1970). A bayesian approach to identification evidence. *Harv. Law Rev.* 83, 489–517.

Redmayne, M. (2001). *Expert Evidence and Criminal Justice*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198267805.001.0001

Robertson, B., and Vignaux, G. A. (1995). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. Chichester: John Wiley & Sons.

Thompson, W. C. (2011). What role should investigative facts play in the evaluation of scientific evidence? *Aust. J. Forensic Sci.* 43, 123–134.

Received: 28 August 2013; accepted: 08 October 2013; published online: 28 October 2013.

*Citation: Thompson WC, Vuille J, Biedermann A and Taroni F (2013) The role of prior probability in forensic assessments. Front. Genet. **4**:220. doi: 10.3389/fgene.2013.00220*

This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics.

DEBORAH TUERKHEIMER

Deborah Tuerkheimer is a Professor of Law at Northwestern University School of Law. From 2009 to 2014 she was Professor of Law at DePaul University College of Law. She earned her undergraduate degree from Harvard College and her law degree from Yale. She teaches and writes in the areas of criminal law, evidence, and feminist legal theory.

Professor Tuerkheimer is the author *Flawed Convictions: "Shaken Baby Syndrome" and the Inertia of Injustice,* (Oxford University Press, 2014). She is also a co-author of the casebook, *Feminist Jurisprudence: Cases and Materials*, and has authored numerous articles on rape and domestic violence.

After clerking for Alaska Supreme Court Justice Jay Rabinowitz, she served for five years as an Assistant District Attorney in the New York County District Attorney's Office, where she specialized in domestic violence and child abuse prosecution.

5 MISSED DIAGNOSIS

In late 2005, Melonie Ware, the mother of two young children, was convicted of murder and sentenced to life in prison. Based on the triad, the state alleged that the Georgia caregiver had shaken to death an eight-month-old boy. Six prosecution doctors dismissed the possibility that that baby, Jaden, had suddenly and inexplicably collapsed during a feeding, as Ware described. The experts were certain that the medical evidence proved her guilt. Even the defense conceded the correctness of the SBS diagnosis, arguing that Jaden had been shaken before he was left in Ware's care. But the jury was convinced by the state's doctors' professed ability to precisely time the baby's injury: it must have been inflicted during the period when he was with his caregiver.[1] The trial was standard fare for these prosecutions.

Shortly after Ware was sentenced, she hired a different lawyer, who promptly moved for a new trial based on the ineffective assistance of trial counsel. After first granting Ware an evidentiary hearing, the same judge who had just sentenced her to a life term made the remarkable decision to set aside the conviction.[2] Ware had been so greatly disadvantaged by her lawyer's subpar performance that, but for it, according to the trial judge, "a reasonable possibility, even a strong possibility, exist[ed]" that she would have been found not guilty.[3] This assessment proved to be correct. At her subsequent retrial, Ware was acquitted. She had served one year in jail on a life sentence imposed wrongly. Though financially wrecked, the family was reunited.[4]

What saved Ware was the discovery that Jaden had apparently died of complications from sickle cell disease. Seven experts, including a pathologist, a pediatric neurologist, an ophthalmologist, a pediatrician, and a hematologist/oncologist (blood cancer specialist) testified to this effect on behalf of the defense. The evidence that sickle cell disease caused Jaden's death was overwhelming. Blood test results were consistent with disseminated intravascular coagulation, which occurs when blood cells meet blood clots that

rip apart the cells. Jaden's blood culture revealed signs of bacterial infection consistent with sickle cell disease. And autopsy slides revealed the presence of sickle cells in the eyes and parts of the brain.[5]

Early in his life, Jaden was actually diagnosed with sickle cell disease; the fact that he suffered from this genetic blood disorder was not in dispute.[6] Even so, Jaden's condition was barely mentioned at Ware's first trial. Rather than pursue a "medical defense," Ware's lawyer decided not to question the SBS diagnosis—he would keep the case "so simple and straightforward," as he later put it, that the jury would be able to follow.[7] By pointing the finger at an unknown abuser, the lawyer hoped to conjure the conventional "whodunit" archetype of criminal investigations. But, as Ware's conviction after her first trial suggests, this model is inapt when no crime whatsoever has taken place. As a substitute for advancing a causal account of a natural disease process—a medical defense, as it were—the "other abuser" strategy is counterproductive.

The decision to forego an attack on the doctors' SBS diagnosis might well have stemmed from the trial lawyer's limited understanding of the relevant scientific research. "I don't think anybody would feel comfortable with the science in this field unless they were a medical doctor or had one sitting next to them," he admitted.[8] An SBS defendant whose lawyer feels daunted by the science surrounding SBS is at a pronounced disadvantage. This is a problem that reaches far beyond Melonie Ware's case. Although a defense attorney's failure to engage the science is not necessarily viewed as ineffective for purposes of constitutional analysis, it still falls short of adequate representation. Yet, for many lawyers, the task of mastering a body of complicated scientific research is overwhelming. Even those who have successfully defended SBS charges remark on the toll it exacts. Articulating a widely shared sentiment, Ware's post-conviction counsel noted, "defending an individual with charges related to Shaken Baby Syndrome is one of the most difficult cases that a criminal defense attorney will face in his or her career."[9]

At Ware's first trial, her lawyer's failure was sufficiently egregious to require a remedy.[10] Not presenting the "critical, material, available medical evidence" amounted to what the post-conviction counsel aptly referred to as a "breakdown in the adversary system."[11] Across this system, equally egregious failures may never come to the attention of a judge, and failures less egregious may not entitle a defendant to relief. But for Ware, a court was prompted to undo the injustice. The case epitomizes the importance of decent defense representation and what that entails in SBS cases.

What happened to Melonie Ware also raises important questions about the functioning of prosecutors and prosecution experts. Consider that evidence

of Jaden's sickle cell disease was not newly "discovered"—rather, it was known at the time of the first trial. Prosecutors were well aware that the baby suffered from this condition; they were even on notice that he had recently been hospitalized for engorgement of the spleen as a result of his disease.[12] The state's response to this evidence was to deny its relevance altogether.

At Ware's first trial, Doctor Beatrice F., a pediatric hematologist, testified that Jaden's death could not have resulted from his sickle cell disease. Doctor F. summarily explained that she was responsible for transfusing Jaden's blood in a prior hospitalization, and that this transfusion eliminated the possibility that complications from sickle cell were the cause of the baby's death.[13] Defense counsel neglected to cross-examine Doctor F. regarding a potential conflict of interest arising from the prior treatment of Jaden and how this might bias her testimony.

The problems with Doctor F.'s testimony became apparent at Ware's second trial. Doctor Michael D. is a pediatric hematologist/oncologist and director of the Sickle Cell Treatment and Educational Center at a leading medical school and a metropolitan children's hospital specializing in the treatment of sickle cell disease. Never having appeared as an expert witness in a criminal case, he testified on Ware's behalf at her retrial. According to Doctor D., there was no evidence that transfusion treatment is effective to prevent engorgement of the spleen, and there are significant risks associated with the particular transfusions that Jaden received, including just the kind of brain injury presented in this case.[14]

Doctor D. suggested that decisions made in the course of caring for Jaden prior to his collapse might have contributed to his death. The child had apparently received too much blood, causing a thickening reaction that can lead to clotting in the brain. There was powerful evidence that this "hyperviscosity syndrome" was a factor in Jaden's collapse: three transfusions had occurred within a four-week period; on the final transfusion, the amount of blood provided doubled; Jaden's CT scans were consistent with bleeding over the course of weeks; and, finally, the autopsy report indicated the presence of thrombosis (clots). In Doctor D.'s estimation, complications from sickle cell disease, or from the treatment of the disease, or from a combination of the two, led to the baby's death.[15]

This opinion cast new light on the testimony of Doctor Beatrice F., again, one of the doctors responsible for Jaden's prior transfusion treatment. In retrospect, it is clear that defense counsel at Ware's first trial should have pursued the argument that sickle cell was causal in Jaden's death. Even at the time, we might have expected the prosecutor to realize that Doctor F. was

potentially conflicted by her role in Jaden's earlier care. But a belief in SBS can stifle critical analysis. In Ware's case, the state was utterly convinced of Ware's guilt and remained so: even after listening to the (unrebutted) testimony of Ware's experts at the hearing on her new trial motion; even after receiving the judge's decision to vacate the conviction based on the "strong probability" that—with the benefit of a fuller evidentiary record—Ware would be acquitted. As evidenced by its decision to retry her, the state still could not see reasonable doubt.

Ware's case places in stark relief the limitations of "differential diagnosis." Less than a week before the alleged shaking incident, Jaden had received a massive blood transfusion in relation to his sickle cell disease. Yet his condition was ruled out as a factor; the presence of a simple medical explanation for Jaden's death did not in any way diminish the confidence of the experts. Their certainty is troubling. It led them to ignore the presence of sickle cells on autopsy slides and to rationalize a problematic blood workup. It kept them from asking difficult questions about Jaden's transfusions and from a truly searching inquiry into why treatment for his sickle cell disease failed.

The state's experts were unanimous that the triad proved Ware's guilt. Doctors can, of course, be convinced and be wrong. But this phenomenon is especially jarring when the presumption of abuse generated by the triad obscures an obvious alternative explanation for neurological decline.

Diagnostic Error

Incorrect medical beliefs are sticky, meaning they endure.[16] Even after a research error is uncovered, its influence can linger.[17] According to Doctor John Ioannidis, a leading expert on the subject, the appropriate response is to invert our collective expectations of science: we ought to stop perceiving the claims of medicine as akin to truth. Physicians in particular should acknowledge the limits of the research endeavor.[18]

But, when it comes to the triad, no such inversion has occurred; expressions of certainty as to its diagnostic significance persist. One consequence is that non-abusive origins of the triad have mostly been a secondary research concern as compared to efforts to establish a connection between the triad and abuse. Even so, understandings of abuse "mimics" have advanced. It might then be supposed that better individual diagnosis would result. Instead, clinical practice has lagged behind research on causes of the triad other than shaking.

The association between quality scientific research and accurate diagnosis is a basic postulate of evidence-based medicine.[19] In theory, there is good reason to believe that this relationship exists. But the reality of clinical diagnosis is more convoluted. Doctors generally struggle to translate the best available scientific knowledge into practice, often reaching conclusions akin to "educated guesses."[20]

In discussing the distinctive epistemology of modern medicine, legal scholar Lars Noah explains that clinicians are disinclined to rely upon the latest scientific research—particularly when it conflicts with previously accepted lore.[21] This is true even when the latest research improves on earlier studies, since physicians treating patients tend not to engage with questions of methodology or research quality.[22] As Noah observes, "we most certainly do not enjoy evidence-based medical practice."[23]

This description of clinical decision making bears directly on SBS. Noah's account hints at why, despite having been universally disavowed by researchers, the pathognomonic triad persists in clinical practice to this day. It suggests how the medical profession might have uncritically perpetuated—through widespread diagnosis—an unproven hypothesis declaring that the triad, without any other signs, could result only from forceful shaking.

But if the role of science in medical diagnosis is more limited than we might like to believe, what *does* affect the "flesh and blood" decision making of physicians?[24] In the past decade, researchers have begun to apply the insights of behavioral science to better understand diagnostic error.[25] We now know that, like everyone else, physicians take shortcuts when processing complex information.[26] Unfortunately, mistakes in diagnosis often result. In the SBS context, these misdiagnoses help to sustain the triad's lasting sway.

Researchers have made great progress in understanding the impact of "cognitive dispositions to respond," which may be defined as "cognitive errors, especially those associated with failures in perception, failed heuristics, and biases."[27] Default to cognitive dispositions to respond is especially likely under certain conditions, such as those encountered by doctors in urgent medical situations.[28] As one literature review noted, "cognitive diagnostic failure is inevitable when the exigencies of the clinical workplace do not allow . . . Olympian cerebral approaches."[29]

What we have learned about cognitive errors in medicine allows for greater insight into the origins of misdiagnoses. A 2005 study of one hundred diagnostic errors provides a helpful taxonomy of causal factors.[30] The study's findings are directly relevant to the diagnosis of SBS. First, faulty processing

of the available information is the most common source of cognitive error, followed by faulty data gathering.[31] Analyzing the relationships between various error sources, researchers were able to identify "clusters of cognitive factors that tended to co-occur."[32] As one illustration, "a mistake relatively early on (e.g., an inadequate history or physical examination) is likely to lead to subsequent mistakes (e.g., in interpreting test results, considering appropriate candidate diagnoses, or calling in appropriate specialists)." Researchers concluded that "diagnostic error is typically multifactorial in origin."[33]

Even so, one error stood out as a primary contributor to misdiagnosis: "[t]he single most common phenomenon was premature closure: the tendency to stop considering other possibilities after reaching a diagnosis."[34] This cognitive process has also been called "satisficing," and it is well documented.[35] Regardless of the term used, this tendency toward too quickly foreclosing alternatives sits uneasily with an idealized notion of differential diagnosis. In SBS cases, the problem is especially salient, since early diagnosis is the rule.

As we will see, premature closure is just one of many cognitive strategies relevant to SBS. The typical progression of a triad-only diagnosis follows a set pattern. A baby presents with acute neurological symptoms and no external signs of abuse. Emergency room doctors promptly discover a subdural hematoma, which triggers a deeply held belief that shaking was causal. An ophthalmologist is called to validate the diagnosis; the ophthalmologist finds retinal bleeding. A child abuse expert intervenes early on and further confirms what the others already strongly suspected: the caregiver with the child at the time of collapse has not provided a satisfactory explanation for the bleeding.

Doctors now believe they know that this person shook the baby. Child protective services and the police are called. Perhaps doctors will make efforts to "rule out" the alternatives perceived as possible. But the presumed diagnosis from the outset is shaking and this intuition is hardly ever disturbed. The trajectory of the diagnostic process seems almost to guarantee that intuition will harden into conviction. In short, when a baby presents with one or more triad symptoms, physicians—emergency room doctors, pediatricians, radiologists, ophthalmologists, and, on autopsy, forensic pathologists—tend to default to SBS.

To understand why, it is helpful to further contemplate the cognitive shortcuts that readily present themselves in this diagnostic setting and to observe how these shortcuts may lead to misdiagnosis. Psychologists have identified over thirty major cognitive dispositions to respond that contribute to error (some are overlapping).[36] The descriptions of these dispositions are striking for their parallels to the conditions that give rise to an SBS diagnosis.

Consider first, errors associated with the initial diagnostic hypothesis. When a baby (who will later be diagnosed with SBS) presents in the emergency room with acute neurological symptoms, the first notable feature discovered is normally subdural bleeding. This raises the prospect of "anchoring," which psychologists describe as a "tendency to perpetually lock onto salient features in the patient's initial presentation too early in the diagnostic process."[37] For the doctor attending the child, the discovery of subdural bleeding prompts an immediate hypothesis: abuse.

Because the triad of symptoms (including subdural bleeding) was once thought to be exclusively diagnostic of shaking, it is not surprising that doctors—most of whom have past experience, some extensive, with these cases—would reflexively associate subdurals and abuse. This dynamic implicates a bias known as "availability," which is "the disposition to judge things as being more likely, or frequently occurring, if they readily come to mind."[38] Whether accurate or not, the past diagnosis of subdural bleeding as SBS may "inflate the likelihood"[39] that subdural bleeding will again be diagnosed as SBS.

In questioning the adults who were caring for the child, the focus remains on the time period closely preceding the baby's collapse, since the likelihood of a lucid interval is considered slim to none. This tendency implicates the "unpacking principle," which warns that a doctor's "failure to elicit all relevant information (unpacking) in establishing a differential diagnosis may result in significant possibilities being missed."[40] The suspect is fast identified, consistent with a diagnostic strategy of "going for the obvious" while giving other possibilities short shrift.[41]

Since SBS is now a diagnosis of exclusion, doctors are supposed to rule out alternatives. Thus begins the process of "differential diagnosis," which tends to implicate a number of common sources of diagnostic error. The explanation for the triad with which medical professionals are most accustomed is SBS. Other possibilities are less defined; they may be complicated, uncertain, and far less attractive than resorting to shaking. This scenario raises the problem of "multiple alternative bias," where doctors "simplify" the diagnostic process "by reverting to a smaller subset" that is "familiar."[42] The result is "inadequate consideration" of other alternatives.[43]

In the typical scenario, a child abuse specialist, if not yet involved, is consulted and becomes responsible for subsequent management of the case. An ophthalmologist is asked to look for retinal hemorrhages. A radiologist is sought for a more expert opinion on a CT scan. Blood is often sent to the lab to test for disease or disorder, though bleeding specialists are not ordinarily called upon for input into what is causing blood on the brain. All of these

decisions suggest the possibility of "triage cuing," which "results in patients being sent in particular directions, [cuing] their subsequent management."[44] The choice of which specialist to consult (or not) may well dictate the ultimate diagnosis—hence the adage that "geography is destiny."[45]

Doctors commonly find no alternative to shaking. In many cases, the dynamics in place suggest the operation of a powerful confirmation bias, which psychologists explain as a "tendency to look for confirming evidence to support a diagnosis rather than look for disconfirming evidence to refute it, despite the latter often being more persuasive and definitive."[46] Put differently, "people selectively focus upon evidence that supports their beliefs or what they want to believe to be true, while ignoring evidence that serves to disconfirm those ideas."[47]

In the clinical context, confirmation bias is a main source of diagnostic error.[48] The doctors' maxim that expresses this phenomenon is "you see what you look for, and you look for what you know."[49] When multiple physicians are involved in a case, as happens with SBS, it is common for each doctor to "verbally confirm a diagnosis or reinforce the initial diagnostic impression," regardless of its accuracy.[50] (This same potential for bias inheres in the autopsy, since pathologists are told what the baby's physicians concluded.) The drive to confirm the original intuition may constrain the workings of the differential diagnosis: the alternatives it includes; the specialists consulted; the tests required; the significance attached to test results; the willingness to withhold judgment for a time; and whether uncertainty is deemed tolerable, or even acknowledged. In a "conspiracy of concurrence" known as groupthink, doctors may avoid voicing dissent.[51]

Too soon, the diagnostic process is cut short. This "premature closure…account[s] for a high proportion of missed diagnoses."[52] Doctors are quickly certain: this is SBS. With few exceptions, once the investigation has concluded, doctors have little impetus to revisit the causal question.[53] In a phenomenon known as "feedback sanction," diagnostic errors are compounded when they go undetected, or when the passage of time blunts the impact of their discovery.[54] The problem of "sunk costs" predicts that clinicians will cling to a diagnosis to preserve investments in "time and energy and, for some, ego."[55]

From start to finish, the prototypical process by which a triad of symptoms comes to be labeled SBS is vulnerable to biasing mechanisms. Of course, the recognition of doctor error as a factor contributing to the stream of SBS diagnoses can coexist with recognition of abuse as one potential cause of head trauma. However, the effects of cognitive dispositions to respond raises

real questions about the rigor of differential diagnosis in this realm, and thus about the reliability of the conclusion so often reached.

A Legal Perspective on Differential Diagnosis

In SBS cases, judges have been willing simply to accept doctors' professed reliance on differential diagnosis. Expert testimony that purports to rest on a proven methodology is thus admitted without analysis. This deferential treatment is problematic, as is a complete judicial failure to probe whether differential diagnosis is an apt justification for the claim of external causation upon which SBS rests. In fact, it is not: reference to differential diagnosis not only obscures, but mischaracterizes, the methodological questions central to the conclusion that a baby was shaken.[56]

Though the faulty description of SBS as a differential diagnosis has been uncritically adopted by criminal courts, this reflexive judicial stance is not inevitable. We see the alternative in the civil realm, where a far more sophisticated framework for evaluating experts' methodological claims has developed. Most important, courts have been able to distinguish between the methodologies of "differential diagnosis" and "differential etiology."[57] The difference is more than semantic, as it points to a key substantive difference between the two processes: the former suggests a measure of validity, while the latter does not.

For one elaboration of this concept, consider a 2007 tort action called *Bowers v. Norfolk Southern Corporation*, which involved injuries sustained by a train operator who then sued the train company.[58] The plaintiff offered the testimony of Doctor Arthur Wardell, an orthopedist who opined, using a method of "differential diagnosis," that the injuries at issue were caused by the vibrations of the locomotive. But the federal district court did not just adopt the doctor's characterization of this methodology. Instead, upon examination, the court observed, "Dr. Wardell did not perform a 'differential diagnosis' on Plaintiff."[59]

The problem was not (or not primarily) related to technical inadequacies in the diagnostic approach. Rather, a proper understanding of differential diagnosis placed the doctor's mode of reasoning outside its ambit. To reach this conclusion, the court relied on two medical dictionary definitions. According to the first, differential diagnosis is "the determination of which one of two or more diseases or conditions a patient is suffering from, by systemically comparing and contrasting their clinical findings."[60] Per the second definition, differential diagnosis is "the determination of which of two or

more diseases with similar symptoms is the one from which the patient is suffering, by a systematic comparison and contrast of the clinical findings."[61] Both meanings demarcate the bounds of differential diagnosis: as a rule, it does not provide an adequate basis for establishing external causation. As the *Bowers* court emphasized, differential diagnosis "focus[es] on diagnosing the disease, not on determining the etiology or cause of the disease."[62]

According to the court—which cited for support the Federal Judicial Center's Reference Manual on Scientific Evidence—differential etiology, by contrast, involves "the investigation and reasoning that leads to the determination of external causation . . . by a process of elimination."[63] This process demands the tools of science and justifies its conclusions by reference to an adequate evidentiary basis. This level of rigor was missing from Doctor Wardell's methodology, which resulted in identification of the locomotive seat's vibrations as the cause of the plaintiff's neck and back injuries. Differential diagnosis could perhaps establish degenerative disk disease as the likely cause of the plaintiff's pain, but that would be the extent of it.

The district court refused to rubber stamp the methodological label of "differential diagnosis," excluding Doctor Wardell's opinion from evidence.[64] The U.S. Court of Appeals for the Eleventh Circuit affirmed, noting that the lower court's analysis was "both thorough and careful."[65] Confronted with differential etiology in the guise of differential diagnosis, other courts in civil cases have reached similar results.[66]

Because it purports to locate an external source of a patient's medical condition, differential etiology implicates validation concerns that are not typically raised by differential diagnosis. As the *Bowers* court emphasized, "the differential diagnosis method has an inherent reliability; the differential etiology method does not."[67] In general, doctors have particular motivations to diagnose accurately. If a doctor misdiagnoses a patient's *condition*, serious health consequences, including death, may follow. Misdiagnoses can also lead to medical malpractice suits.[68] But mistakes regarding external causation are unlikely to result in either of these outcomes. (The *Bowers* court offered these explanations for the reliability differential between the two methodologies.)

There are other reasons to treat skeptically conclusions derived from differential etiology, particularly in the medical context.[69] Rejecting an effort to characterize differential etiology as differential diagnosis, one court pointed to physician expertise as the key factor.[70] It was wrong to "conflate[] a doctor's expertise in diagnosis with a doctor's expertise in etiology," noted the court, adding, "[m]ost treating physicians have more training in and experience with diagnosis than etiology."[71]

Even when physicians do think about etiology in the clinical setting, they do so in ways that tend to undermine arguments for admissibility.[72] Causation in clinical practice reflects what one court described as a "precautionary principle." By way of explanation, the court offered that, "[i]f a particular factor *might* cause a disease, and the factor is readily avoidable, why not advise the patient to avoid it? Such advice—telling a welder, say, to use a respirator—can do little harm, and might do a lot of good."[73] This precautionary principle, however appropriate for the clinician, far from assures the reliability required of expert testimony. The "low threshold for making a decision serves well in the clinic but not in the courtroom," warned the court, "where decision requires not just an educated hunch but at least a preponderance of the evidence."[74] (In civil cases, the burden of proof is much lower than in the criminal context; and still expert testimony on causation is often excluded.)

For good reasons, then, when a doctor's methodology is accurately classified as differential etiology, many courts in civil cases have been extra cautious about admitting the expert opinion.[75] This perspective on causation bears directly on SBS, which is a diagnosis of external causation: shaking (or abuse of some sort) caused these symptoms.[76] Whatever its ostensible label, a diagnosis of causation raises the very concerns underlying judicial reluctance to admit opinions based on differential etiology.[77] In civil cases, the presumption of reliability that attends *true* differential diagnosis (when performed adequately) is often suspended when external causation is at issue. Thus far, this same judicial attitude has not applied when the triad leads doctors to conclude that a baby was shaken—in effect, that the baby's condition was caused by a specific human act. But insisting that SBS entails "differential diagnosis" does not make it reliable.

The etiological foundations of SBS warrant greater scrutiny of expert claims, regardless of whether they are styled as differential diagnosis. In civil cases—even those that stay within the framework of "differential diagnosis"—this cautious orientation is reflected by rigorous judicial analysis of opinions on causation.[78] A developed, albeit inconsistent, body of law limits the admissibility of testimony when it purports to identify a source of injury.[79] In contrast, criminal courts afford almost total deference to these same opinions when advanced by prosecution experts.[80] Given that testimony regarding the triad is functionally the same in kind as the expert opinions routinely excluded in civil cases (where, again, the burden of proof is lower), the automatic admission of this testimony in SBS cases is striking.[81]

There are multiple barriers to the admission of causation testimony in civil cases. Here, courts demand proof of both general causation and specific causation[82]—what, in the parlance of differential diagnosis, might be called "rule in" and "rule out." By distinguishing between general causation and specific causation and insisting that each be demonstrated, courts have developed a rather sophisticated framework for assessing the adequacy of a proffered expert opinion.

Consider one court's overview:

> The process of differential diagnosis is undoubtedly important to the question of "specific causation." If other possible causes of an injury cannot be ruled out, or at least the possibility of their contribution to causation minimized, then the "more likely than not" threshold for proving causation may not have been met. But, it is also important to recognize that a fundamental assumption underlying this method is that the final, suspected "cause" remaining after this process of elimination must actually be capable of causing the injury. That is, the expert must "rule in" the suspected cause as well as "rule out" other possible causes. And, of course, expert opinion on this issue of "general causation" must be derived from scientifically valid methodology.[83]

On the need to first "rule in" in order to satisfy admissibility standards, expert testimony must establish that the cause in question *could* contribute to the result in question. To do this, judges have recognized the importance of identifying the specific casual mechanism at issue. As one court observed, "the underlying predicates of any cause-and-effect medical testimony are that medical science understands the physiological process by which a particular disease or syndrome develops."[84] In this regard, both clinical experience[85] and case reports[86] may be insufficient to demonstrate the causal mechanism with the requisite precision.[87] Where general causation has not been adequately proven, courts have excluded expert testimony that purports to rest on a process of differential diagnosis.[88]

In SBS cases, lurking questions involving general causation have gone unaddressed by courts. Even discounting the biomechanical research that challenges the validity of a triad-only diagnosis,[89] prosecution experts admittedly don't know *how* shaking causes the triad (physiologically), or whether impact or another "abusive" mechanism might instead be at issue. This level of knowledge falls short of what courts have generally required for admissibility in the civil realm.

In similar fashion, experts' inability to scientifically determine the force levels involved in SBS (or AHT) is somewhat analogous to problems of dosage arising in toxic tort actions.[90] Here, a high degree of precision is required before expert opinion on causation can be admitted. As one court remarked, "a fundamental tenet of toxicology is that the 'dose makes the poison' and that all chemical agents, including water, are harmful if consumed in large quantities, while even the most toxic substances are harmless in minute quantities.... Therefore, in determining whether plaintiffs' exposure to PCBs could have caused any illness that they have, it is necessary to establish the dose/response relationship between PCBs and those particular illnesses."[91]

Scientific understandings of the causal mechanism supposedly responsible for the triad have largely unraveled in recent years—hence the move from shaking-only to AHT. The new unknowns are reflected in the testimony of prosecution experts who posit an array of options to explain the triad: the baby was thrown on a bed, or banged on the floor, or shaken, or perhaps some combination. Testimony regarding the forces required to cause the triad is likewise unmoored from science. "A reasonable person would know that this kind of shaking could cause injury," which is a common refrain, is not a scientific standard. In civil litigation, it is unlikely that a court would allow this kind of expert testimony on causation. In SBS cases, courts have yet to fasten on deficiencies in evidence of general causation.

When juxtaposed with admissibility requirements in tort actions, proof of specific causation in SBS prosecutions is also weak. Like "ruling in," the "ruling out" aspect of expert testimony is subjected to a rather stringent analysis in the civil context, where large amounts of money are often at stake.[92] There are various facets to this more exacting review in civil actions.

First, courts are unwilling to defer to assertions that an expert eliminated reasonable possibilities other than the hypothesized cause. Whether the process of foreclosing alternatives was adequate is for a judge to decide, not a doctor.[93] According to one court, "[a]n expert who supplies nothing but a bottom line supplies nothing of value to the judicial process."[94] According to another court, "in evaluating the reliability of an opinion based on a differential diagnosis, courts look at the substance of the expert's analysis, rather than just the label."[95] Testimony is excluded if the quality of the diagnostic process is unacceptable.[96]

Courts are even willing to question whether the method supposedly at issue was relied upon at all.[97] As one court noted, "simply claiming that an expert used the 'differential diagnosis' method is not some incantation that opens the *Daubert* gate to allow an expert's opinions to be admitted at trial.

Indeed, it can easily amount to nothing more than medico-legal sophistry used in an attempt to avoid the Court's reliability analysis."[98] (*Daubert v. Merrell Dow Pharmaceuticals* is the U.S. Supreme Court case establishing the commonly accepted standard governing the admissibility of expert testimony.)

Before admitting an opinion based on a process of elimination, courts require a high level of methodological precision. An expert must "systemically and scientifically rul[e] out specific causes until a final, suspected cause remains."[99] Where an expert "does not explain how or why he ruled out" alternatives, his opinion may be excluded.[100] An "analysis" is needed to satisfy the reliability standard.[101] Courts will not merely accept an expert's summary conclusion—even one based on professional judgment[102]—that all causes but the cause left standing were ruled out.

In many cases, it turns out that experts' causal claims are based (explicitly or implicitly) on the chronology of events, often referred to as temporal order.[103] Judges in civil cases have been wary of this reasoning. These reservations are illustrated by a product liability action against the manufacturer of herbal weight loss supplements, Metabolife, for causing serious injuries to its users.[104] In its discussion of reliability, the court was careful to identify what is widely known as the *post hoc ergo propter hoc* fallacy:

> [P]roving a *temporal* relationship between taking Metabolife and the onset of symptoms does not establish a *causal* relationship. In other words, simply because a person takes drugs and then suffers an injury does not show causation. Drawing such a conclusion from temporal relationships leads to the blunder of the *post hoc ergo propter hoc* fallacy.
>
> The *post hoc ergo propter hoc* fallacy assumes causality from temporal sequence. It literally means, "after this, because of this." It is called a fallacy because it makes an assumption based on the false inference that a temporal relationship proves a causal relationship.[105]

Courts tend to guard against expert opinion that suffers from the *post hoc ergo propter hoc* fallacy, particularly when medical science cannot explain "the physiological process by which a particular disease or syndrome develops."[106] This is true even when the doctor engaged in a delineated protocol for ruling out possible causes of an injury. For instance, in an action alleging that a slip-and-fall injury led to hormonal damage and ultimately fibromyalgia, the U.S. Court of Appeals for the Fifth Circuit rejected the expert's methodology as unsound and held it was properly subject to exclusion. The doctor, Mary Reyna, who was certified in pain medicine, followed a "protocol" for

diagnosing fibromyalgia that included taking a medical history, ruling out prior or subsequent causes of the condition, and performing or reviewing physical tests, which were all negative. Doctor Reyna then "deduced" that the plaintiff's fall was the only possible remaining cause of her illness.[107]

The court was unimpressed. "This is not an exercise in scientific logic," it observed, "but in the fallacy of *post-hoc ergo propter-hoc* reasoning, which is as unacceptable in science as in law. By the same 'logic,' Doctor Reyna could have concluded that if [the plaintiff] had gone on a trip to Disney World and been jostled in a ride, that event could have contributed to the onset of fibromyalgia."[108] Absent a "specific train of medical evidence," the court refused to accept the reliability of the diagnosis, despite the doctor having "ruled out" a number of possibilities.

In SBS prosecutions, this "specific train of medical evidence" is also missing—noticeably so, now that the medical establishment has moved away from shaking as the exclusive (or even an identifiable) causal mechanism.[109] Notwithstanding this evidentiary deficit, prosecution experts maintain that the triad could result only from abuse inflicted immediately before the baby's collapse. With respect to alternative causes that were ruled out, rarely are doctors called upon to describe their methodological choices, much less explain them to an inquiring judge. The expert testimony is admitted without regard to the universe of possibilities considered, how each was eliminated, what might still remain, and whether the chosen cause—be it shaking, or impact or, simply, abuse—rests adequately on a scientific foundation. This lax judicial oversight of expert claims regarding the triad is even more notable when viewed in wider legal context.

Proponents of the diagnosis often remark (rightly of course) that scientists are unable to test their hypotheses in SBS cases by shaking babies. But this cannot explain the disparity we see in the evidentiary treatment of causation testimony. For the civil realm, too, presents "situations of irreducible causal uncertainty."[110] Judges have nevertheless been unreceptive to the notion that these situations call for less rigorous evidentiary standards.[111] It is true, as one court has suggested, that speculative hypotheses (which often result from "irreducible causal uncertainty") serve a function in the medical realm, where "if the costs of action are low, doctors may want to act . . . without further support."[112] Even so, courts in civil cases have emphasized that inadequately supported opinions should be excluded from a trial.[113] As Judge Richard Posner, who is among the most influential jurists, once pronounced, "the courtroom is not the place for scientific guesswork, even of the inspired sort."[114]

Sustained examination of causation evidence, the rule in civil cases, has not penetrated criminal court, where judges are faced seemingly unaware with comparable admissibility decisions. The diagnosis of SBS rests entirely on claims of causation, as do prosecutions based on the triad. The state's experts continue to insist that the baby's neurological symptoms must have resulted from some type of abuse. When these opinions are given a pass, the convictions that result are not secure.

Anatomy of a Missed Diagnosis

In a given SBS case, if the triad resulted from a factor other than abuse, then there has been no crime, and the defendant is necessarily innocent. Even with a diligent search, it will not always be possible to identify the cause of neurological impairment. A more practical problem is that, once a baby presents with a triad, medical investigations tend to fall short. In rare instances, an alternative cause becomes obvious, but usually too late. Expert certainty that attends the triad tends to stand in for thorough consideration of causes other than shaking.

How mistakes are made, and how they are unearthed, implicates both the likelihood of error in SBS diagnosis and the unlikelihood of discovery. We do not know how often preexisting conditions in a baby are overlooked, as occurred when Melonie Ware, the Georgia caregiver, was found guilty of murder. But a look at cases where missed diagnoses ultimately were identified shows that we cannot rely on our adversary system of justice to forestall the conviction of innocents.

Julie Baumer

In 2005, Julie Baumer was found guilty of violently shaking her six-week-old nephew, Ben. Baumer, who worked as a mortgage loan officer in a southeast Michigan town, had been caring for Ben since his birth. (Ben's mother struggled with drug addiction, and his biological father was not in the picture.) Upon her conviction for felony child abuse, Baumer was sentenced to ten to fifteen years in prison.[115]

Two years later, her appeal was denied.[116] Baumer's claims and the reasons for their failure are typical of SBS appeals, as we will see. The primary challenge was to the sufficiency of the evidence. In particular, Baumer's appellate lawyer emphasized the trial testimony of the defense expert, a pathologist,

who suggested that birth trauma might explain Ben's collapse. The prosecution experts denied this was a possibility. In their view, the meaning of the triad was unambiguous.[117] This was enough to satisfy the standard of appellate review. The jury was entitled to credit the state's experts, who unequivocally established Baumer's guilt. The court concluded that "when the evidence is viewed in the light most favorable to the prosecution, sufficient circumstantial evidence was presented from which the jury could reasonably infer that defendant knowingly or intentionally caused serious physical harm to the victim."[118]

Baumer further contended that her lawyer was ineffective for failing to pursue the defense of birth trauma, instead speculating that another family member had caused Ben's injuries. This rationale for a new trial was also rejected. Citing the "strong presumption that counsel's performance constituted sound trial strategy," even when it did not prove successful, the appeals court refused to "second-guess with the benefit of hindsight" the defense lawyer's decision not to argue that the baby's brain was bleeding since birth. Her appeal denied, Baumer faced another eight to thirteen years of incarceration.[119]

Many defendants, lacking the resources to fund a legal challenge, resign themselves to their fates when an appeal fails. Those in a position to do so may opt to attack their convictions collaterally. Baumer chose to file a petition for a writ of habeas corpus in federal district court, raising claims of ineffective assistance of counsel, insufficiency of the evidence, and actual innocence.[120] Because she had not exhausted her available state-court remedies, however, the petition was dismissed on purely procedural grounds.[121] According to the court, Baumer was required to move for post-conviction relief under the applicable state statute authorizing such claims. Upon a denial of her motion, Baumer would need to appeal first to the Michigan Court of Appeals, and then to the Michigan Supreme Court before she could refile her federal habeas petition.[122] While she served her time in prison, Baumer's case would have to run its course in state court.

She had by now secured new legal representation. Baumer's lead defense attorney was the county prosecutor when charges against her were initiated.[123] (Because he never dealt directly with the case as a prosecutor, and the trial was held after he left office, he was permitted to represent Baumer.) In August 2009, four years after she was convicted, a three-day evidentiary hearing was held before the trial judge that sentenced her. Later that fall, the judge vacated Baumer's conviction based on the ineffective assistance of her trial lawyer.[124]

According to the court, defense counsel's retention of a single expert to testify at trial was not enough to meet the minimal standard of competent

representation. The expert, a pediatric forensic pathologist, by her own admission was unqualified to interpret the key radiological evidence. "There was no strategic reason for defense counsel's failure to investigate and hire" *the right* expert—in this case, a radiologist—wrote the court. Because the lawyer's failure to do so "was based solely on financial concerns," his performance was legally deficient.[125]

What made the lawyer's failings worse, according to the judge, was that—despite appearances to the contrary at trial—Baumer had a valid medical defense. Ben suffered from venous sinus thrombosis (VST), a condition in which a blood clot forms in the sinuses that drain blood from the brain, which led to the neurological symptoms that doctors mistook for SBS. Three defense experts testified at the post-conviction hearing that VST was a missed diagnosis, and that Ben was not abused. In the opinion setting aside Baumer's conviction, the trial judge stressed the shared opinion of the defense experts: Ben's radiology reports did not indicate any traumatic injury; rather, the bleeding and retinal hemorrhaging was "clearly and solely due to VST."[126]

The court recognized that medical testimony was the essence of the state's case against Baumer, and that the prosecution doctors relied upon CT/MRI scans to diagnosis abuse. Given this, a competing interpretation of the radiological evidence would have been critical to a trial defense. Because there was a "reasonable probability that but for counsel's error," Baumer would have been acquitted, the court determined that she suffered actual prejudice warranting a grant of post-conviction relief. Baumer was conditionally released from prison while the state appealed the ruling.

Since the evidence that Ben had suffered a childhood stroke was enough to induce a judge to vacate Baumer's conviction, her new lawyer, the former prosecutor, expressed hope that the same proof would persuade the County Attorney not to retry the case. "We think the evidence is overwhelming in favor of her innocence," he remarked.[127] But after the Michigan Supreme Court declined to hear its appeal, the state chose to proceed once again.

The second trial lasted over three weeks. The prosecution presented five experts—three who read the radiology reports, and two who treated Ben—and the defense offered six.[128] This time, the SBS triad was not enough to convict. "There was absolute reasonable doubt," said the foreperson. "We had two sets of experts with two different opinions. Who do you believe? We had to set that aside and say, 'Is Julie responsible for this?' And the answer is 'no.' "[129]

Baumer spent four years in prison. But she was fortunate to have happened upon lawyers, students at the University of Michigan Law School's Innocence Clinic, and doctors willing to donate their time to exonerate her.

Few defendants have the resources to afford Baumer's defense—which would have cost more than $150,000 in fees, were it not for the pro bono efforts of her dedicated team.[130]

When the case ended, there were those who continued to believe that the triad proved guilt: prosecutors, their experts, Ben's adoptive family.[131] For the trial judge and for a jury, the radiology scans used to diagnose SBS were in fact powerful evidence of a natural disease process—one that the state's doctors completely failed to detect. According to one of the defense neuroradiologists, Doctor Michael K., "while this condition has been recognized for decades, it is a difficult diagnosis that is often missed, particularly on CT scan."[132] In Ben's case, doctors should have ordered a prompt MRI scan of the venous sinuses, suggested Doctor K. But even with "good imaging, this diagnosis may still be missed."[133]

Like other "mimics" of abuse, VST provides jurors with an alternative cause of the three symptoms that, according to the state, have their origin in shaking alone. But it is difficult for a defendant to unearth this condition. More often, the presumptive diagnosis of SBS is presented as the only real possibility. When the state's doctors default to abuse if presented with the triad, or when the defense fails to hire experts qualified to review the imaging (or other relevant tests, for that matter), an innocent explanation may well go uncovered. About VST, Baumer's lawyer remarked after the acquittal, "If you're not looking for it, you won't see it."[134]

Drayton Witt

In 2002, just convicted by a jury of shaking his four-month-old son Steven to death, Drayton Witt addressed the court for the first time. "Your Honor, [I would] like to introduce myself, first off. To everybody in the courtroom, I am the defendant. But to people that know myself, I am Drayton Shawn. I know you get a lot of people in front of you daily saying I am sorry, asking for mercy. I am different. I am not sorry, for I didn't do no wrong. But I am up here to tell you how much my son meant to me."[135]

After hearing from Witt, the judge was ready to pronounce sentence. "I did preside at the trial. I would say the expert testimony was overwhelming that this was not the result of any illness and anyone who sat in this courtroom for those two weeks, listened to those individuals, would be likewise convinced."[136]

Yet those who sat in the courtroom during Witt's trial, however convinced, were not privy to important medical evidence. Despite "overwhelming"

expert testimony that "this was not the result of illness," in fact the baby's death was quite likely the result of illness. But this would only become apparent a decade later. Without any reason to believe that Witt's conviction would ultimately be vacated, the judge sentenced the defendant to twenty years in prison.

In 2012, Witt's new lawyers at the Arizona Justice Project contacted Doctor A. L. M., the forensic pathologist who conducted Steven's autopsy, and asked that he review the case.[137] After doing so, Doctor M. made an extraordinary proclamation: "I have determined that I cannot stand by my previous conclusion and trial testimony that Steven's death was a homicide. . . . If I were to testify today, I would state that I believe Steven's death was likely the result of a natural disease process, not SBS."[138]

In his sworn declaration submitted on behalf of Witt, Doctor M. explained why, a decade before, he ruled Steven's death a homicide:

> By the time of the autopsy, I was notified that physicians at Phoenix Children's Hospital suspected that Steven Witt had been a victim of child abuse and, more specifically, Shaken Baby Syndrome (SBS). I did not find any outward signs of abuse (or violent impact) on Steven's body, but, during this autopsy, I observed that Steven Witt had retinal hemorrhages and optic nerve sheath hemorrhages (bleeding within the eyes and around the optic nerve), subdural hemorrhage (bleeding in the subdural area overlying the brain), and cerebral edema (brain swelling). Based upon these observations during the autopsy and the consensus of medical and scientific research and knowledge known to me at the time, I concluded that Steven died from SBS.[139]

At Witt's 2002 trial, Doctor M. testified for the prosecution in a manner consistent with these autopsy findings. He was now disavowing this testimony. In explaining his turnaround, Doctor M. stated under oath that, since Witt's trial, there had been "significant developments in the medical community's understanding of SBS, most of which serve to undermine the reliability of the SBS diagnosis." Many conditions, he added, could create the "very symptoms and injuries once thought to be nearly exclusively attributable to SBS." New perspectives on the triad cast doubt on the version of the diagnosis that convicted Witt of murder.[140]

As Doctor M. also observed, the baby had a "complicated medical history, including unexplained neurological problems." In his view, these problems were overlooked clues to what happened to Steven.[141] Other experts,

reviewing Witt's file at the request of his lawyers years after he was convicted, thought the same. One of these doctors was Doctor A. Norman Guthkelch, the author of the 1971 paper, "Infantile Subdural Hematoma and Its Relationship to Whiplash Injuries," among the first studies to advance the hypothesis that would later become SBS. Doctor Guthkelch, in a sworn affidavit filed on behalf of Witt,[142] criticized the assumption underlying the classic diagnosis—namely, that the triad meant that a baby had been shaken. As Doctor Guthkelch remarked, there was "not a vestige of proof when the name [SBS] developed that shaking alone causes the triad—subdural hematoma, retinal hemorrhages, and brain swelling. . . . In fact, it is likely that many other things besides shaking can cause the triad." Doctor Guthkelch concluded, a "diagnosis of non-accidental death, such as 'shaken baby syndrome,' is not justified when the only evidence of abuse available is the triad."[143]

Steven's SBS diagnosis exemplified the danger of equating the triad and abuse. The case demanded a far "more thorough review" than was given.[144] Indeed, Doctor Guthkelch cited a host of "confounding factors" arising during the baby's short life that might have contributed to his death: several attacks of persistent seizures (one requiring a six-day hospitalization at the same children's hospital where doctors would only a month later diagnose SBS); a flawed intubation in which the tube was misplaced into the esophagus, depriving the baby of oxygen; a recent infection; a possible metabolic disorder; and severe dehydration.[145] When asked in a deposition whether there was enough evidence to say that Steven was abused, Doctor Guthkelch responded with an unqualified "No."[146] Five other doctors reached a similar conclusion: this was a death from natural causes.[147] In the experts' opinion, Steven probably died from venous thrombosis.[148]

Based on this new evidence, Witt's lawyers petitioned in early 2012 for a new trial. Under state procedural rules, a defendant is entitled to relief when newly discovered evidence would probably have changed the verdict. Here, Witt's attorneys argued that there was a "significant shift in medical opinion" regarding the cause of Steven's death, and an evolution in SBS generally. Witt's conviction rested on medical testimony that could "be demonstrated to be, in material respects, false and in other respects subject to a fierce medical and scientific debate." With the new evidence before it, a jury might well view Witt as innocent—or, at the very least, possess reasonable doubt about his guilt.[149]

For months before their baby's death, Witt and Steven's mother, Maria, sought medical explanations and treatment for Steven's obvious neurological problems. After his six-day stint at the children's hospital, the infant was released without an explanation for his continuing seizures. Steven's health

continued to prompt concern and calls by Maria to the hospital emergency room and to the family pediatrician, who noted the possibility of a sepsis infection in the wake of the baby's recent hospitalization. Maria repeatedly asked the pediatrician for a referral to a pediatric neurologist who was covered by her insurance, to no avail. In the weeks before his death, Steven was feverish and experienced frequent bouts of vomiting.[150]

One night, Witt became worried that his baby was deteriorating. He drove to the restaurant where Maria was working and the two brought Steven to the hospital. On the way, Steven had another major seizure. After a failed effort to insert a tube in the trachea, the infant's heart stopped beating. Before his transfer back to the hospital where he had been admitted for recurring seizures the month before, Steven became severely dehydrated. On top of cardiopulmonary arrest, he was diagnosed with possible sepsis.[151]

Upon his arrival at the hospital, however, doctors became fixated on a diagnosis of SBS. Within about an hour, according to the trial testimony of Doctor Patricia T., the pediatric critical care doctor, she discovered bilateral retinal hemorrhages.[152] At this point, she explained, "I suspected and would be concerned that there was a head injury." Asked by the prosecutor, "what would that be," Doctor T. answered, "The entity called shaken baby syndrome. He came in with really a catastrophic, unexplained event with little history to support it, quit breathing, and then retinal hemorrhages. And it was really the only finding of significance that I could find on Steven."[153]

Doctor T. proceeded to order a CT scan, which showed subdural hygroma (pooling of cerebrospinal fluid into the subdural region), subdural bleeding, and cerebral edema, all of which confirmed for Doctor T., "it's a traumatic injury." Soon after, child protective services, the police, and the hospital's child abuse specialist were notified. The case followed the standard course, with doctors along the way confirming the early suspicion of SBS. Even Steven's documented medical history did not disrupt the conventional diagnostic approach to the triad. The hospital's neurologist who had treated Steven during his hospitalization the month before was never consulted. Steven died the next day.[154]

Present during the autopsy were a child abuse pediatrician and a police detective.[155] No attention was given, it seems, to what experts would later identify as a thrombosed [clotted] vein in one of the autopsy photographs.[156] Upon the classification of Steven's death as a homicide, Witt was charged with first-degree murder.

The state's proof of guilt was the routine testimony of doctors regarding the definitive meaning of the triad. Steven's injuries were caused by whiplash

forces as powerful as the forces generated by high-speed motor vehicle accidents. As a result of this violent shaking, bridging veins tore and caused immediate neurological collapse. Apart from shaking, only a "severe head-on car accident" could bring about this type of retinal hemorrhage, which was described as "large globules of blood with sharp edges and acute looking bright red." In short, as told by the state's five doctors, Steven's injuries were caused by violent shaking that could only have occurred while he was in his father's care. The baby's difficult birth, his persistent and unexplained seizures, his past fevers and infection, all of which were documented, were not relevant to his final collapse. Doctors were certain that SBS was the correct diagnosis. At trial, Witt's only expert was a forensic pathologist who, without really undermining the state's testimony regarding SBS, suggested that Steven had probably died of dehydration.[157]

Witt's conviction was cast in doubt only after he had served a decade in prison, when effective expert review of Steven's medical records uncovered fundamental weaknesses in the evidence against Witt—the very evidence that once seemed "overwhelming," in the words of the sentencing judge. Unlike when Witt was tried, his lawyers were able to identify a plausible alternative cause of the baby's death. Its experts were of the collective view that Steven died from an ongoing disease process, one that may have led to venous thrombosis.[158]

An explanation for the baby's death was not all that was new since the trial. As Witt's lawyers emphasized in their new trial motion, establishment consensus had shifted regarding the mechanics of SBS, its "mimics," and its unknowns.[159] With the benefit of time, it was clear that many of the opinions of the state's experts were weakly supported; others were wrong.

The state did not file a response to Witt's petition for post-conviction relief, and in the spring of 2012, Witt was released from prison.[160] Maricopa County Attorney Bill Montgomery initially promised to retry Witt, explaining that his office continued to believe Witt was guilty. (This is the case where the prosecutor explained, "Obviously we believed it the first time around."[161]) Though the diagnosis used to prove guilt at the time of Witt's conviction had been revised substantially, Montgomery commented, "I think we're still looking at cases where children were injured." But, he acknowledged, "how we prove that may change."[162]

In court, guilt is proven with evidence—in SBS prosecutions, medical evidence. Regardless of what prosecutors (or doctors) may happen to *believe* about the meaning of the triad, any "change" in how guilt is established depends on the continued willingness of experts to testify to the requisite

degree of certainty about its diagnostic worth in a particular case. Faced with this reality, it is understandable that the Maricopa County Attorney later moved to dismiss all charges against Witt.[163] Whatever prosecutors may have believed, they could not prove Witt's guilt.

Abigail Tiscareno

Abigail Tiscareno's first trial, in 2004, was much like those of others whose prosecutions rest on the triad.[164] When one-year-old Nathan was left at Tiscareno's Park City, Utah, day care that morning, he appeared to be healthy. Just hours later, Tiscareno called 911 to report that the baby was having trouble breathing. Nathan survived but suffered permanent brain damage, and Tiscareno was charged with felony child abuse.[165]

At her trial by jury, prosecution experts testified that the injury to Nathan's brain was so severe that it would necessarily have been inflicted immediately before the child collapsed. As one doctor explained, "he would have been severely injured and it's just not consistent that he would have done anything after the injury occurred."[166] Because Tiscareno was with Nathan when he fell unconscious, she was deemed guilty.

Although a CT scan showed two colors of blood, raising the prospect of older bleeding in Nathan's brain, the prosecution witnesses dismissed this possibility. The child abuse specialist who directed the hospital's child abuse program, Doctor Lori F., testified that "there was no evidence that I could determine in consulting with all of the other physicians that there was a preexisting chronic bleed in Nathan's head so that it was all very acute or all very, very fresh." Doctor Marion W., the neurosurgeon, stressed that "there was no old blood at all that we could see. Everything we saw was fresh."[167]

In addition to the medical evidence, the prosecutor introduced Tiscareno's account to police investigators. The caregiver described finding Nathan in his crib in a semiconscious state, gasping for air. She attempted to rouse him by calling his name, twice, and jostling him back and forth. This was presented as her confession to the abuse. At trial, Tiscareno maintained that she never shook Nathan other than in the course of revival efforts. But, as is typical, she could provide no satisfactory explanation for the baby's neurological symptoms.[168]

In contrast, the prosecution experts offered certainty: Nathan's injuries were acute. Admittedly, parts of the hematoma evidenced by CT scan and later removed from his brain were suggestive of old bleeding. But a clot was sent to pathology and, according to the uncontroverted testimony of

prosecution doctors, the results of this microscopic analysis confirmed that the bleeding was entirely new.[169]

In closing argument, the prosecutor emphasized the doctors' insistence that "this injury could not have occurred at any other time but in the morning," when Nathan was in Tiscareno's care. Discounting a defense expert's testimony that the baby's brain may already have been bleeding when he was left with the caregiver, the prosecutor remarked, "Doctor [W.], the man who opened up Nathan's skull and looked inside . . . would have a firsthand account of whether or not there's new or old blood." There was no old blood, the prosecutor maintained.[170]

Tiscareno was convicted. A mother of three school-aged children, she faced fifteen years in prison when her new team of lawyers discovered a pathology report that had never been disclosed to the defense.[171] The report, which was inexplicably missing from the medical records provided by the state, revealed that the hematoma found in Nathan's brain—the one that had been sent to the laboratory for pathological analysis—in fact contained old bleeding. Microscopic testing showed that the chronic subdural hematoma observed on the CT scan (then denied by the doctors who testified against Tiscareno) was indeed real. There was old blood.[172]

This report had just been found when the trial judge granted the defense motion for a new trial, on unrelated grounds. (The jury had been improperly instructed.) Despite the emergence of a central fact that directly contradicted its experts' repeated assurances, and upon which the caregiver's conviction rested, the prosecution nonetheless decided to try the caregiver a second time. This time, there was no denying that Nathan's brain was bleeding before he was placed in Tiscareno's care on the day in question. But the state's doctors still were certain that the baby had been shaken immediately before he collapsed. Now that the presence of old bleeding was a given, the experts minimized its diagnostic significance. The chronic subdural was a nonfactor, its origins unknown and its import negated by the newer blood.[173]

At Tiscareno's second trial, Doctor Lori F., the child abuse specialist, adhered to the position that Nathan must have been abused on the defendant's watch. "[A]t the time that he was normal," she explained, "he couldn't have been injured this severely." The "configuration of retinal hemorrhages" excluded all non-traumatic possibilities. And because of Nathan's "severe neurological deterioration, severe edema, massive hemorrhage," this was necessarily "a new injury." Asked whether there was "anything in your evaluation of Nathan, including all of the tests you—that you reviewed, the doctors you consulted with, your experience, to suggest that this was—that his injuries

were a result of significant trauma, a lucid interval, and then some kind of spontaneous reoccurrence of bleeding," Doctor F. answered simply, "no." She later added that it was "the degree of seriousness and amount of trauma that's reported at surgery and the amount of brain injury that he suffered at that moment that caused me to make the diagnosis that he was quickly symptomatic. His injury was so severe as to be as close to fatal as you can get."[174]

Because of the presence of old blood, now conceded, there could be no question that Nathan experienced an interval of lucidity while his brain bled. But Doctor F. nevertheless insisted that new bleeding was what caused Nathan to collapse, and that *it* (the new bleeding) was inconsistent with a period of consciousness. On cross-examination, defense counsel pressed Doctor F. on her underlying reasoning, which proved rather circular.

Q: So the fact that there are disrupted or broken axons is the reason that the person is immediately symptomatic?
A: Right.
Q: What evidence, I want you to tell me everything that you rely on that shows that Nathan [] had axons that were disrupted or broken?
A: We only have clinical evidence. The acute traumatic unconsciousness that he experienced.
Q: He was unconscious?
A: He was unconscious. We don't have histologic [microscopic] data. We have brain edema, which shows that there's damage.
Q: Swelling. What else?
A: Those are the main reasons.[175]

Doctor F. conceded that chronic subdural hematomas can re-bleed spontaneously, but declared that this had not occurred with Nathan. Chronic subdurals, she explained, "don't generally cause traumatic unconsciousness. The symptoms are much more indolent. They progress much more slowly, you get an idea that the kid is becoming more irritable, slowly becoming more lethargic or they may just have no symptoms at all." While acknowledging that the re-bleeding of a chronic subdural can be asymptomatic, Doctor F. nevertheless remained confident that Nathan's new hemorrhaging was fully unrelated to the old. Because he was a healthy child with no indications of neurological impairment—no irritability, no lethargy, no vomiting, no lack of appetite, no fever—it could not have been the case that Nathan was experiencing the re-bleeding of a chronic subdural. Doctor F.'s understanding in this regard came from the baby's father, who reported that Nathan had been "fine all week."[176]

Defense counsel probed further:

Q: So in making that diagnosis, you are relying on the accuracy of what he tells you?
A: Yes.
Q: You didn't talk to anyone else?
A: No.
Q: In front of you is a book of exhibits.... Do you see that book?
A: This one?
Q: Yes.
A: Okay.
Q: Let me represent to you that the tabs one, two, four—one, two, and four are Nathan's medical records prior to the time he was admitted to the [hospital].... Have you reviewed those?
A: No.
Q: Never seen them?
A: Not from his pediatrician, no.
Q: So in making the diagnosis as to what happens to Nathan and telling the police investigators... what occurred, you didn't think it was important to review his pediatric records?
A: No.[177]

Doctor F. admitted that, within hours of first seeing Nathan, she told the police that Nathan had been severely shaken and that he would have lost consciousness immediately thereafter. She well understood the significance of this diagnosis for law enforcement purposes. As she testified:

Q: You created a time line for the officers, correct?
A: Correct.
Q: As to [how] this abuse could have occurred?
A: Yes.
Q: And it wasn't a time line as to this is what I think it probably is, you said it was impossible for it to have occurred any time outside of your time line, agreed?
A: It was based on symptoms, yes.
Q: Not asking you—I'm just asking you, that's what you conveyed to the officers?
A: That's correct.
...

Q: At the time you told them that, you had a CT scan which had been read by a neuroradiologist as indicating that Nathan had an acute subdural superimposed upon a chronic subdural, agreed?
A: Uh-huh.[178]

Doctor F. was sure that Nathan's bleeding was caused by recent shaking—so sure that she proceeded despite CT scan evidence to the contrary. At the very least, ambiguity surrounding the scan would seem to have suggested the need for agnosticism regarding the cause of the baby's symptoms, particularly since a pathologist would soon be examining the tissue on a microscopic level. But the same certainty that would later allow Doctor F. to reject the relevance of the old bleeding might explain why she declined to reserve judgment pending review of the pathology finding.

Q: It didn't occur to you to even inquire, to go, well, goll, we've got the CT scan that says it's a chronic…what [is] the pathology report [] going to say?
A: I didn't know a piece of tissue had been sent to pathology. It would depend on the neurosurgeon to do that.
Q: I didn't ask you if you knew this had been sent. I asked you if you knew that it's an important piece of the puzzle.
A: It is important.
Q: And yet you narrow this time line and sent everyone on their way and said it can only be at this time frame without ever getting that pathology report?
A: Yes.
Q: Not only did you not get it that day…or that month, you never got it?
A: That's correct.[179]

Even when she did read the report and learned that there was old bleeding, Doctor F. held steadfast to her belief in the correctness of her original conclusion. She admitted as much in her testimony at the second trial. When asked "in terms of the time line as to who could have perpetrated this, in your mind, that is an absolute; nothing is going to change that?" Doctor F. responded, "In this scenario, nothing [is] going to change my mind." Queried, "that's an absolute," she answered, "yes, that's an absolute, yes."[180]

This time around, though, a trial judge acquitted.[181] Defense experts had successfully challenged the notion that the old bleeding was irrelevant to Nathan's condition. In their estimation, Nathan was likely experiencing

a protracted neurological decline.[182] The judge also heard about the limitations of CT findings, both for establishing a time frame for bleeding and for identifying its origins.[183] No single, definitive explanation for Nathan's condition was provided by the defendant. But the prosecutor could not overcome evidence that the baby's brain was already bleeding when he was delivered to Tiscareno.

Many other prosecutions involving unexplained chronic bleeding (and caregivers with equally impeccable records) have resulted in convictions. This mother was returned to her children.

Of her eighteen-month ordeal, which included the contemplation of fifteen years in prison, Tiscareno says, "I no longer had a life. I just wanted to be with my kids and my husband and pray it would be over."[184] She recounts how her children were mocked at school and her family's savings were spent on her defense. She describes her own public humiliation and the loss of a career in child care. And she grieves for Nathan's diminished existence. As Tiscareno's husband, Guillermo, tells it, "now everything is gone. I don't care about the money. Thank God I have hands to work. Maybe someday we can recover, I don't know. But they threw my wife's reputation out the window. It's going to be hard."[185]

This is the relatively happy ending.

LUCY B. RORKE-ADAMS

Lucy B. Rorke-Adams is a graduate of the University of Minnesota (BA 1951, MA 1952, BS 1955 and MD 1957). Her internship, residency (anatomical pathology) and fellowship (neuropathology) were done at the Philadelphia General Hospital (PGH) from 1957 through 1962 following which she became Board Certified in Anatomic Pathology and Neuropathology. Dr. Rorke-Adams joined the staff of PGH in 1962 as Chief of Pediatric Pathology and Assistant Neuropathologist. In 1965, while still at PGH, she also assumed responsibility for neuropathology at The Children's Hospital of Philadelphia where she is currently Senior Neuropathologist.

Dr. Rorke-Adams became forensic neuropathologist for the Office of the Medical Examiner of Philadelphia in 1977 and served in this capacity until 2004 and then from 2007 through 2014. One of her major responsibilities involved evaluation of cases of putative child abuse. This led to consultation from other jurisdictions and a seat on the Committee of Medico-Legal Aspects of Child Abuse of the Attorney General of Pennsylvania.

Dr. Rorke-Adams is author/co-author of 294 peer-reviewed publications, 40 non-peer-reviewed (invited) publications, and 25 editorials, reviews and chapters as well as two books.

Dr. Rorke-Adams is the recipient of numerous honors, a few of which include the following: President, American Association of Neuropathologists, Bronze Plaque for Meritorious Contributions to Advancement of Neuropathology from Am. Assn of Neuropathologists, Provost's Award for Distinguished Teaching from Univ of Pennsylvania, Honorary Member American Assn of Neuroradiology, Secretary of College of Physicians of Philadelphia, Honorary Member of Neuropathological Societies of Australia/New Zealand, Argentina and Spain, and Establishment of the Lucy Balian Rorke-Adams Chair in Pediatric Neuropathology at The Children's Hospital of Philadelphia.

tected. Failure to look beyond the simplistic and increasingly untenable shaking hypothesis risks incalculable damage by wrongfully removing children from loving parents or incarcerating innocent people. Further, by focusing on shaking or inflicted trauma to the exclusion of accidental and natural causes, we are almost certainly missing opportunities to save babies through prevention, early diagnosis and treatment.

References

1. Crown Prosecution Service. Non-accidental Head Injury (NAHI, formerly referred to as Shaken Baby Syndrome [SBS])-Prosecution Approach. http://www.cps.gov.uk/legal/l_to_o/non_accidental_head_injury_cases/. 2011.
2. Duhaime AC, Gennarelli TA, Sutton LN, Schut L. 'Shaken Baby Syndrome': a misnomer? J Pediatr Neurosciences. 1988;4(2):77–86.
3. Ommaya AK, Goldsmith W, Thibault L. Biomechanics and neuropathology of adult and paediatric head injury. Br J Neurosurg. 2002 Jun;16(3):220–42.
4. Leestma JE. Case analysis of brain-injured admittedly shaken infants: 54 cases, 1969–2001. Am J Forensic Med Pathol. 2005 Sep;26(3):199–212.
5. Shannon P, Smith CR, Deck J, Ang LC, Ho M, Becker L. Axonal injury and the neuropathology of shaken baby syndrome. Acta Neuropathol (Berl). 1998 Jun;95(6):625–31.
6. Winter SC, Quaghebeur G, Richards PG. Unusual cervical spine injury in a 1 year old. Injury. 2003;34(4):316–9.
7. Barnes PD, Krasnokutsky MV, Monson KL, Ophoven J. Traumatic spinal cord injury: accidental versus nonaccidental injury. Semin Pediatr Neurol. 200815(4):178–84.
8. Christian CW, Block R. Abusive head trauma in infants and children. Pediatrics. 2009;123(5):1409–11.
9. Adamsbaum C, Grabar S, Mejean N, Rey-Salmon C. Abusive head trauma: judicial admissions highlight violent and repetitive shaking. Pediatrics. 2010;126(3):546–55.
10. Browder J, Kaplan HA, Krieger AJ. Venous lakes in the suboccipital dura mater and falx cerebelli of infants: surgical significance. Surg Neurol. 1975;4(1):53–5.
11. Mack J, Squier W, Eastman JT. Anatomy and development of the meninges: implications for subdural collections and CSF circulation. Pediatr Radiol. 2009;39(3):200–10.
12. Matshes E. Retinal and optic nerve sheath haemorrhages are not pathognomonic of abusive head injury. Presentation G1 (Pathobiology). American Academy of Forensic Sciences. Seattle, 2010:p272.

The triad of retinal haemorrhage, subdural haemorrhage and encephalopathy in an infant unassociated with evidence of physical injury is not the result of shaking, but is most likely to have been caused by a natural disease

NO

It has been the practice of physicians to organise historical, physical and laboratory findings which occur with some frequency into syndromes or specific disease entities, and contributions by pathologists often provide a morphological base for the disorder. Thus, in the century and a half interval since Rudolf Virchow's studies earned him the sobriquet of 'Father of Pathology', innumerable diseases have been recognised, although unfamiliar constellations continue to challenge the diagnostic acumen of physicians, requiring ongoing clinical and pathological investigations to establish their place in the spectrum of disease.

Among this group are those that appear to be associated with child abuse. Although there is ample historical documentation of child abuse throughout the ages, a scientific approach to define the nature and extent of such abuse is a relatively recent phenomenon.[1] Whereas abuse may take many forms, the majority do not cause death, e.g. psychological or sexual abuse, but infliction of injury to the central nervous system (CNS) is among the most lethal; about two-thirds of child abuse victims who die do so because of CNS trauma.[2]

Clinical and pathological studies have documented three features associated with CNS trauma that occur so frequently they are commonly referred to as 'the triad', specifically, subdural haemorrhage (SDH), retinal haemorrhage (RH), and encephalopathy.

This triad is found in infants who may/may not exhibit other injuries, such as bruising and/

Lucy B Rorke-Adams
MD, Senior Neuropathologist, The Children's Hospital of Philadelphia; Consultant Forensic Neuropathologist, Office of the Medical Examiner, City of Philadelphia and Clinical Professor of Pathology, Neurology and Pediatrics, University of Pennsylvania School of Medicine, USA
rorke@email.chop.edu

or fractures. Pathogenesis of the triad has been ascribed to severe acceleration–deceleration forces consequent to shaking, plus or minus impact.

An enormous body of evidence based upon peer-reviewed studies has established the high frequency of association between the triad and shaken impact syndrome, with the caveat that this triad may not be pathognomonic for inflicted trauma.[3] Specifically, one or more components may signal a naturally occurring disease, including among others, a variety of haematological/coagulopathic disorders, rare metabolic diseases, vascular malformations, etc.

Routine diagnostic evaluation of infants who present with one or more features of the triad therefore includes a search for one of the known diagnostic possibilities in the context of history and ancillary investigations.[4,5]

Those who challenge the triad as a sentinel of possible nonaccidental trauma have advanced alternative disease states to explain its occurrence. Their list includes hypoxia-ischemia, birth injury, excessive coughing/vomiting, infections, vaccinations and venous thromboses.[4] It is of note that these alternative suggestions purporting to account for the features of the triad have been extant for a relatively short time, first appearing in 2003.[6]

This was a publication by Geddes et al, who theorised that pathogenesis of SDH and retinal haemorrhage was hypoxia-ischemia and not trauma. The study upon which this extraordinary claim was based was severely flawed, including, for example, no clinical or pathological examination of the eyes; two years later it was retracted by Geddes, but by that time, the evil genie had escaped Pandora's box, repercussions of which have been far-ranging. A considerable literature has since accumulated with contributions both from Geddes's supporters (even after her retraction) and a host of challengers.[7] Of primary importance is the fact that, to date, no reliable evidence base supporting a pathogenetic relationship between hypoxia-ischemia and subdural bleeding or retinal haemorrhages has been forthcoming.

Also lacking is evidence-based literature supporting the assertion that late consequences of 'birth injury' may be mistaken for nonaccidental head trauma. Experienced paediatric pathologists have documented falcine and small SDH in perinates dying of problems unrelated to the CNS, e.g. congenital anomalies, infections etc., and recent radiological studies have confirmed these observations.[8] The majority of the haemorrhages have resolved by one month of age, and if the infant comes to postmortem after a month or more, a delicate avascular membrane is sometimes found. The assertion that it is highly vascularised and may bleed spontaneously or consequent to minor trauma has no documented factual base.

It is also well established that retinal haemorrhages occur peripartum and these, too, disappear by four weeks of age.[9]

The claim that venous thromboses cause the triad is blatantly false. Although intracerebral haemorrhages are common, no standard texts of radiology or pathology document association of thromboses with SDH, although it is conceivable that small posterior pole retinal haemorrhages may result from increased intracranial pressure.[9]

Although subdural effusions and retinal haemorrhages are sometimes found in infants with bacterial meningitis, SDHs are exceptionally rare, even if the agent is haemolytic 'strep'. The retinal haemorrhages are basically caused by increased intracranial pressure and distinguishable by an experienced ophthalmologist from those consequent to trauma.[9]

Assertions that vaccinations or excessive coughing/vomiting cause subdural and retinal haemorrhages are clearly ludicrous. There is, in fact, strong evidence to the contrary concerning coughing/vomiting.[10-12] Surridge et al.[10] studied 72 patients who required intensive care because of pertussis, 97% of whom were less than 12 months of age, and reported CNS complications to include seizures and encephalopathy; three patients died. They found neither SDH nor RH clinically or pathologically.

A companion study by Cherry[11] of children with severe croup with/without pneumonia (including some with diphtheria) made no mention of SDH/RH as a complication in severely affected

The triad of retinal haemorrhage, subdural haemorrhage and encephalopathy in an infant unassociated with evidence of physical injury is not the result of shaking but is most likely to have been caused by a natural disease—the 'no' case. J Prim Health Care. 2011;3(2):161–163.

patients. Similarly, Fitzpatrick et al., who studied a group of children with cyclical vomiting syndrome, found none with complicating SDH/RH.[12]

The scientific base for shaken impact syndrome has accumulated over a period of at least 150 years, although sporadic writings of physicians, anatomists and writers commenting about effects of CNS trauma, in particular concussion, appeared long before that time.

The concept that SDH was a consequence of shaking was advanced in 1930, and innumerable observations of traumatised infants by Caffey, Kempe, Gutkelch and countless others, laid the foundation for the objective base of shaken impact syndrome upon which contemporary investigators continue to build.

Contributions by paediatricians, neuroradiologists, neurosurgeons, clinical and forensic pathologists, physiologists, ophthalmologists, biomechanical engineers, social workers, and law enforcement agents have formed the evidence base that currently supports the diagnosis of shaken impact syndrome.

Although components of the syndrome include the triad, the diagnosis is actually based upon a complex constellation of clinical-pathological-investigative findings. These include:

1. investigative data
2. clinical history, examination and therapeutic requirements
3. laboratory studies to rule out natural disease, and
4. radiological, ophthalmological and pathological findings, all of which are evaluated against a knowledge base of clinical disease and features of accidental versus nonaccidental trauma.

The triad is an important component within this complex constellation, but does not stand alone.

Specialists involved in the tragic field of child abuse remain ever mindful of the wisdom of John Dewey who said: "Intelligence is not something possessed once and for all. It is in constant process of forming, and its retention requires constant alertness in observing consequences, an open-minded will to learn and courage in readjustment." Those who offer untested hypotheses to defend individuals who have harmed infants do considerable disservice to science and to the victims.

References

1. Block H. Abandonment, infanticide and filicide. Am J Dis Child. 1988;142:1058–1060.
2. Rorke LB. Neuropathology. In: Ludwig S, Komberg AE, editors. Child abuse. 2nd ed. New York: Churchill Livingston; 1992.
3. Munns RA, Brown JK, eds. Shaking and other non-accidental head injuries in children. Cambridge: Mac Keith/Cambridge University Press; 2005.
4. Reece RM. Differential diagnosis of inflicted childhood neurotrauma. In: Reece RM, Nicholson CE, editors. Inflicted childhood neurotrauma. United States: Amer Acad Pediatrics; 2003.
5. Chiesa A, Duhaime A-C. Abusive head trauma. Pediatr Clin N Am. 2009;56:317–331.
6. Geddes JF, Tasker RC, Hackshaw AK, et al. Dural haemorrhage in non-traumatic infant deaths: does it explain bleeding in 'shaken baby syndrome'? Neuropathol Appl Neurobiol. 2003;29:14–22.
7. Jaspan T. Current controversies in the interpretation of non-accidental head injury. Pediatr Radiol. 2008;38 Suppl 3:S378–87.
8. Rooks VJ, Eaton JP, Ruess L, Petermann GW, Keck-Wherley J, Pedersen RC. Prevalence and evolution of intracranial hemorrhage in asymptomatic term infants. Am J Neuroradiol. 2008;29(6):1082–9.
9. Levin AV. Retinal hemorrhages: Advances in understanding. Pediatr Clin N Am. 2009;56:333–344.
10. Surridge J, Segedin ER, Grant CC. Pertussis requiring intensive care. Arch Dis Child. 2007;92:970–975.
11. Cherry JA. Croup. NEJM. 2008;358:384–391.
12. Fitzpatrick E, Bourke B, Drumm B, et al. Outcome for children with cyclical vomiting syndrome. Arch Dis Child. 2007;92:1001–1004.

J Neurosurg Pediatrics 3:232–239, 2009

Neck injuries in young pediatric homicide victims

Clinical article

Laura K. Brennan, M.D.,[1,2] David Rubin, M.D., M.S.C.E.,[1,2] Cindy W. Christian, M.D.,[1,2] Ann-Christine Duhaime, M.D.,[3] Haresh G. Mirchandani, M.D.,[4] and Lucy B. Rorke-Adams, M.D.[4,5]

[1]Division of General Pediatrics, The Children's Hospital of Philadelphia; [2]Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, Pennyslvania; [3]Department of Neurosurgery, Dartmouth-Hitchcock Medical Center, Lebanon, New Hampshire; [4]Office of the Medical Examiner of the City of Philadelphia; and [5]Division of Anatomic Pathology, The Children's Hospital of Philadelphia, Pennsylvania

Object. In this study, the authors estimate the prevalence of injuries to the soft tissue of the neck, cervical vertebrae, and cervical spinal cord among victims of abusive head trauma to better understand these injuries and their relationship to other pathophysiological findings commonly found in children with fatal abusive head trauma.

Methods. The population included all homicide victims 2 years of age and younger from the city of Philadelphia, Pennyslvania, who underwent a comprehensive postmortem examination at the Office of the Medical Examiner between 1995 and 2003. A retrospective review of all available postmortem records was performed, and data regarding numerous pathological findings, as well as the patient's clinical history and demographic information, were abstracted. Data were described using means and standard deviations for continuous variables, and frequency and ranges for categorical variables. Chi-square analyses were used to test for the association of neck injuries with different types of brain injury.

Results. The sample included 52 children, 41 (79%) of whom died of abusive head trauma. Of these, 29 (71%) had primary cervical cord injuries: in 21 there were parenchymal injuries, in 24 meningeal hemorrhages, and in 16, nerve root avulsion/dorsal root ganglion hemorrhage were evident. Six children with abusive head trauma had no evidence of an impact to the head, and all 6 had primary cervical spinal cord injury (SCI). No child had a spinal fracture. Six of 29 children (21%) with primary cervical SCIs had soft-tissue (ligamentous or muscular) injuries to the neck, and 14 (48%) had brainstem injuries. There was a significant association of primary cervical SCI with cerebral edema ($p = 0.036$) but not with hypoxia-ischemia, infarction, or herniation.

Conclusions. Cervical SCI is a frequent but not universal finding in young children with fatal abusive head trauma. In the present study, parenchymal and/or root injury usually occurred without evidence of muscular or ligamentous damage, or of bone dislocation or fracture. Moreover, associated brainstem injuries were not always seen. Although there was a significant association of primary cervical cord injury with cerebral edema, there was no direct relationship to brainstem herniation, hypoxia-ischemia, or infarction. This suggests that cervical spinal trauma is only 1 factor in the pathogenesis of these lesions. *(DOI: 10.3171/2008.11.PEDS0835)*

Key Words • **abusive head trauma** • **cervical cord injury** • **neck injury**

Since John Caffey first identified maltreatment as the cause of unexplained skeletal and brain injuries in infants in 1946,[11] considerable research has been published describing the epidemiology, patterns, and mechanisms of injury associated with abusive head trauma in young children. Although underrecognition still exists,[24] abusive head trauma affects ~ 17 in 100,000 US children annually.[27] The exact mechanism of injury in abusive head trauma, and the relative contributions of shaking and impact continue to be debated. Although some researchers report that shaking alone is sufficient to cause injury,[8,9,12,23,31,43] others conclude that blunt impact is necessary to cause significant primary brain trauma.[14]

Nevertheless, the clinical manifestations and outcomes of abusive head trauma have been well described. Many studies have documented that the majority of young infants who suffer abusive head trauma have extraparenchymal hemorrhage,[5,13,14,34,35] > 80% have retinal hemorrhages,[14,35] and 30–50% have skeletal injuries of various ages.[7,13,34] Between 60 and 85% show evidence of impact injury, mani-

Abbreviations used in this paper: OME = Office of the Medical Examiner; SCI = spinal cord injury; SDH = subdural hemorrhage.

fested by skull fractures or cranial soft-tissue contusions, some of which are not apparent until postmortem examination.[8,14,17,34] The mortality rate after abusive head trauma approaches 20%,[5,30] and in those who survive, > 75% will have permanent neurological impairments.[2,3,5,30]

Recent research on abusive head trauma has explored the relationship between abusive head trauma and hypoxic cerebral injury. Hypoxic-ischemic injury is common in patients with abusive head trauma[7,17,18,28,41] and it is theorized that trauma-induced apnea leads to cerebral hypoxia and/or ischemia.[17,18,25,28] Some investigators have speculated that secondary injury consequent to hypoxia-ischemia, edema, or infarction may be a stronger determinant of a patient's ultimate neurological outcome than the primary injury or the associated presence of subdural and subarachnoid hemorrhage, or diffuse axonal injury.[15,25,28]

Although patterns of brain injury in infants with abusive head trauma have been well described, there has been far less attention directed toward the prevalence and clinical significance of neck injuries in these infants. Because many investigators have postulated that a shaking whiplash event, with or without head impact, may be the primary mechanism in many infants with inflicted brain injuries,[12,14,23] it is reasonable to hypothesize that the necks of these young infants may also be injured. Additionally, infants and young children are at an increased risk of flexion, distraction, and rotational injuries to the cervical spine due to their horizontally oriented facet joints, incompletely formed uncovertebral joints, increased laxity of the associated spinous ligaments, immature paraspinous musculature, and a relatively large head-to-body ratio.[19] Some researchers have suggested that with purely rotational injuries (such as shaking), injury to the neck and cervical spinal cord should be universal.[1]

Despite concerns about the contributions of neck injury to the sequelae of abusive head trauma, the few case series in the literature have been limited in sample size and scope of data reported. We therefore report the postmortem findings in a cohort of young homicide victims in whom complete pathological analyses of the brains and spinal columns were performed. We have focused on estimating the prevalence of injuries to the soft tissue of the neck, the cervical vertebrae, and the cervical spinal cord to better understand these injuries and their relationship to other pathophysiological findings common in children who have sustained fatal abusive head trauma.

Methods

The target population for this study was all homicide victims 2 years of age and younger from the city of Philadelphia, Pennsylvania, in whom a comprehensive postmortem examination was done at the OME between 1995 and 2003. Children were eligible if the chief medical examiner in consultation with the senior neuropathologist onsite at the OME determined the manner of death to be homicide; all available information provided to the OME was used before reaching this decision. Findings were retrospectively reviewed for analysis. In our review, we included certification of identification reports, case registration summaries (including case histories), general autopsy reports, and neuropathology reports. Limited medical records, including imaging studies, were available in some cases. The general autopsy was completed by a forensic pathologist at the OME, and a full neuropathological examination was performed by a forensic neuropathologist (L.B.R.-A.). Examination of the nervous system included removal of the brain, spinal cord, and eyes in continuity, followed by gross and microscopic examination according to the standard techniques described elsewhere.[26] Demographic information, clinical history, and injury data were collected. Demographic variables included sex, race, and age at death. Race was categorized as Caucasian, African-American, Asian, Hispanic, Native American, or other, as recorded in the case registration summaries.

In addition to the documentation of cause and manner of death (homicide), the clinical history included the time interval between injury and death, and an assessment of the mechanism of death. Where possible, the time interval was calculated in days from reported symptom onset to death, as recorded in the case registration summaries. The mechanism of death was determined and reported by the chief medical examiner based on assessment of the available investigative data and results of the postmortem examination. Based on this report, we separated the cases into 2 categories: abusive head trauma and death consequent to non-CNS injury, including asphyxiation and abdominal trauma.

Table 1 lists all the pathological findings that were recorded. Neck injury data included the presence of cervical spine dislocation or fracture, ligament or muscle injury, or any soft-tissue injury of the neck (such as hematoma or bruising). Primary cervical SCIs included any cervical cord contusion, laceration, or transection; vertebral artery injury; nerve root avulsion/dorsal root ganglia hemorrhage; and meningeal hemorrhage (epidural, intradural, subdural, and/or subarachnoid). Given its proximity to the cervical cord, any traumatic injury to the brainstem (such as a laceration or hemorrhage) was also recorded.

Primary traumatic brain lesions included contusions or lacerations, parenchymal hemorrhage, and meningeal hemorrhage. Cerebral contusions and lacerations were further classified as superficial contusions or lacerations (defined as olfactory bulb/tract injury or cortical contusions/lacerations) and deep contusions and lacerations (defined as axonal injury, gliding injury, injury to the corpus callosum, or ventricular tears/lacerations). The presence or absence of cerebral edema, cerebral hypoxia-ischemia, a cerebral infarction, and herniation were determined as follows according to the usual neuropathological criteria[22] and were noted for each patient.

Cerebral Edema. The diagnosis of cerebral edema was determined by comparing the brain weight of the homicide victim with the expected value for a child of that age. Absolute figures cannot be given here because brain weight changes from birth to 24 months of life. Beginning with the brain weight determination, the pathologist then examined the specimen for sulcal effacement consequent to gyral widening and flattening. Microscopic diagnosis of cerebral edema rests on the presence of status spongiosis, acute swelling of oligodendroglia, and exaggeration of perivascular and pericellular shrinkage artifact.

TABLE 1: Summary of pathological findings in 52 infant homicide victims*

Neck	Cervical Spinal Cord	Retinal Hemorrhage	Cranial Findings	Secondary Cerebral Findings	Extracranial Findings
ligamentous injury	parenchymal cord injury	bilat	meningeal hemorrhage	cerebral edema	extracranial fx
	meningeal hemorrhage	unilat	epidural	brainstem herniation	rib fx
	nerve root avulsion/dorsal root		subdural	hypoxia-ischemia	
			subarachnoid	infarction	
muscle injury			intracerebral bleeding		concomitant visceral injury
other soft-tissue injury			cerebral contusions/lacerations		facial bruising
			superficial		body bruising
			deep		
cervical spine dislocation or fracture			DAI		
			evidence of BFT		
			skull fx		
			galeal/subgaleal hematoma		
			bruises to scalp		
			superficial cerebral contusions/lacerations		
			brainstem trauma		

* BFT = blunt force trauma; DAI = diffuse axonal injury; fx = fractures.

Hypoxia-Ischemia. The diagnosis of hypoxia-ischemia was based on established gross and microscopic features of the specimen. These often, but not always, included severe superficial and deep congestion, deep pink to purple discoloration of the gray matter (consequent to the pathophysiological mechanism of cerebral autoregulation), and microscopic identification of acute neuronal necrosis.

Cerebral Infarction. Cerebral infarction may take various forms that differ according to whether the infarction is acute, subacute, or chronic, and whether it is in a vascular distribution, such as in the middle cerebral artery, a border zone lesion, or laminary necrosis.

Herniation. Evaluation of a brain for evidence of herniation is a routine part of gross examination. Criteria for determination of herniation included grooving of unci and/or parahippocampal gyri, compression of the third cranial nerve in association with herniation of these structures, cingulate herniation beneath the falx cerebri, cerebellar tonsillar grooving, the presence or absence of associated brainstem swelling, presence of Duret hemorrhages, identification of the Kernohan notch phenomenon, and rarely, herniation of the temporal poles over the sphenoid ridge into the anterior fossa.

Meningeal hemorrhage was characterized as epidural, subdural, and/or subarachnoid if present. Evidence of blunt or impact trauma to the head, such as skull fractures, galeal or subgaleal hematomas, superficial contusions and lacerations, and any scalp hematomas or bruising was recorded. Based on injury identification, children with abusive head trauma were then classified as either having visible evidence of impact or no visible evidence of impact. The presence of retinal hemorrhages, determined through examination of the eyes, was also recorded.

Collected extracranial injury data included the presence of other injuries, such as rib fractures, extremity fractures, abdominal trauma (such as lacerations or contusions to internal organs or the presence of a hemoperitoneum), and cutaneous bruising or hematomas.

All data were abstracted from the autopsy reports and entered into an Access database (Microsoft Corporation). Data were then imported into STATA version 8.2 software (STATA Corporation) for analysis. Data were described using means and standard deviations for continuous variables, and frequencies and ranges for categorical variables. Chi-square analyses were used to test for the association of neck injuries with different types of brain injury.

This study was reviewed and approved by the Institutional Review Board of the Children's Hospital of Philadelphia.

Results

There were a total of 52 homicide victims 2 years of age or younger in the city of Philadelphia, Pennsylvania, between 1995 and 2003 who underwent postmortem examinations at the OME. Of this group, 41 (79%; 95% CI 65.3–88.9) died of abusive head trauma, and 10 died of other mechanisms, including 3 who died of asphyxiation and 7 who died of blunt abdominal or body trauma. An additional child died of complex injuries sustained when her mother jumped from a second story window with the infant in her arms. Although the young child had significant neurotrauma along with other injuries, including organ laceration and multiple fractures, the case was unique and therefore classified separately.

Young children with abusive head trauma were significantly younger at death than those killed by other means ($p = 0.036$). Infants with abusive head trauma were

TABLE 2: Population demographics of 52 infant homicide victims*

Characteristic	Abusive Head Trauma (41 children)	Other MOD (11 children)	p Value
sex (%)			
male	15 (36.6)	6 (54.6)	0.28
female	26 (63.4)	5 (45.4)	
age at death (%)			
<1 yr	22 (53.7)	2 (18.2)	0.036
1–2 yrs	19 (46.3)	9 (81.8)	
race (%)			
African-American	31 (75.6)	7 (63.6)	0.243
Caucasian	8 (19.6)	2 (18.2)	
Hispanic	1 (2.4)	2 (18.2)	
Asian	1 (2.4)	0 (0)	
length of time btwn injury & death (%)			
≤1 day	22 (57.9)†	9 (81.8)	0.147
≥2 days	16 (42.1)†	2 (18.2)	

* MOD = mechanism of death.

† Due to insufficient data, 3 children were not included in this group.

more often female and survived longer after injury, but these differences were not statistically significant. There were no racial differences between the children who died of head trauma and those who died of other mechanisms (Table 2).

Medical records, some including imaging results, were available for 22 of the 52 children. Four were dead on or shortly after arrival at the hospital, and none had undergone MR imaging of the neck or cervical spinal cord. Six children had CT scans of the cervical spine, all of which were negative for fractures, subluxation, or soft-tissue swelling. Magnetic resonance imaging of the brain was done in 5, yielding abnormal results in all cases with varying degrees of hemorrhage, infarction, and hypoxic injury. Of these 5, the cause of death in 4 was abusive head trauma, and the fifth child died of asphyxiation.

Neck Injuries

Twenty-nine of 41 children with abusive head trauma (71%; 95% CI 54–84) had primary injuries to the cervical spinal cord; 2 had cervical SCIs alone without associated primary traumatic brain injuries, and 1 of these was pronounced dead on arrival to the emergency department. No clinical history was available for this child, who also had multiple other injuries, including rib fractures, a femur fracture, liver lacerations, multiple subgaleal hemorrhages of scalp, and diffuse superficial contusions and abrasions to the chest and head. No cervical spinal cord images were obtained in this child. The second child died 3 days after presenting with extreme lethargy and respiratory distress. She initially received a diagnosis of acidemia and encephalopathy from a possible inborn error of metabolism, and was found on postmortem examination to have only a traumatic cervical SCI. There was insufficient clinical data to determine whether this patient had any limb movement or respiratory effort before death.

Among the 29 children with cervical SCIs, there were some similarities in pathological findings: 21 (72%) had parenchymal injuries, such as cord contusions, lacerations, or transections; 24 (83%) had meningeal hemorrhages; and 16 (55%) had nerve root avulsions or dorsal root ganglion hemorrhages. Five children had parenchymal injuries without meningeal hemorrhaging, 8 had meningeal hemorrhaging without parenchymal injuries, and 16 had both parenchymal and meningeal injuries. Of the 16 with nerve root avulsions or dorsal root ganglion hemorrhaging, 14 also had meningeal hemorrhages. Ten children had meningeal hemorrhaging without nerve root avulsion or dorsal root ganglion hemorrhaging.

Only 6 (21%) of 29 children with primary cervical SCIs had soft-tissue injuries to the neck. Among these, 4 had muscle, 3 had ligamentous, 2 had other soft-tissue injuries, and 3 had both muscle and ligamentous injuries. Soft-tissue injury to the neck was uncommon overall, as only 10 (19%) of 52 homicide victims had such an injury. Of these 10, 9 were victims of abusive head trauma and only 1 was the victim of another mechanism of death (asphyxiation). Among the 9 children with abusive head trauma and soft-tissue injuries to the neck, 6 had muscle injuries, 3 had ligamentous injuries, and another 3 had other soft-tissue injuries to the neck. Overall, therefore, 9 (22%) of 41 children had abusive head trauma, 1 (9%) of 11 children had other mechanisms of death (not head trauma), and 1 of 3 victims of asphyxiation had soft-tissue injuries to the neck. There were no children with cervical spine fractures or dislocations (Table 3).

We examined the relationship between the cervical spinal cord and the mechanism of abusive head trauma (impact vs no evidence of impact). We found a trend toward universal SCI in children without blunt impact; of the 6 children without visible signs of impact, all had primary cervical spinal cord and regional injuries, compared with 23 (65.7%) of the 35 with evidence of blunt trauma ($p = 0.088$).

TABLE 3: Pathological findings in the necks of infant homicide victims

Finding	MOD		
	Abusive Head Trauma (41 infants)	Cervical SCI (29 infants)	Other (11 infants)
any soft-tissue injury to neck (%)	9 (22)	6 (20.7)	1 (9.1)
muscle injury	6 (14.6)	4 (13.8)	1 (9.1)
ligamentous injury	3 (7.3)	3 (10.3)	0 (0)
other soft-tissue injury	3 (7.3)	2 (6.9)	0 (0)
cervical spine dislocation or fracture	0 (0)	0 (0)	0 (0)

The presence of primary cervical SCI was examined for any association with cerebral edema, infarction, hypoxia-ischemia, and herniation among the children with abusive head trauma. No association was found between primary cervical SCI and hypoxia-ischemia ($p = 0.853$), infarction ($p = 0.44$), or herniation ($p = 0.16$). There were insufficient clinical data to correlate these findings with clinical manifestations typically associated with SCI, such as apnea or paralysis. However, there was a significant association between primary cervical cord injury and cerebral edema ($p = 0.036$; Table 4).

Head Injuries

Thirty-seven (90%) of 41 infants and young children with abusive head trauma had intracranial meningeal hemorrhaging. This included SDHs in 34 (92%), and subarachnoid hemorrhages in 34 (92%). Epidural hemorrhage was found in 8 children (21%); all epidural hemorrhages were reported as small and none were clinically apparent. Seven of the 8 were associated with overlying skull fractures. Twenty-seven of the 41 children (66%) had evidence of intracerebral bleeding. In 1 child, the condition of the brain was such that the presence of cerebral contusions and lacerations could not be ascertained, but of the remaining 40 children, 32 (80%) had some cerebral contusion or laceration, whether superficial or deep. Twenty-six (65%) of 40 had superficial cerebral contusions and lacerations, and 23 (58%) of 40 had deep cerebral contusions and lacerations. Six children (15%) had diffuse traumatic axonal injury.

Among the 41 children with abusive head trauma, 35 (85%) had evidence of blunt head trauma. Evidence of blunt trauma included skull fractures in 13 (37%), galeal or subgaleal hematomas in 24 (69%), scalp bruising in 19 (54%), and superficial cerebral contusions and lacerations in 23 (54%).

Eleven of the 41 children (27%) with fatal abusive head trauma had extracranial fractures (such as rib or extremities). Eight of 11 (73%) had 1 or more rib fractures. Among those 8, rib fractures were the only fractures in 7 (88%), with the other child also having a femur fracture. Three of 11 children (27%) had only extremity fractures without rib fractures; these consisted of a tibial metaphyseal fracture, a clavicle fracture, and fractures of the ulna and radius. Twenty-one children (51%) had concomitant visceral injuries such as liver, splenic, or renal injuries. Twenty-seven children (66%) had facial bruising, and 20 (49%) had bruising to the rest of the body.

Of the 41 children with abusive head trauma, 27 (66%) had evidence of cerebral edema, and 23 (56%) had hypoxic-ischemic injury. Due to the physical condition of the brain specimens, determination of cerebral herniation could not be made in 1 case and determination of infarction could not be made in another. However, of the remaining children, 9 (23%) had herniations and 6 (15%) had cerebral infarctions. Cerebral edema, hypoxia-ischemia, infarction, and herniation were not specific to children who died of abusive head trauma; in fact, there was no statistically significant difference in the proportions of children with these findings by mechanism of death (Table 5).

Brainstem Injuries

Brainstem trauma was found in 16 (40%) of 40 children with abusive head trauma (1 specimen could not be examined due to poor condition). Fourteen of 16 (88%) also had primary cervical SCI.

Retinal Hemorrhages

Thirty of the 41 children (73%) with abusive head trauma had retinal hemorrhages, which were bilateral in 21 cases and unilateral in 9. Of the 11 children who died without evidence of CNS injuries, 7 had eyes available for neuropathological examination, none of which had retinal hemorrhages. There were 3 children who had both SDHs and retinal hemorrhaging but no external or other injuries.

TABLE 4: Association of primary cervical SCIs with cerebral edema, hypoxia-ischemia, infarction, and brainstem herniation among infants with abusive head trauma

Finding	Cervical SCI (29 infants)	No SCI (12 infants)	OR*	p Value
cerebral edema	22 (75.9%)	5 (41.7%)	4.4 (1.09–17.7)	0.036
brainstem herniation	8 (28.6%)	1 (8.33%)	4.4	0.160
hypoxia-ischemia	16 (55.2%)	7 (58.3%)	0.88 (0.24–3.31)	0.853
infarction	5 (17.9%)	1 (8.33%)	2.4	0.44

* Numbers in parentheses are 95% CIs.

TABLE 5: Association of cerebral edema, hypoxia-ischemia, infarction, and brainstem herniation lesions with MOD among 52 infant victims of homicide

Finding	Abusive Head Trauma (%)	Other MOD (%)	p Value
cerebral edema	27 (65.9)	7 (63.6)	0.89
brainstem herniation	9 (22.5)	1 (9.1)	0.321
hypoxia-ischemia	23 (56.1)	6 (54.6)	0.93
infarction	6 (15)	1 (9.1)	0.614

Although retinal hemorrhages were found in children who had evidence of isolated blunt abusive head trauma without any cervical cord injury, children who had both cervical cord injuries and evidence of blunt head trauma had the highest rate of retinal hemorrhages (92%; Table 6).

Discussion

In this study of homicide victims 2 years of age or younger, cervical SCIs were commonly found in those who died of abusive head trauma and occurred in those with and without visible evidence of impact injury. There have been several reports of spinal cord and neck injuries in children with abusive head trauma, but our case series is the largest and most complete analysis of these injuries reported to date.[16–18,20,23,33,38–42,44,45] Several studies of pediatric cervical cord injury of various causes have included abuse as a mechanism of trauma, but these accounted for 4% or less in most case series.[10,21,29,36,37] Currently, it seems that radiographic evaluation (primarily MR imaging) of the apparently intact neck has not been useful in identifying SCI. The presence of cervical SCI, hematomas, and nerve root damage has been documented postmortem in infants who have sustained abusive head trauma;[16–18,23,25,41] however, MR imaging failed to identify the cervical injuries among inpatients.[16]

Our findings corroborate those of others who report that the majority of infants with fatal abusive head trauma exhibit external evidence of blunt trauma[14] and that a high proportion have SDHs[10–14] and retinal hemorrhages.[14,35] A considerable number had concomitant skeletal injuries.[7,13,34]

Damage to the cervical spinal cord and roots has also been reported by others. Feldman et al.,[16] in a study of 5 victims of abusive head trauma who underwent autopsy, found that 1 had diffuse thin subdural blood (in continuity with thin cranial subdural blood) overlying the upper cervical cord, and 3 had subarachnoid blood overlying the cord (associated either with cranial subarachnoid blood or extensively distributed subarachnoid blood). Hadley et al.[23] found that 5 of 6 patients with abusive head trauma had epidural and/or SDHs of cervical spinal cord at the cervicomedullary junction and 4 of 6 patients had evidence of ventral spinal cord contusions at high cervical levels on postmortem examination. In their case series of 4 victims of abusive head trauma who underwent autopsy, Johnson and colleagues[25] found that 1 had spinal cord contusion and laceration, and another had a cervical SDH.

Additionally, Geddes et al.[17,18] and Shannon et al.[41] reported significant rates of cervical cord injury based on expression of β-amyloid precursor protein utilizing the immunoperoxidase technique in infants who sustained fatal abusive head trauma. Geddes and associates[17,18] found that 11 of 37 infants had epidural cervical hemorrhages and focal axonal damage involving the brainstem and spinal nerve roots. The findings of Shannon and colleagues[41] were even more striking, as 7 of 11 infants with no evidence of impact injury exhibited axonal injury of the cervical spinal cord. This was particularly prominent at the root entry zone.

One investigator has suggested that in fatal abusive head trauma in infants, cervical cord injury would be a universal finding because cord injury occurs at lower distraction forces than does primary cerebral injury.[1] Other investigators, however, have identified errors in these mathematical calculations.[32] While we have found that cervical cord injury is common, it is not universal. It was identified in the 6 children who had no visible evidence of

TABLE 6: Retinal pathology in infants with abusive head trauma and cervical SCIs*

Type of Injury	Retinal Pathology			
	None (%)	Minimal RH (%)	Bilat RH (%)	Retinal Detachment (%)†
BFT w/o cervical SCI (15 infants)	9 (60)	2 (13.3)	4 (26.6)	1/4 (25)
BFT & nerve root/meningeal trauma (15 infants)‡	3 (20)	1 (6.6)	11 (73)	5/11 (45)
cervical SCI (15 infants)	2 (13.3)	4 (26.6)	9 (60)	3/13 (23)
cervical SCI & nerve root trauma (2 infants)	0 (0)	0 (0)	2 (100)	1/2 (50)
cervical SCI & BFT (12 infants)	1 (8.3)	2 (16.6)	9 (75)	4/9 (44)
no cervical SCI or root trauma (16 infants)	11 (62)	0 (0)	5 (38)	1/5 (20)

* RH = retinal hemorrhage.

† The denominator represents cases in which retinal detachment was specifically noted and could be separated from artificial detachment secondary to processing of the specimen.

‡ No isolated nerve root/meningeal injury was present in the absence of blunt force trauma.

blunt impact trauma. Given results by Feldman et al.[16] that among 12 children with abusive head injury, no cervical cord injury was detected by MR imaging, these results may not be generalizable to nonfatally injured children. It may be that cervical cord injury is a marker for more severe injury.

There is currently some controversy regarding relationship of soft-tissue injuries and spinal cord trauma. Our data indicate that although cervical cord injury is common, adjacent soft-tissue injury occurs less frequently. Moreover, there was no evidence of fracture or dislocation. Absence of bone injury may perhaps be explained by the relative laxity and flexibility of spinal ligaments and musculature in the infant neck, which may be able to withstand more flexion and extension rotational forces than the spinal cord itself.

It has been postulated that cervical spinal/root injury initiates apnea, hypoxic-ischemic injury, and the subsequent death of these infant victims. Although this may be correct, there is evidence that infants with other types of neural and nonneural trauma also exhibit secondary CNS abnormalities such as hypoxic-ischemic lesions and edema. These may occur consequent to cerebral perfusion failure or biochemical abnormalities resulting from traumatic head injuries such as increased oxidative stress, which can mediate several cellular changes, any or all of which may lead to neuronal injury.[4,6,46]

The exact relationship between these commonly seen cervical injuries and clinical symptoms remains unclear. Unfortunately, we did not have adequate clinical data to correlate with our pathological findings, but it is possible that cervical SCIs may contribute significantly to apnea and other clinical correlates of abusive injuries. It is also possible that some infants have less severe SCIs that cause transient dysfunction and result in apnea or hypoventilation, but that are not visible on the postmortem examination or do not cause visible changes on MR images in survivors.

The present study is not without limitations. Our sample population is small, which limits our subgroup analysis. Clinical histories in these patients were spotty and often unavailable, so we were unable to perform any analysis of injuries in relation to the presenting complaint. Only 1 pediatric forensic neuropathologist performed the postmortem examinations and described the findings, and although this neuropathologist is very experienced, there was no second reviewer to support her findings objectively. Lastly, this study is generalizable to fatally injured children only, and the significance of these findings to the presentation and diagnosis of neck injuries in nonfatally abused children is unclear.

Future directions include correlating clinical and radiographic findings to pathological findings, as well as further exploration of the relationship and pathway between cervical SCI and death in a larger population.

Conclusions

Abusive head trauma is the most common mechanism of death among infant homicide victims. Victims exhibit a high frequency of SDH, retinal hemorrhages, and skeletal injuries. Cervical SCIs are also common, but not universal. In the present study, parenchymal and/or root injuries usually occurred without evidence of muscular or ligamentous damage, or bone dislocation or fracture. Moreover, associated brainstem injury was not always seen. Although there was a significant association of primary cervical SCI with cerebral edema, there was no direct relationship with brainstem herniation, hypoxia-ischemia, or infarction. This finding suggests that cervical spinal trauma is only 1 factor in the pathogenesis of these lesions. Future study may help determine whether SCI plays a major role in the common findings of apnea and hypoxic-ischemic brain injury in infants who have sustained abusive head trauma.

Disclaimer

The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper.

References

1. Bandak FA: Shaken baby syndrome: a biomechanics analysis of injury mechanisms. **Forensic Sci Int 151:**71–79, 2005
2. Barlow K, Thompson E, Johnson D, Minns RA: The neurological outcome of non-accidental head injury. **Pediatr Rehabil 7:**195–203, 2004
3. Barlow KM, Thompson E, Johnson D, Minns RA: Late neurologic and cognitive sequelae of inflicted traumatic brain injury in infancy. **Pediatrics 116:**e174–e185, 2005
4. Bayir H, Kochanek PM, Kagan VE: Oxidative stress in immature brain after traumatic brain injury. **Dev Neurosci 28:** 420–431, 2006
5. Benzel EC, Hadden TA: Neurologic manifestations of child abuse. **South Med J 82:**1347–1351, 1989
6. Berger RP, Adelson PD, Richichi R, Kochanek PM: Serum biomarkers after traumatic and hypoxic brain injuries: insight into the biochemical response of the pediatric brain to inflicted brain injury. **Dev Neurosci 28:**327–335, 2006
7. Biousse V, Suh DY, Newman NJ, Davis PC, Mapstone T, Lambert SR: Diffusion-weighted magnetic resonance imaging in shaken baby syndrome. **Am J Ophthalmol 133:**249–255, 2002
8. Biron D, Shelton D: Perpetrator accounts in infant abusive head trauma brought about by a shaking event. **Child Abuse Negl 29:**1347–1358, 2005
9. Bonnier C, Mesples B, Gressens P: Animal models of shaken baby syndrome: revisiting the pathophysiology of this devastating injury. **Pediatr Rehabil 7:**165–171, 2004
10. Brown RL, Brunn MA, Garcia VF: Cervical spine injuries in children: a review of 103 patients treated consecutively at a level 1 pediatric trauma center. **J Pediatr Surg 36:**1107–1114, 2001
11. Caffey J: Multiple fractures in the long bones of infants suffering from subdural hematoma. **AJR Am J Roentgenol 56:** 163–173, 1946
12. Caffey J: The whiplash shaken infant syndrome: manual shaking by the extremities with whiplash-induced intracranial and intraocular bleedings, linked with residual permanent brain damage and mental retardation. **Pediatrics 54:**396–405, 1974
13. Duhaime AC, Alario AJ, Lewander WJ, Schut L, Sutton LN, Seidl TS, et al: Head injury in very young children: mechanisms, injury types, and opthalmologic findings in 100 hospitalized patients younger than 2 years of age. **Pediatrics 90:** 179–185, 1992
14. Duhaime AC, Gennarelli TA, Thibault LE, Bruce DA, Margulies SS, Wiser R: The shaken baby syndrome: a clinical, pathological, and biomechanical study. **J Neurosurg 66:**409–415, 1987

15. Ewing-Cobbs L, Prasad M, Kramer L, Landry S: Inflicted traumatic brain injury: relationship of developmental outcome to severity of injury. **Pediatr Neurosurg 31:**251–258, 1999
16. Feldman KW, Weinberger E, Milstein JM, Fligner CL: Cervical spine MRI in abused infants. **Child Abuse Negl 21:**199–205, 1997
17. Geddes JF, Hackshaw AK, Vowles GH, Nickols CD, Whitwell HL: Neuropathology of inflicted head injury in children. I. Patterns of brain damage. **Brain 124:**1290–1298, 2001
18. Geddes JF, Vowles GH, Hackshaw AK, Nickols CD, Scott IS, Whitwell HL: Neuropathology of inflicted head injury in children. II. Microscopic brain injury in infants. **Brain 124:** 1299–1306, 2001
19. Ghatan S, Ellenbogen R: Pediatric spine and spinal cord injury after inflicted trauma. **Neurosurg Clin N Am 13:**227–233, 2002
20. Gleckman AM, Kessler SC, Smith TW: Periadventitial extracranial vertebral artery hemorrhage in a case of shaken baby syndrome. **J Forensic Sci 45:**1151–1153, 2000
21. Grabb PA, Pang D: Magnetic resonance imaging in the evaluation of spinal cord injury without radiographic abnormality in children. **Neurosurgery 35:**406–414, 1994
22. Graham DI, Lantos PL (eds): **Greenfield's Neuropathology, ed 6. Vol I.** New York: Oxford University Press, 1997
23. Hadley MN, Sonntag VK, Rekate HL, Murphy A: The infant whiplash-shake injury syndrome: a clinical and pathological study. **Neurosurgery 24:**536–540, 1989
24. Jenny C, Hymel KP, Ritzen A, Reinert SE, Hay TC: Analysis of missed cases of abusive head trauma. **JAMA 281:**621–626, 1999
25. Johnson DL, Boal D, Baule R: Role of apnea in nonaccidental head injury. **Pediatr Neurosurg 23:**305–310, 1995
26. Judkins AR, Hood IG, Mirchandani HG, Rorke LB: Rationale and technique for examination of the nervous system in suspected infant victims of abuse. **Am J Forensic Med Pathol 25:**29–32, 2004
27. Keenan HT, Ryan DK, Marshall SW, Nocera MA, Merten DF, Sinal SH: A population-based study of inflicted traumatic brain injury in young children. **JAMA 290:**621–626, 2003
28. Kemp AM, Stoodley N, Cobley C, Coles L, Kemp KW: Apnoea and brain swelling in non-accidental head injury. **Arch Dis Child 88:**472–476, 2003
29. Kewalramani LS, Kraus JF, Sterling HM: Acute spinal-cord lesions in a pediatric population: epidemiological and clinical features. **Paraplegia 18:**206–219, 1980
30. King WJ, MacKay M, Sirnick A, Canadian Shaken Baby Study Group: Shaken baby syndrome in Canada: clinical characteristics and outcomes of hospital cases. **CMAJ 168:**155–159, 2003
31. Leestma JE: Case analysis of brain-injured admittedly shaken infants: 54 cases, 1969-2001. **Am J Forensic Med Pathol 26:** 199–212, 2005
32. Margulies S, Prange M, Myers BS, Maltese MR, Ji S, Ning X, et al: Shaken baby syndrome: a flawed biomechanical analysis. **Forensic Sci Int 164:**278–279, 2006
33. McGrory BE, Fenichel GM: Hangman's fracture subsequent to shaking in an infant. **Ann Neurol 2:**82, 1977
34. Merten DF, Osborne DRS, Radkowski MA, Leonidas JC: Craniocerebral trauma in the child abuse syndrome: radiological observations. **Pediatr Radiol 14:**272–277, 1984
35. Munger CE, Peiffer RL, Bouldin TW, Kylstra JA, Thompson RL: Ocular and associated neuropathologic observations in suspected whiplash shaken infant syndrome. **Am J Forensic Med Pathol 14:**193–200, 1993
36. Pang D, Pollack IF: Spinal cord injury without radiographic abnormality in children—the SCIORWA syndrome. **J Trauma 29:**654–664, 1989
37. Pang D, Wilberger JE: Spinal cord injury without radiographic abnormalities in children. **J Neurosurg 57:**114–129, 1982
38. Piatt JH Jr, Steinberg M: Isolated spinal cord injury as a presentation of child abuse. **Pediatrics 96:**780–782, 1995
39. Ranjith RK, Mullett JH, Burke TE: Hangman's fracture caused by suspected child abuse. A case report. **J Pediatr Orthop B 11:**329–332, 2002
40. Rooks VJ, Sisler C, Burton B: Cervical spine injury in child abuse: report of two cases. **Pediatr Radiol 28:**193–195, 1998
41. Shannon P, Smith CR, Deck J, Ang LC, Ho M, Becker L: Axonal injury and the neuropathology of shaken baby syndrome. **Acta Neuropathol 95:**625–631, 1998
42. Sneed RC, Stover SL: Undiagnosed spinal cord injuries in brain-injured children. **Am J Dis Child 142:**965–967, 1988
43. Starling SP, Patel S, Burke BL, Sirotnak AP, Stronks S, Rosquist P: Analysis of perpetrator admissions to inflicted traumatic brain injury in children. **Arch Pediatr Adolesc Med 158:**454–458, 2004
44. Swischuk LE: Spine and spinal cord trauma in the battered child syndrome. **Radiology 92:**733–738, 1969
45. Thomas NH, Robinson L, Evans A, Bullock P: The floppy infant: a new manifestation of nonaccidental injury. **Pediatr Neurosurg 23:**188–191, 1995
46. Wagner AK, Bayir H, Ren D, Puccio A, Zafonte RD, Kochanek PM: Relationship between cerebrospinal fluid markers of excitotoxicity, ischemia, and oxidative damage after severe TBI: the impact of gender, age, and hypothermia. **J Neurotrauma 21:**125–136, 2004

Manuscript submitted March 19, 2008.
Accepted November 21, 2008.
Address correspondence to: Laura K. Brennan, M.D., Children's Hospital of Philadelphia, Division of General Pediatrics, 34th Street and Civic Center Boulevard, Philadelphia, Pennsylvania 19104. email: brennanl@email.chop.edu.

Current Concepts

Nonaccidental Head Injury in Infants — The "Shaken-Baby Syndrome"

Ann-Christine Duhaime, M.D.,
Cindy W. Christian, M.D.,
Lucy Balian Rorke, M.D.,
and Robert A. Zimmerman, M.D.

TRAUMA is the most common cause of death in childhood, and inflicted head injury is the most common cause of traumatic death in infancy.[1-3] Beginning with the classic descriptions of Kempe et al.[4] and Caffey[5] and with subsequent clinical, biomechanical, and radiologic studies, the diagnostic features of nonaccidental head injury in infants and toddlers have become widely recognized. This review outlines the mechanisms, typical features, differential diagnosis, and acute management of the most frequently encountered form of infantile inflicted head injury, the so-called shaken-baby syndrome.

BIOMECHANICS AND TERMINOLOGY

The names applied to the syndromes of inflicted head injury in infancy reflect the evolving and sometimes controversial understanding of the actions necessary to cause the types of injuries seen, such as shaking an infant held by the arms or trunk or forcefully striking an infant's head against a surface. Although there is considerable controversy, the available evidence suggests that it is the sudden deceleration associated with the forceful striking of the head against a surface that is responsible for most, if not all, severe, inflicted brain injuries. Because the histories given when infants with such injuries present for medical attention are often vague or unreliable, the events must be inferred from knowledge of the causative forces in witnessed cases of accidental trauma and experimental models of injury. Studies of the biomechanics of brain injury have established that forces applied to the head that result in a rotation of the brain about its center of gravity cause diffuse brain injuries. It is this type of movement that is responsible for the diffuse axonal injury and subdural hematoma seen, for example, in cases of motor vehicle accidents that result in severe disability or death. In contrast, forces that result in a translation, or straight-line, movement of the center of gravity are generally less injurious to the brain, with the effects largely determined by the specific focal contact forces.[6] The type and severity of the injury are determined both by the type of deceleration and by its magnitude. In infants and young children, household falls causing head injuries mainly involve low-velocity translational forces; rotational (or angular) deceleration is distinctly uncommon.[3]

The term "whiplash shaken-baby syndrome" was coined by Caffey to explain the constellation of infantile subdural and subarachnoid hemorrhage, traction-type metaphyseal fractures, and retinal hemorrhages and was based on evidence that angular (rotational) deceleration is associated with cerebral concussion and subdural hematoma.[7-10] Because the type but not the magnitude of deceleration was addressed in early reports of the syndrome, it was postulated that injuries could be inflicted unwittingly by caretakers through generally acceptable child care practices. More recent biomechanical studies of these injuries show that the magnitude of angular deceleration is 50 times as great when the head of an infant model held by the trunk forcefully strikes a surface as when shaking alone occurs, and it only reaches injury thresholds calculated for infants at the moment of impact. When the surface is soft, the force of the impact is widely dissipated and may not be associated with visible signs of surface trauma, even though the brain itself decelerates rapidly.[11] It is the sudden angular deceleration experienced by the brain and cerebral vessels, not the specific contact forces applied to the surface of the head, that results in the intracranial injury. This angular force is distinct from the forces generated in most cases of accidental trauma in infants. The majority of abused infants in fact have clinical, radiologic, or autopsy evidence of blunt impact to the head.[11-13] Thus, the term "shaking–impact syndrome" may reflect more accurately than "shaken-baby syndrome" the usual mechanism responsible for these injuries.[14] Whether shaking alone can cause the constellation of findings associated with the syndrome is still debated, but most investigators agree that trivial forces, such as those involving routine play, infant swings, or falls from a low height are insufficient to cause the syndrome. Instead, these injuries appear to result from major rotational forces, which clearly exceed those encountered in normal child-care activities.[3,13,15-19]

EPIDEMIOLOGY

The shaking–impact syndrome is largely restricted to children under three years of age, with the majority of cases occurring during the first year of

From the Divisions of Neurosurgery (A.-C.D.), General Pediatrics (C.W.C.), Neuropathology (L.B.R.), and Neuroradiology (R.A.Z.), Children's Hospital of Philadelphia and the University of Pennsylvania School of Medicine, Philadelphia. Address reprint requests to Dr. Duhaime at Neurosurgery, Children's Hospital of Philadelphia, 34th and Civic Center Blvd., Philadelphia, PA 19104.

Reprinted from *The New England Journal of Medicine*
338:1822-1829 (June 18), 1998

life.[11,20,21] In a prospective study of consecutively admitted children under two years of age who had head injuries, 24 percent of the injuries resulted from inflicted trauma; among infants with severe injuries, the proportion was even higher.[2,3,22] Frequently, such children have evidence of previous abuse.[13] At our institution, more traumatic deaths result from child abuse involving head injuries than from any other single cause.

Risk factors for nonaccidental injuries in children include young parents, unstable family situations, low socioeconomic status, and disability or prematurity of the child.[23,24] Starling et al. found that the perpetrators were, in descending order of frequency, fathers, boyfriends, female babysitters, and mothers.[25]

HISTORY, PHYSICAL EXAMINATION, AND LABORATORY FINDINGS

With inflicted head injuries, an accurate history is rarely provided at presentation. The information most commonly reported involves the child's symptoms or a history of blunt impact to the head, usually of a minor nature.[11,20] A history of shaking is obtained in a minority of cases.[11-13,20] The history may be vague or may vary with time, or a mechanism of injury that is incompatible with the developmental capacity of the child may be described.

Common symptoms include lethargy, irritability, seizures, increased or decreased tone, impaired consciousness, vomiting, poor feeding, breathing abnormalities, and apnea. Milder neurologic findings include lethargy, irritability, and meningismus. Approximately half of all patients with the shaking–impact syndrome have severe impairment, are unresponsive, have opisthotonos, or are moribund.[12] The fontanelle may be full. Seizures are reported in 40 to 70 percent of patients.[20,26]

Retinal hemorrhages, best seen with the use of mydriatic agents, are found in 65 to 95 percent of patients.[2,20,21,27,28] The hemorrhages may be unilateral or bilateral, and retinal folds or detachments may be seen. The exact biomechanical forces necessary to cause retinal hemorrhages are unknown, but several mechanisms have been postulated, including increased retinal venous pressure, extravasation of subarachnoid blood, and traction of retinal vessels at the vitreoretinal interface due to angular deceleration.[29,30] Although strongly associated with inflicted head injury, retinal hemorrhages are not specific for the diagnosis, nor can they be dated with precision. Such hemorrhages have been reported in some cases of accidental trauma (especially subdural hematoma) and, in rare cases, after resuscitation; they can also occur with papilledema.[3,27,31,32] Retinal hemorrhages are seen in up to 40 percent of vaginally delivered newborns but resolve by one month of age.[33] Nontraumatic causes include subarachnoid hemorrhage, sepsis, coagulopathy, galactosemia, severe hypertension, and other rare conditions.[34-37] The diagnosis of inflicted head injury cannot rest on the finding of retinal hemorrhage alone, but the finding of severe bilateral retinal hemorrhage with retinal folds or detachments is particularly suggestive of the diagnosis.

General physical findings may include bruising, swelling, a pattern of cutaneous marks, and burns. In some patients no extracranial injuries are detected. Some cutaneous injuries become visible only after admission. In some patients, soft-tissue injuries, including scalp hemorrhages, are noted only at autopsy.[11,13]

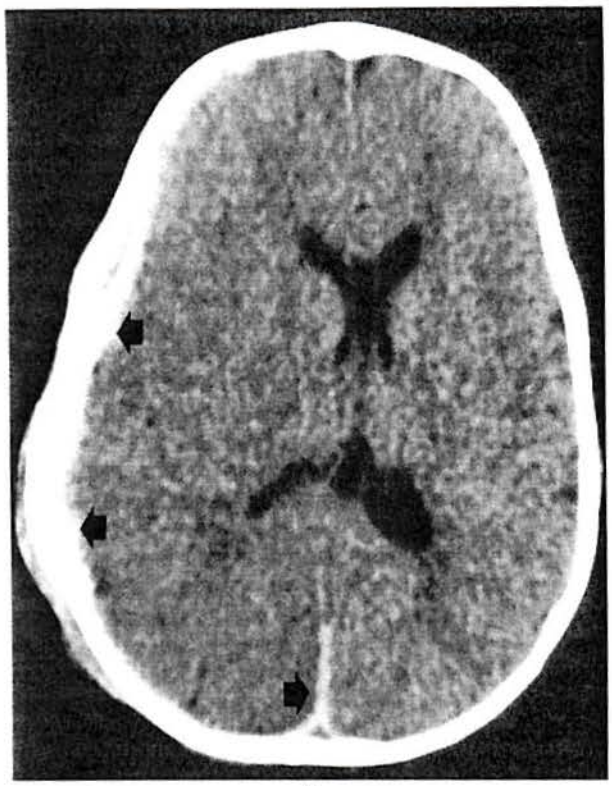
Lumbar puncture, typically performed as part of an evaluation for sepsis in infants with nonspecific findings, reveals bloody fluid. Hemoglobin values may be decreased.[21] Elevated coagulation factors do not necessarily indicate a primary coagulopathy but may reflect the underlying brain injury.[38]

RADIOLOGIC FINDINGS

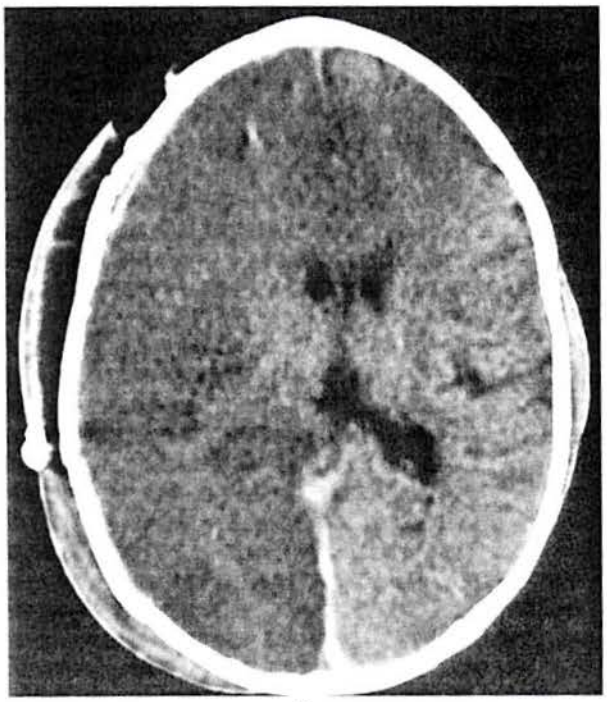
Computed tomographic (CT) scanning is the mainstay of the diagnosis of the shaking–impact syndrome. Subdural or subarachnoid hemorrhage can nearly always be detected on CT scans, although the more subtle findings may be missed by less experienced observers. Hemorrhages most often appear as unilateral or bilateral high-density collections of fresh blood that are thin but extensive; a particular propensity for the interhemispheric fissure, especially posteriorly, is well documented.[39,40]

A peculiar and poorly understood CT finding that is uniquely associated with subdural hematoma in infancy is extensive loss of gray–white differentiation and diffuse hypodensity. This finding can be unilateral or bilateral. The basal ganglia and posterior fossa structures are relatively spared and thus appear hyperdense as compared with the surrounding cerebrum, which is abnormally hypodense ("reversal sign").[41] In unilateral cases, an additional wedge-shaped area of hypodensity in the contralateral frontal lobe, probably reflecting subfalcine herniation, is usually noted (Fig. 1B). Diffuse hypodensity is not always apparent on the initial CT scan (Fig. 1A) but appears within the first few days in infants with severe neurologic symptoms (i.e., unresponsiveness).[42] This finding is not specific for abuse, but since abuse is the most common cause of subdural hematoma in infancy, it is seen most often in association with abuse.

Magnetic resonance imaging (MRI) is useful in detecting and characterizing small extraaxial hemorrhages in infants with equivocal CT findings. The identification of parenchymal contusions on MRI scans may also be helpful in differentiating the shaking–impact syndrome from the rare case of spontaneous subarachnoid hemorrhage (Fig. 2).[43] Soft-tissue swelling may be noted on CT scans, MRI scans, or plain skull films. Plain films are superior to CT



A



B

Figure 1. Axial Cranial CT Scans in an Eight-Month-Old Unresponsive Boy Found at Home.

A skeletal survey showed fractures of the skull, multiple ribs, arms, and legs. The initial CT scan (Panel A), obtained without the administration of contrast material, shows a right-convexity subdural hematoma extending onto the posterior falx cerebri (arrows). There is mass effect with a midline shift. A scan obtained 24 hours after surgical evacuation of the hematoma (Panel B) shows hypodensity throughout the right cerebral hemisphere, involving both the cortex and the white matter. The contralateral anterior frontal lobe is also characterized by decreased density. Blood remains visible along the falx cerebri.

scans for the detection of skull fractures, which are found most commonly in the occipital or parietooccipital regions. Multiple or complex skull fractures have been associated with abuse.[40,44,45]

A skeletal survey is essential in the evaluation of a child for the shaking–impact syndrome, since extracranial abnormalities are detected in 30 to 70 percent of abused children with head injuries.[40,46] A wide variety of skeletal injuries have been described. Although none are strictly pathognomonic of abuse, multiple posterior or lateral rib fractures and metaphyseal fractures are characteristic. In some patients, delayed repeated films or radionuclide bone scans are necessary to detect sites of subtle injury.[47,48]

Some infants with previous inflicted injuries present with chronic subdural hematomas, although data from studies of such infants are scarce. In Parent's series, 44 percent of infants with chronic subdural hematomas were thought to have sustained previous inflicted injuries.[49] The diagnosis in this population rests largely on the finding of unexplained skeletal or other injuries indicative of abuse. Treatment of symptomatic chronic subdural hematoma usually includes surgical drainage or shunting.

INITIAL MANAGEMENT, CLINICAL COURSE, AND OUTCOME

The initial treatment of infants with markedly impaired consciousness includes intubation, ventilation, fluid resuscitation, and anticonvulsant therapy. Surgical evacuation should be considered in the case of a large acute hematoma.[50,51] The value of aggressive management of intracranial hypertension has been questioned on the basis of outcome studies, which show that infants who present with poor prognostic indicators, especially bilateral diffuse hypodensity on CT scans, have dismal outcomes regardless of treatment. Less severely injured infants are treated with anticonvulsant agents and closely observed; recovery is variable in such cases.[26,52]

In infants who succumb, the cause of death is uncontrollable intracranial hypertension. Remarkable cortical and white-matter atrophy is seen consistently on follow-up neuroradiologic studies (Fig. 3) in

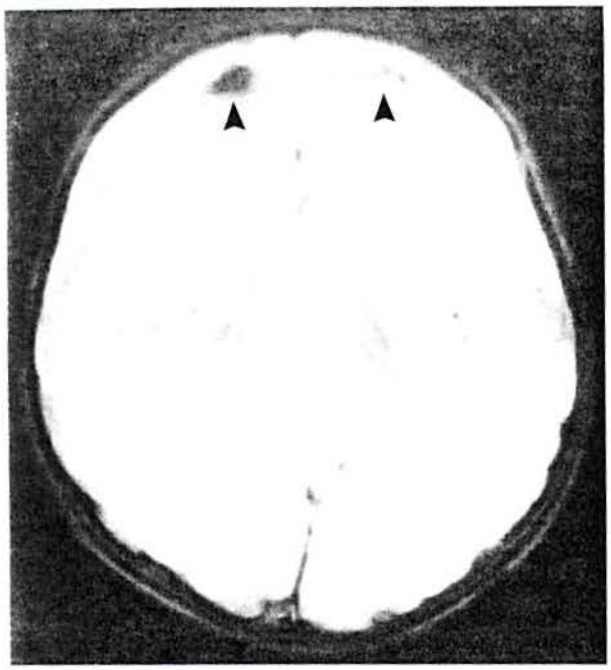


Figure 2. T_2-Weighted Axial Cranial MRI Scan in a Four-Month-Old Girl Reported to Have Fallen from a Low Height.

Gradient-echo imaging was used with the technique of the fast low-angle shot (FLASH) to demonstrate blood products. Areas of acute hemorrhage on both frontal cortical surfaces can be seen (arrowheads), along with proteinaceous extraaxial collections. This infant also had multiple rib fractures.

survivors with diffuse hypodensity during the acute period.

TIMING OF THE INJURY

Since the history is often unreliable in cases of the shaking–impact syndrome, information about the timing of the injury must be extrapolated from data on accidental trauma. Acute subdural hematoma associated with severe neurologic compromise, brain swelling, or death occurs in the setting of a clear injury involving a major mechanical force and is followed by the immediate or rapid onset of neurologic symptoms.[53] In a series of 95 children who died from accidental head injuries, all but 1 of the children had an immediate decrease in the level of consciousness; the exception was a patient with an expanding epidural hematoma.[54] This type of injury, generated by contact forces to the skull and dura, is usually not associated with a serious primary brain injury and is rarely associated with child abuse.[18] Other reports of delayed deterioration after pediatric head injury have primarily involved the onset of seizures, followed by recovery.[55] On the basis of these data, it can be discerned that there is no evidence of a prolonged interval of lucidity between the injury and the onset of symptoms in children with acute subdural hematoma and brain swelling — the injuries also seen in severe cases of the shaking–impact syndrome (i.e., those associated with coma or death). Thus, an alert, well-appearing child has not already sustained a devastating acute injury that will become clinically obvious hours to days later.

The timing of the traumatic event is more difficult to establish in patients with mild neurologic injuries and is determined on the basis of general physical and radiologic findings. These methods can indicate only a general time frame.

A separate issue concerns the possibility of a subclinical injury that is later exacerbated by a relatively minor second mechanical trauma. Such rare events have been reported in older children and adults, usually in the setting of acute subarachnoid and subdural hemorrhage and brain swelling related to recurrent impact to the head involving well-documented concussive forces during sports activities.[56,57] This pattern of injury, with a clear time line and rapid, well-described acute deterioration, stands in sharp contrast to the vague histories of previous episodes of trivial trauma that are sometimes suggested as possibly causative in the shaking–impact syndrome. There is no evidence that traumatic acute subdural hematoma, particularly that leading to death, occurs in otherwise healthy infants in an occult or subclinical manner.

AUTOPSY FINDINGS

Although there are some variations, pathological findings in infants who have been shaken and battered are remarkably consistent. Evidence of external injury has been found in up to 85 percent of such infants and is most often located in the head and neck. Scalp trauma is sometimes visible only after the hair has been shaved. Autopsy detects fractures in 25 percent of affected infants.[58] Fractures involving the skull are most common in the posterior parietal bone or occipital bone or both.

Subdural hemorrhage, usually localized at the parieto-occipital convexity or posterior interhemispheric fissure, is the most consistent autopsy finding in shaking–impact syndrome.[11,21] Such hemorrhages typically range from 2 to 15 ml in volume and almost never cause death because of direct mass effect.[58] In most fatal cases, the hemorrhage is acute and involves liquid blood or a small clot resembling currant jelly.

With fatal injuries, estimates of the time at which the injury occurred rely on clinical, radiologic, and postmortem findings. Hirsch has provided guidelines for determining the age of a subdural hematoma on the basis of its gross features and microscopical characteristics.[59] Various factors limit the reliability of these methods; for example, reduced cerebral blood flow may impede the cellular re-

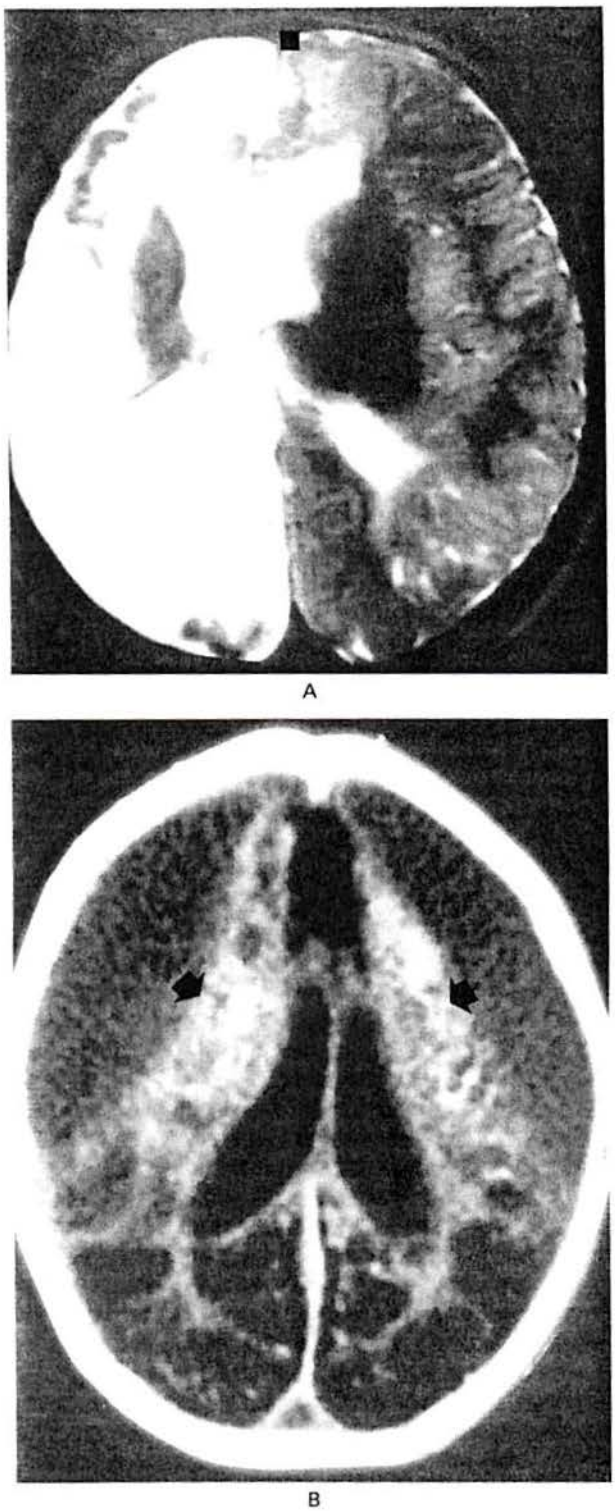


Figure 3. Follow-up Brain Images in Two Infants Injured at Four Months of Age.

A T_2-weighted axial MRI scan in one infant (Panel A) shows severe encephalomalacia involving the entire right cerebral hemisphere, with subcortical cystic changes. The ventricles are enlarged because of atrophy. The left frontal lobe is atrophic, whereas the basal ganglia and the remainder of the left hemisphere are relatively spared. An axial CT scan in the other infant (Panel B) shows bilateral diffuse encephalomalacia involving the supratentorial cortical and subcortical regions, with cysts and calcifications. The markedly atrophic brain (arrows) is surrounded by proteinaceous subdural fluid.

sponse. Iron staining must be performed to detect hemosiderin, if previous or old hemorrhage is suspected. Most infants who die within a few days after presentation have no evidence of organization of the hematoma.

Superficial contusions are most frequent in the olfactory bulbs and tracts and underlying gyrus rectus.[58] Of greater mechanical importance are gliding contusions, tears of the corpus callosum, and diffuse axonal injury, which result from extreme rotational force.[60-62] Occasionally, the ventricular wall, the vein of Galen, or even the vertebral artery may be torn. The rostral brain stem may be damaged as a consequence of the forces of angular deceleration.[63] Cerebral edema is common in infants who survive for hours or days, and necrosis is often observed.

Acute hemorrhage along the sheath of the optic nerve is typically most obvious at the junction of the nerve and the globe. Retinal hemorrhages occupy any or all layers of the retina and may be preretinal or subretinal as well. Occasionally, large vitreous hemorrhages are present.[58,64,65]

Careful dissection of the cervical region is essential. The prosector must remove the brain and spinal cord in continuity, since the most common site of cervical injury is C1 to C4. Tissue sampling for microscopical study should include typical sites of diffuse axonal injury, with the realization that gross hemorrhage may be absent. The exact time course for the development of axonal retraction balls during the initial hours after injury is a matter of debate.[66-68]

Some affected infants survive with intellectual and neurologic deficits (including blindness) for weeks, months, or years after the injury. The findings in these cases include well-demarcated cavities, primarily in the frontal lobes, representing the residua of the gliding contusions; more widespread cystic or noncystic gray-matter damage; scars in the centrum ovale, corpus callosum, or both; and chronic retinal damage with secondary optic-nerve degeneration.

PATHOPHYSIOLOGIC FACTORS

The causes of the severe brain swelling and subsequent extreme tissue loss in infants with the shaking–impact syndrome who survive are incompletely

understood and are unique to this age group. Most accidental subdural hematomas in infants are caused by motor vehicle collisions or falls from substantial heights, but in these cases, both diffuse brain swelling and fatal outcomes have been reported.[3,19,27,69]

Johnson et al. have suggested that when crying infants are shaken until apnea renders them silent, hypoxia is the primary pathophysiologic event.[26] However, the finding of unilateral hypodensity in one third of cases suggests that global hypoxia is not the only factor, nor does the pattern of delayed atrophy match that seen in survivors of isolated hypoxic injury from other causes. Cervical trauma has been reported on the basis of autopsy findings in cases of inflicted head injuries, although clinical signs of spinal cord injury are rare.[21,70] Vessel occlusion, perhaps resulting from concomitant strangulation, has been suggested to explain the cases of more unilateral tissue loss.[71] However, this explanation is rarely borne out by the findings on MRI angiography or autopsy.[58]

The most consistent finding in cases of the shaking–impact syndrome is the presence of subdural and subarachnoid blood. Hemorrhage therefore is both a marker for the threshold of force required to cause the injury and a likely pathophysiologic contributor to the resultant brain damage. It thus appears that some combination of mechanical trauma, hemorrhage, hypoxia, and possibly seizure activity overwhelms the compensatory mechanisms of the immature brain, resulting in massive swelling and widespread neuronal loss. A further understanding of these processes will require more scrutiny and better experimental models.[72-75]

DIFFERENTIAL DIAGNOSIS

No other medical condition fully mimics all the features of the shaking–impact syndrome. Several patterns of clinical and radiographic findings allow a definitive diagnosis. These include a history of trivial or no trauma, acute subdural hemorrhage, and unexplained extracranial bony injuries or clearly inflicted soft-tissue injuries; and a definite history of no possibility of trauma with clear physical or radiologic evidence of head impact with subdural hemorrhage. Although not necessary for the diagnosis, the findings of retinal hemorrhages or multiple fractures in different stages of healing make the diagnosis more certain.[3,76] It is clear that some suspicious cases will have insufficient findings with which to make a firm diagnosis. An algorithm has been developed to help differentiate injuries that can be assumed to be inflicted from those that are suspicious, for the purpose of classifying cases in clinical research. The results with this algorithm are closely correlated with the determinations of the child-abuse team at our institution.[3]

The single most common diagnosis mimicking nonaccidental trauma is accidental injury. Small epidural hemorrhages and traumatic subarachnoid hemorrhages can be mistaken for subdural hematomas; MRI may be helpful in these instances.[77] Accidental subdural hemorrhages have been reported in infants after motor vehicle collisions or falls involving substantial angular deceleration.[19,27,69,77] Infants with enlarged extraaxial spaces, such as may be seen in some cases of shunted hydrocephalus, appear to be at increased risk for subdural or subarachnoid hemorrhage with lesser degrees of trauma.[78] In cases of accidental head injury, the history is clear and consistent, the infant's symptoms reflect the forces described, and no unexplained skeletal injuries are identified.

A variety of coagulopathies are associated with intracranial hemorrhage in infants, including hemophilia and hypoprothrombinemia caused by vitamin K deficiency.[79,80] These disorders are suggested by the clinical history, physical findings, and laboratory tests. Transient prolongation of the prothrombin time and disseminated intravascular coagulopathy have been associated with the presence of parenchymal brain injury in infants with accidental trauma and in those with inflicted trauma.[38,81] Recommended screening tests include assessment of the platelet count, prothrombin time, activated partial-thromboplastin time, and bleeding time; abnormal values merit further evaluation.

Osteogenesis imperfecta is a rare inherited disorder of connective tissue that results from an abnormal quantity or quality of type I collagen. This disorder is usually readily distinguished from injuries caused by child abuse, although the physical features of osteogenesis imperfecta may be subtle. In addition to fractures, suggestive findings include blue sclerae, hearing impairment, dentinogenesis imperfecta, hypermobility of the joints, bruising, short stature, radiographic evidence of wormian bones, osteopenia, bowing and angulation of healed fractures, and progressive scoliosis. Although uncertainty about the diagnosis is usually related to unexplained skeletal injuries, subdural hemorrhage is a rare complication of the disease.[82] Fractures associated with osteogenesis imperfecta usually involve the diaphyses of long bones, but rib fractures, fractures of varying ages, and in rare cases, metaphyseal fractures can occur.[83,84] Biochemical analysis of cultured skin fibroblasts is diagnostic in approximately 85 percent of patients with the disease.[85] Clinical and radiographic evaluation by an experienced examiner is usually sufficient to distinguish osteogenesis imperfecta from injuries caused by child abuse, with biochemical testing reserved for cases in which the diagnosis is uncertain.[84]

Glutaric aciduria type I is a metabolic disorder caused by a defect of glutaryl–coenzyme A dehydrogenase. The onset of clinical symptoms may be acute or insidious, and the findings may include developmental delay, hypotonia, dyskinesia, cortical at-

rophy, and subdural collections.[86] Skeletal injuries and retinal hemorrhages have not been described as part of the disease. Urinary screening for this disorder should be considered in infants with appropriate clinical findings.

CONCLUSIONS

The shaking–impact syndrome is a common, serious injury resulting from major mechanical forces. If the history and the physical and radiologic findings are suggestive of this diagnosis, the patient should be admitted to the hospital for treatment. A thorough, unbiased evaluation is essential. If abuse is suspected, the law requires that the appropriate child-welfare and law-enforcement agencies be notified. Caretakers should be informed, in a nonaccusatory manner, that the diagnosis is suspected and that investigative procedures will be necessary for the welfare of the child. The medical record has great legal importance, and careful documentation will later benefit the physician, who may be subpoenaed to testify in court.

The future safety of a child with the shaking–impact syndrome rests on the physician's ability to recognize its characteristic features. Effective prevention strategies must be guided by an improved understanding of the pathophysiology and causes of this common disorder.

REFERENCES

1. Centers for Disease Control. Childhood injuries in the United States. Am J Dis Child 1990;144:627-46.
2. Billmire ME, Myers PA. Serious head injury in infants: accident or abuse? Pediatrics 1985;75:340-2.
3. Duhaime AC, Alario AJ, Lewander WJ, et al. Head injury in very young children: mechanisms, injury types, and ophthalmologic findings in 100 hospitalized patients younger than 2 years of age. Pediatrics 1992;90:179-85.
4. Kempe CH, Silverman FN, Steele BF, Droegemueller W, Silver HK. The battered-child syndrome. JAMA 1962;181:17-24.
5. Caffey J. On the theory and practice of shaking infants: its potential residual effects of permanent brain damage and mental retardation. Am J Dis Child 1972;124:161-9.
6. Gennarelli TA, Thibault LE. Biomechanics of head injury. In: Wilkins RH, Rengachary SS, eds. Neurosurgery. Vol. 2. New York: McGraw-Hill, 1985:1531-6.
7. Caffey J. The whiplash shaken infant syndrome: manual shaking by the extremities with whiplash-induced intracranial and intraocular bleedings, linked with residual permanent brain damage and mental retardation. Pediatrics 1974;54:396-403.
8. Ommaya AK, Faas F, Yarnell P. Whiplash injury and brain damage: an experimental study. JAMA 1968;204:285-9.
9. Ommaya AK, Gennarelli TA. Cerebral concussion and traumatic unconsciousness: correlation of experimental and clinical observations on blunt head injuries. Brain 1974;97:633-54.
10. Guthkelch AN. Infantile subdural hematoma and its relationship to whiplash injuries. BMJ 1971;2:430-1.
11. Duhaime AC, Gennarelli TA, Thibault LE, Bruce DA, Margulies SS, Wiser R. The shaken baby syndrome: a clinical, pathological, and biomechanical study. J Neurosurg 1987;66:409-15.
12. Hahn YS, Raimondi AJ, McLone DG, Yamanouchi Y. Traumatic mechanisms of head injury in child abuse. Childs Brain 1983;10:229-41.
13. Alexander R, Sato Y, Smith W, Bennett T. Incidence of impact trauma with cranial injuries ascribed to shaking. Am J Dis Child 1990;144:724-6.
14. Bruce DA, Zimmerman RA. Shaken impact syndrome. Pediatr Ann 1989;18:482-4.
15. Gilliland MGF, Folberg R. Shaken babies — some have no impact injuries. J Forensic Sci 1996;41:114-6.
16. Helfer RE, Slovis TL, Black MB. Injuries resulting when small children fall out of bed. Pediatrics 1977;60:533-5.
17. Hanigan WC, Peterson RA, Njus G. Tin ear syndrome: rotational acceleration in pediatric head injuries. Pediatrics 1987;80:618-22.
18. Shugerman RP, Paez A, Grossman DC, Feldman KW, Grady MS. Epidural hemorrhage: is it abuse? Pediatrics 1996;97:664-8.
19. Reiber GD. Fatal falls in childhood: how far must children fall to sustain fatal head injury? Report of cases and review of the literature. Am J Forensic Med Pathol 1993;14:201-7.
20. Ludwig S, Warman M. Shaken baby syndrome: a review of 20 cases. Ann Emerg Med 1984;13:104-7.
21. Hadley MN, Sonntag VKH, Rekate HL, Murphy A. The infant whiplash-shake injury syndrome: a clinical and pathological study. Neurosurgery 1989;24:536-40.
22. Goldstein B, Kelly MM, Bruton D, Cox C. Inflicted versus accidental head injury in critically injured children. Crit Care Med 1993;21:1328-32.
23. Klein M, Stern L. Low birth weight and the battered child syndrome. Am J Dis Child 1971;122:15-8.
24. Sills JA, Thomas LJ, Rosenbloom L. Non-accidental injury: a two-year study in central Liverpool. Dev Med Child Neurol 1977;19:26-33.
25. Starling SP, Holden JR, Jenny C. Abusive head trauma: the relationship of perpetrators to their victims. Pediatrics 1995;95:259-62.
26. Johnson DL, Boal D, Baule R. Role of apnea in nonaccidental head injury. Pediatr Neurosurg 1995;23:305-10.
27. Luerssen TG, Huang JC, McLone DG, et al. Retinal hemorrhages, seizures, and intracranial hemorrhages: relationships and outcomes in children suffering traumatic brain injury. In: Marlin AE, ed. Concepts in pediatric neurosurgery. Vol. 11. Basel, Switzerland: Karger, 1991:87-94.
28. Harcourt B, Hopkins D. Ophthalmic manifestations of the battered-baby syndrome. BMJ 1971;3:398-401.
29. Greenwald MJ, Weiss A, Oesterle CS, Friendly DS. Traumatic retinoschisis in battered babies. Ophthalmology 1986;93:618-25.
30. Massicotte SJ, Folberg R, Torczynski E, Gilliland MGF, Luckenbach MW. Vitreoretinal traction and perimacular retinal folds in the eyes of deliberately traumatized children. Ophthalmology 1991;98:1124-7.
31. Goetting MG, Sowa B. Retinal hemorrhage after cardiopulmonary resuscitation in children: an etiologic reevaluation. Pediatrics 1990;85:585-8.
32. Gilliland MGF, Luckenbach MW. Are retinal hemorrhages found after resuscitation attempts? A study of the eyes of 169 children. Am J Forensic Med Pathol 1993;14:187-92.
33. Baum JD, Bulpitt CJ. Retinal and conjunctival haemorrhage in the newborn. Arch Dis Child 1970;45:344-9.
34. Wetzel RC, Slater AJ, Dover GJ. Fatal intramuscular bleeding misdiagnosed as suspected nonaccidental injury. Pediatrics 1995;95:771-3.
35. Levy HL, Brown AE, Williams SE, de Juan E Jr. Vitreous hemorrhage as an ophthalmic complication of galactosemia. J Pediatr 1996;129:922-5.
36. McLellan NJ, Prasad R, Punt J. Spontaneous subhyaloid and retinal haemorrhages in an infant. Arch Dis Child 1986;61:1130-2.
37. Skalina MEL, Annable WL, Kliegman RM, Fanaroff AA. Hypertensive retinopathy in the newborn infant. J Pediatr 1983;103:781-6.
38. Hymel KP, Abshire TC, Luckey DW, Jenny C. Coagulopathy in pediatric abusive head trauma. Pediatrics 1997;99:371-5.
39. Zimmerman RA, Bilaniuk LT, Bruce D, Schut L, Uzzell B, Goldberg HI. Computed tomography of craniocerebral injury in the abused child. Radiology 1979;130:687-90.
40. Merten DF, Osborne DRS, Radkowski MA, Leonidas JC. Craniocerebral trauma in the child abuse syndrome: radiological observations. Pediatr Radiol 1984;14:272-7.
41. Han BK, Towbin RB, De Courten-Myers G, McLaurin RL, Ball WS Jr. Reversal sign on CT: effect of anoxic/ischemic cerebral injury in children. AJNR Am J Neuroradiol 1989;10:1191-8.
42. Duhaime AC, Bilaniuk L, Zimmerman R. The "big black brain": radiographic changes after severe inflicted head injury in infancy. J Neurotrauma 1993;10:Suppl 1:S59. abstract.
43. Sato Y, Yuh WTC, Smith WL, Alexander RC, Kao SCS, Ellerbrock CJ. Head injury in child abuse: evaluation with MR imaging. Radiology 1989;173:653-7.
44. Cohen RA, Kaufman RA, Myers PA, Towbin RB. Cranial computed tomography in the abused child with head injury. AJR Am J Roentgenol 1986;146:97-102.
45. Meservy CJ, Towbin R, McLaurin RL, Myers PA, Ball W. Radiographic characteristics of skull fractures resulting from child abuse. AJR Am J Roentgenol 1987;149:173-5.
46. Lazoritz S, Baldwin S, Kini N. The whiplash shaken infant syndrome: has Caffey's syndrome changed or have we changed his syndrome? Child Abuse Negl 1997;21:1009-14.

47. Kleinman PK. Diagnostic imaging in infant abuse. AJR Am J Roentgenol 1990;155:703-12.
48. Smith FW, Gilday DL, Ash JM, Green MD. Unsuspected costo-vertebral fractures demonstrated by bone scanning in the child abuse syndrome. Pediatr Radiol 1980;10:103-6.
49. Parent AD. Pediatric chronic subdural hematoma: a retrospective comparative analysis. Pediatr Neurosurg 1992;18:266-71.
50. Cho DY, Wang YC, Chi CS. Decompressive craniotomy for acute shaken/impact baby syndrome. Pediatr Neurosurg 1995;23:192-8.
51. Duhaime AC, Sutton LN, Christian C. Child abuse. In: Youmans JR, ed. Neurological surgery: a comprehensive reference guide to the diagnosis and management of neurosurgical problems. 4th ed. Vol. 3. Philadelphia: W.B. Saunders, 1996:1777-91.
52. Duhaime AC, Christian CW, Moss E, Seidl T. Long-term outcome in infants with the shaking-impact syndrome. Pediatr Neurosurg 1996;24:292-8.
53. Gennarelli TA. Head injury in man and experimental animals: clinical aspects. Acta Neurochir Suppl (Wien) 1983;32:1-13.
54. Willman KY, Bank DE, Senac M, Chadwick DL. Restricting the time of injury in fatal inflicted head injuries. Child Abuse Negl 1997;21:929-40.
55. Snoek JW, Minderhoud JM, Wilmink JT. Delayed deterioration following mild head injury in children. Brain 1984;107:15-36.
56. Kelly JP, Nichols JS, Filley CM, Lillehei KO, Rubinstein D, Kleinschmidt-DeMasters BK. Concussion in sports: guidelines for the prevention of catastrophic outcome. JAMA 1991;266:2867-9.
57. Cantu RC. Cerebral concussion in sport: management and prevention. Sports Med 1992;14:64-74.
58. Rorke LB. Neuropathology. In: Ludwig S, Kornberg AE, eds. Child abuse: a medical reference. 2nd ed. New York: Churchill Livingstone, 1992:403-21.
59. Hirsch CS. Craniocerebral trauma. In: Froede RC, ed. Handbook of forensic pathology. Northfield, Ill.: College of American Pathologists, 1990:182-90.
60. Lindenberg R, Freytag E. Morphology of brain lesions from blunt trauma in early infancy. Arch Pathol 1969;87:298-305.
61. Vowles GH, Scholtz CL, Cameron JM. Diffuse axonal injury in early infancy. J Clin Pathol 1987;40:185-9.
62. Adams JH, Doyle D, Graham DI, Lawrence AE, McLellan DR. Gliding contusions in nonmissile head injury in humans. Arch Pathol Lab Med 1986;110:485-8. [Erratum, Arch Pathol Lab Med 1986;110:1075.]
63. Adams JH, Graham DI. Diffuse brain damage in non-missile head injury. In: Anthony PP, Macsween RNM, eds. Recent advances in histopathology. Edinburgh, Scotland: Churchill Livingstone, 1984:241-57.
64. Riffenburgh RS, Sathyavagiswaran L. Ocular findings at autopsy of child abuse victims. Ophthalmology 1991;98:1519-24.
65. Budenz DL, Farber MG, Mirchandani HG, Park H, Rorke LB. Ocular and optic nerve hemorrhages in abused infants with intracranial injuries. Ophthalmology 1994;101:559-65.
66. Blumberg PC, Jones NR, North JB. Diffuse axonal injury in head trauma. J Neurol Neurosurg Psychiatry 1989;52:838-41.
67. Adams JH, Doyle D, Ford I, Gennarelli TA, Graham DI, McClellan DR. Diffuse axonal injury in head injury: definition, diagnosis and grading. Histopathology 1989;15:49-59.
68. Pilz P. Axonal injury in head injury. Acta Neurochir Suppl (Wien) 1983;32:119-23.
69. Chiaviello CT, Christoph RA, Bond GR. Stairway-related injuries in children. Pediatrics 1994;94:679-81.
70. Feldman KW, Weinberger E, Milstein JM, Fligner CL. Cervical spine MRI in abused infants. Child Abuse Negl 1997;21:199-205.
71. Bird CR, McMahan JR, Gilles FH, Senac MO, Apthorp JS. Strangulation in child abuse: CT diagnosis. Radiology 1987;163:373-5.
72. Inglis FM, Bullock R, Chen MH, Graham DI, Miller JD, McCulloch J. Ischaemic brain damage associated with tissue hypermetabolism in acute subdural haematoma: reduction by a glutamate antagonist. Acta Neurochir Suppl (Wien) 1990;51:277-9.
73. Bullock R, Butcher SP, Chen M, Kendall L, McCulloch J. Correlation of the extracellular glutamate concentration with extent of blood flow reduction after subdural hematoma in the rat. J Neurosurg 1991;74:794-802.
74. Chen MS, Bullock R, Graham DI, Miller JD, McCulloch J. Ischemic neuronal damage after acute subdural hematoma in the rat: effects of pretreatment with a glutamate antagonist. J Neurosurg 1991;74:944-50.
75. Shaver EG, Duhaime AC, Curtis M, Gennarelli LM, Barrett R. Experimental acute subdural hematoma in infant piglets. Pediatr Neurosurg 1996;25:123-9.
76. Luerssen TG, Bruce DA, Humphreys RP. Position statement on identifying the infant with nonaccidental central nervous system injury (the whiplash-shake syndrome). Pediatr Neurosurg 1993;19:170.
77. Duhaime AC, Christian CW, Armonda R, Hunter J, Hertle R. Disappearing subdural hematomas in children. Pediatr Neurosurg 1996;25:116-22.
78. Duhaime AC, Sutton LN. Head injury in the pediatric patient. In: Tindall GT, Cooper PR, Barrow DL, eds. The practice of neurosurgery. Vol. 2. Baltimore: Williams & Wilkins, 1996:1553-77.
79. Yoffe G, Buchanan GR. Intracranial hemorrhage in newborn and young infants with hemophilia. J Pediatr 1988;113:333-6.
80. Ryan CA, Gayle M. Vitamin K deficiency, intracranial hemorrhage, and a subgaleal hematoma: a fatal combination. Pediatr Emerg Care 1992;8:143-5.
81. Miner ME, Kaufman HH, Graham SH, Haar FH, Gildenberg PL. Disseminated intravascular coagulation fibrinolytic syndrome following head injury in children: frequency and prognostic implications. J Pediatr 1982;100:687-91.
82. Tokoro K, Nakajima F, Yamataki A. Infantile chronic subdural hematoma with local protrusion of the skull in a case of osteogenesis imperfecta. Neurosurgery 1988;22:595-8.
83. Astley R. Metaphyseal fractures in osteogenesis imperfecta. Br J Radiol 1979;52:441-3.
84. Steiner RD, Pepin M, Byers PH. Studies of collagen synthesis and structure in the differentiation of child abuse from osteogenesis imperfecta. J Pediatr 1996;128:542-7.
85. Wenstrup RJ, Willing MC, Starman BJ, Byers PH. Distinct biochemical phenotypes predict clinical severity in nonlethal variants of osteogenesis imperfecta. Am J Hum Genet 1990;46:975-82.
86. Haworth JC, Booth FA, Chudley AE, et al. Phenotypic variability in glutaric aciduria type 1: report of fourteen cases in five Canadian Indian kindreds. J Pediatr 1991;118:52-8.

The Americas:
The New England Journal of Medicine
Publishing Division of the Massachusetts Medical Society
1440 Main Street, Waltham, MA 02451-1600 USA
Tel: (1) 781 893 3800 x 1199, **Fax:** (1) 781 893 0413
E-mail: customer@nejm.massmed.org

All Other Countries:
The New England Journal of Medicine
c/o European Magazine Distribution (EMD) GmbH,
Knesebeckstrasse 96, 10623 Berlin, GERMANY
Tel: (49) 30 3123883, **Fax:** (49) 30 3132032

The New England Journal of Medicine *(ISSN 0028-4793) is published weekly in the English language by the Massachusetts Medical Society (Waltham, MA, USA). Material printed in* The NEJM *is covered by copyright.*

KEITH A. FINDLEY

Keith A. Findley, a 1985 graduate of the Yale Law School and a 1981 graduate of Indiana University, is an assistant professor at the University of Wisconsin Law School, where he is also co-founder and co-director of the Wisconsin Innocence Project, and where he teaches courses in Criminal Procedure, Evidence, and Wrongful Convictions.

Professor Findley just completed a five-year term as president of the Innocence Network, an affiliation of nearly 70 innocence projects in the United States, Canada, the United Kingdom, Ireland, Australia, New Zealand, the Netherlands, France, Italy, South Africa, Israel, and Argentina. He has written and published numerous articles and book chapters on topics related to wrongful conviction of the innocent, including both flawed Shaken Baby Syndrome convictions and more generally the effects of cognitive biases in the investigation and litigation of criminal cases. Through the Wisconsin Innocence Project, he represented Audrey Edmunds in the first case to overturn an SBS conviction based on new developments in the medical science. He has also served as an Assistant State Public Defender in Madison, and he has litigated hundreds of appeals at all levels of state and federal courts.

Profiles in Patient Safety: Confirmation Bias in Emergency Medicine

Jesse M. Pines, MD, MBA

Abstract

Confirmation bias is a pitfall in emergency care and may lead to inaccurate diagnoses and inappropriate treatments and care plans. Because of the increasing severity and volume of emergency care, emergency physicians often must rely on heuristics, such as rule-out protocols, as a guide to diagnosing and treating patients. The use of heuristics or protocols can be potentially misleading if the initial diagnostic impression is incorrect. To minimize cognitive dissonance, clinicians may accentuate confirmatory data and ignore nonconfirmatory data. Clinicians should recognize confirmation bias as a potential pitfall in medical decision making in the emergency department. Reliance on the scientific method, Bayesian reasoning, metacognition, and cognitive forcing strategies may serve to improve diagnostic accuracy and improve patient care.

ACADEMIC EMERGENCY MEDICINE 2006; 13:90–94

Keywords: diagnostic accuracy, confirmation bias, emergency medicine, medical error

Mr. W is a 51-year-old diabetic male who presents to the emergency department (ED) with a seven-day history of lumbar lower back pain that occurred immediately after lifting a heavy box at work. He is triaged at 2:00 AM and is seen by Dr. J at 2:45 AM. He reports radiation of pain down the front of his leg and denies trauma, and bowel or bladder abnormalities. He has been using high-dose Motrin (600 mg every 6 hours) to relieve the pain. He reports a pain severity of 10/10. He has no other medical problems, smokes marijuana occasionally, and has a distant history of IV drug abuse. Triage vitals are as follows: blood pressure, 150/91; heart rate, 105 beats per minute; temperature, 100.5°F; and respiratory rate, 16 respirations per minute. He took 600 mg of Motrin 1 hour before ED arrival. He reports that he has been unable to work all week and needs a written excuse for his boss.

The nurse approaches the emergency physician (EP) and states, "Mr. W is here again. He is here all the time requesting pain medicine and work excuses for lower back pain. He was even here yesterday and was seen by your colleague, Dr. S, [was] diagnosed as having a muscle strain or a herniated disk, [was] given two Percocet orally, and [was] told to follow up with his primary physician. Let's get him out of here." Because it is a busy night, no rooms are available and Mr. W is examined in the hall. He states that he was told to return if he had a fever. Mr. W states that he thought he had a fever at home but did not have a thermometer. On exam, he is very uncomfortable lying recumbent on a stretcher next to his wife, who looks very concerned. Head, neck, heart, lung, and abdominal examination are normal. Back examination reveals diffuse lumbar bony and paraspinous tenderness. He is unable to tolerate a straight-leg raise because of pain. The neurological examination is grossly nonfocal, and he has no major deficits in sensation or motor ability. No rectal or perineal examinations are performed because he is in the hall. He states that the Percocet that he received last night helped "a little" with the pain but did not relieve it completely.

From the Department of Emergency Medicine, University of Pennsylvania (JMP), Philadelphia, PA.

Received August 6, 2004; revision received March 21, 2005; accepted July 26, 2005.

Series editors: Pat Croskerry, MD, PhD, Dartmouth General Hospital Site, Dalhousie University, Halifax, Nova Scotia, Canada; and Marc J. Shapiro, MD, Rhode Island Hospital, Brown University School of Medicine, Providence, RI.

Address for correspondence and reprints: Jesse M. Pines, MD, MBA, Department of Emergency Medicine, University of Pennsylvania, 3400 Spruce Street, Ground Ravin, Philadelphia, PA 19104. Fax: 215-662-3953; e-mail: pinesjes@uphs.upenn.edu.

Scenario 1

While the ED nurse is writing the chart, she again approaches Dr. J: "Come on…Dr. S saw him last night and thought he was fine. I dipped his urine again tonight and it was normal. Let's discharge him. There are 10 people in the waiting room." Dr. J agrees that this is likely drug-seeking behavior and discharges the patient, giving him two Percocet to go, and instructs him again to see his primary physician. Two ED technicians help Mr. W to his car so that his wife can drive him home.

ISSN 1069-6563
PII ISSN 1069-6563583

doi: 10.1197/j.aem.2005.07.028

Scenario 2

Dr. J insists that Mr. W's pain be controlled in the ED and that he be examined in a private room. His urine is dipped and is negative. Mr. W changes into a gown, and he is found to have severe pain with standing and then becomes diaphoretic. Rectal tone and perineal sensation and skin examination are normal. Four milligrams of IM morphine sulfate are ordered. On reexamination, the EP notices that Mr. W is diaphoretic and that his temperature now has risen to 102.2°F. He states that his pain now rates 9/10. On further questioning, he admits using IV heroin 3 weeks before onset of the pain. Complete blood count, chemistries, erythrocyte sedimentation rate (ESR), a chest radiograph, and lumbar plain films are ordered. His white blood cell count (WBC) is 11.8 × 10^3/mm^3, his ESR is 47 mm/h, and the rest of his labs are unremarkable. Chest radiograph shows no acute disease. Lumbar films show vertebral endplate and disk destruction. Emergency magnetic resonance imaging with gadolinium enhancement is ordered and reveals findings consistent with epidural abscess. The neurosurgery consultant is immediately notified of the results and decides to take Mr. W to the operating room for emergent drainage.

DISCUSSION

It is easy to imagine how the first scenario might happen in a busy ED. EPs repeatedly are challenged to rapidly diagnose and treat multiple patients, some of whom present with potentially life-threatening illness. EPs increasingly are being forced to work in crowded conditions and to focus on efficiency of patient throughput while attempting to maintain the highest possible quality of care. Because of the depth, scope, and volume of cognitive thinking required to manage patient information, medical errors of cognition are a significant issue in emergency medicine (EM) practice.[1–5] EM particularly is susceptible to cognitive errors, because clinicians are required to integrate their knowledge base with new situations to create a diagnostic and management plan.[6] EPs face a very high cognitive load and frequently manage many patients simultaneously who have life-threatening and potentially life-threatening conditions. Many studies have confirmed that the major cause of malpractice claims in EDs is a failure to diagnose.[7–9] A 1993 study found that about 2% of patients with acute myocardial infarction mistakenly are sent home.[10]

Because of the rapidity with which EPs must work and the importance of an accurate diagnosis, it is important that EPs be cognizant of the possibility that diagnoses may be compromised by *confirmation bias*. Put simply, this means that one may have an initial or a preconceived idea about something and interpret subsequent information or data so as to confirm that idea (or in the case of EPs, to confirm the diagnoses). As a specialty, EPs have developed skills that open them to potential errors in cognition such as confirmation bias.

In the case presentation, an initial biased approach may be for Dr. J to confirm Dr. S' diagnosis of musculoskeletal back pain without further in-depth examination and investigation. Certain elements in the history confirm his judgment. The natural inclination of a busy EP is to sort patients quickly by categorizing them by diagnostic or treatment strategy.[11] In this case, the EP may accentuate the historical elements confirming the diagnosis of musculoskeletal back pain (preceded by injury, previous diagnosis, and history of many ED visits) and not investigate further pertinent historical elements (e.g., when pressed, Mr. W admitted to recent intravenous drug abuse).

Confirmation bias is related closely to *anchoring bias*, which can come into play when there is an incorrect initial impression and the focus of the evaluation is centered on that initial impression. For example, in this case, our patient who presents with classic musculoskeletal back pain (and multiple visits) actually has an epidural abscess, or a patient treated frequently for migraine headaches actually has an acute subarachnoid hemorrhage.

Because of the volume and acuity of care in an ED, quick sorting and categorization can serve to reduce the already high cognitive load required to manage multiple patients.[12] One study of trauma patients found that there were reasoning errors in 100% of trauma resuscitations.[13]

The use of heuristics is a necessary evil in caring for ED patients. The use of heuristics is inevitable to allow clinicians to maintain efficiency and not chase the metaphorical zebras (a colloquial term designating those possible diagnoses that are least likely and most difficult to confirm on the basis of given clinical data, as in the saying, "when you hear hoofbeats, think horses, not zebras"). Thus, attaching safeguards to the heuristics, rather than avoiding the heuristics, has been a solution for error avoidance. However, sometimes the initial clinical suspicion is not borne out by results of diagnostic tests, repeated examinations, and observation. When the initial clinical suspicion is high for a particular illness, the EP may place more emphasis on confirmatory data than on nonconfirmatory data. For example, in this case, because the nurse and Dr. S both see the likely diagnosis as musculoskeletal pain, Dr. J may preferentially search for information that confirms that diagnosis and not approach Mr. W as if he were a new case of severe lower back pain. Thus, the influence of confirmation bias can lead to errors in medical decision making. This can be even more powerful when, in a clinician's judgment, a constellation of signs and symptoms appears pathognomonic of a particular illness, or when another physician has already made a diagnosis. The tendency to overemphasize confirmatory data (confirmation bias) often can compromise the ability of EPs to accurately diagnose and treat patients. This can lead to EPs overlooking vital information and not asking all the right questions needed to diagnose and treat patients accurately.

Confirmation bias occurs when people selectively focus upon evidence that supports their beliefs or what they want or believe to be true, while ignoring evidence that serves to disconfirm those ideas. Confirmation bias is a very human way of thinking. Francis Bacon described confirmation bias as follows in 1620:

> The human understanding when it has once adopted an opinion (either as being the received opinion or as being agreeable to itself) draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these it either neglects and despises, or else by some distinction sets aside

> and rejects; in order that this great and pernicious predetermination the authority of its former conclusions may remain inviolate.[14]

Confirmation bias is well documented in the behavioral and economic literatures.[15–17] It empirically is even stronger when information is presented sequentially, as it is in clinical emergency care, compared with when all the information is available up front.[18] Variability in the temporal processing and receipt of information may influence decision making because the longer a person holds onto a decision or approach, the more difficult it becomes for him or her to break from that thinking.

When multiple providers are caring for a patient, confirmation bias can have a variable effect on guiding accurate diagnosis and treatment. Other providers caring for the same patient may verbally confirm a diagnosis or reinforce the initial categorization of a patient by the initial diagnostic impression. In the case of an EM resident presenting a patient in the assessment-oriented way, in which the assessment precedes the presentation, confirmatory data may be highlighted to reinforce the assessment, whereas nonconfirmatory data (that may or may not have been asked for) may be omitted.[19] However, the presence of multiple providers may help prevent confirmation bias because one provider may get a critical bit of information that differs from the first providers', and he or she accordingly changes the plan of care. When a medical student spends an hour taking an exhaustive history and physical examination, that student's lack of direction (and lack of knowledge of heuristics) ultimately can lead to pertinent information being found that may not have been found in a briefer encounter.

Further examples of confirmation bias affecting a conclusion on a less individual basis include a drug vendor's interpretation of a study designed to validate the use of its product (and publication of the same). In addition, when clinical policies or pathways designed by clinicians occur in a hospital with a particular research interest in a mode of therapy, or sponsorship by a particular vendor, confirmatory data may be accentuated in a nonscientific manner.

An additional level of complexity in the EM decision-making process occurs when a clinical impression is strong enough to guide diagnosis without the support of confirmatory data. An example of this is a patient with typical features of chest pain resembling an acute coronary syndrome. Adjunctive data may not support this, such as a negative cardiac marker or electrocardiogram (that may be normal in the early stages of acute myocardial infarction), but often definitive diagnostic workup may not be immediately available to emergency healthcare providers (cardiac catheterization), and a judgment must be made on clinical grounds. Confirmatory tests in this situation must be taken with an in-depth understanding of both the value (sensitivity and specificity) and limitations of available historical information, ED testing, and appropriate use of all the available data in clinical decision making. Here confirmation bias can be helpful because even in the face of negative data (ECG, cardiac markers), a high level of clinical suspicion is not changed by the objective data.

A related concept is *cognitive dissonance,* which holds that it is psychologically uncomfortable to hold contradictory cognitions.[20] It can be confusing when an unexpected result (usually negative) comes back on a patient in whom the illness in question was highly suspected. This can lead to disposition issues. When the initial impression is highly suspicious for serious illness and the initial search for a cause is not fruitful, EPs sometimes may accentuate any positive data to support a justification for hospital admission. This again is a beneficial effect of confirmation bias when it leads to appropriate patient care.

Physicians and scientists are prone to confirmation bias, as are practitioners in many other academic disciplines. The more that researchers believe that they are right, the greater weight they place on confirmatory information. One study in which journal reviewers were asked to evaluate manuscripts that described identical experimental procedures reporting variable results (positive, negative, or mixed) found that reviewers were strongly biased against manuscripts that reported results contrary to their theoretical perspective.[21,22]

One solution to managing data and decision making in high workload situations is the presence of automation. Automation in clinical medicine is analogous to clinical guidelines, such as protocols that may be present in a chest pain center (rule-out protocols).[23] These may even be built into clinical information systems or ED protocols. A recent study in the aviation literature showed that automation was helpful in guiding initial plans but found that one third of pilots failed to revise flight plans as a result of change in conditions.[24] One could argue that the presence of automation may even hinder the ability to reconsider alternative diagnoses when there exists a location bias (chest pain center). This may lead to a patient who has chest pain secondary to cholelithiasis being misdiagnosed after a cardiac evaluation. Automation in clinical information systems may be very useful in reducing medication errors,[25] but built-in forcing strategies such as "trauma labs," "toxicology labs," or "rule out cholecystitis with labs and ultrasound," may result in missed diagnosis if the initial impression (anchoring bias) is incorrect.

Additionally, the use of so-called screening labs should be used principally in the way they were designed: as screening tests (i.e., not as diagnostic tests). For example, when a patient who is a poor historian presents with nonspecific symptoms, the use of screening labs most often is not helpful to the patient, and further historical evaluation must be done to identify a source for the complaint (i.e., calling family and other providers). Aside from consuming health care dollars by performing tests that are very unlikely to help the patient, when abnormal test results return that were not appropriately ordered in the first place, it might lead physicians to jump to inappropriate diagnostic conclusions.

Schermer[26] stated, "Smart people believe weird things because they are skilled at defending beliefs they arrived at for nonsmart [sic] reasons." Because of the increasing complexity of cases and the cognitive load required to take care of them, EPs must rely on heuristics to care for many patients. One solution to confirmation bias is the application of the scientific method, in which the intent is to disprove a belief as opposed to searching for only confirmatory evidence. Although this makes sense

in theory, it takes an experienced clinician to be unaffected by the confirmation bias of the initial diagnostic impression. The scientific method also can have its own pitfalls. In the case of chest pain of potential cardiac origin, one cannot start with the impression that it is present and then look for evidence to disprove it, because obtainable data beyond the clinical impression are not strong enough to overrule the initial impression. Evidence that disproves can be just as suspect as evidence that confirms; it depends on the likelihood ratios (or sensitivity and specificity) of the evidence used.

A particularly robust approach that is particularly useful in EM involves Bayesian reasoning, in which known data on tests are combined with initial clinical impressions (pretest probabilities) to derive accurate diagnostic probabilities of disease (posttest probability).[27] For example, when the clinician has a high clinical suspicion (pretest), nonconfirmatory data are more likely to be false (false negative), and confirmatory data are more likely to be true, than if the pretest clinical suspicion was low. It becomes appropriate to emphasize nonconfirmatory data to a lesser degree. Changing the diagnostic impression as a result of a negative test is more likely to cause a diagnostic error. Examples of this include using a normal WBC to exclude the diagnosis of appendicitis in a classic clinical presentation, or excluding pulmonary embolism in a patient with a moderate clinical probability but a normal D-dimer. Key skills in Bayesian reasoning are deciding whether and how the test will contribute to diagnostic certainty before ordering it and interpreting the result in light of the pretest probability. If the test is ordered specifically to confirm a positive diagnosis, then the clinician should disregard a negative result. For example, in the hypotensive trauma patient, free abdominal fluid on the FAST exam should lead directly to laparotomy (if there is no other reason for hypotension); however, if no free fluid is seen, the test is not sensitive enough to rule out intraabdominal injury, and the patient should be further evaluated (laparotomy if the patient remains unstable; abdominal CT if they are stable). If it is done to refute a diagnosis (i.e., as rule-out), a positive diagnosis similarly only calls for further tests. An example of this is the use of the ESR for temporal arteritis in an otherwise low-risk patient. If the ESR is elevated, a biopsy still is needed before a definite diagnosis can be made; if the ESR is low, clinicians can be fairly certain that temporal arteritis is not present. In these ways, the use of heuristics and confirmation bias actually may prevent misdiagnosis in that they may prevent the physician from leaving the most probable diagnosis to chase a zebra.

Confirmation bias is an issue for clinicians taking the initial history when the first impression steers the history in such a way that the physician poses questions that confirm the impression and may not ask the ones that might suggest a different diagnosis. The physician is not necessarily ignoring relevant data; however, the chain of thought that he or she follows simply is not allowing him or her to steer in the direction of seeking truth.

Sometimes, because of distractions, such as thoughts of other patients who are being treated concurrently, the EP may gloss over certain things in the history while seeking the list of typical presenting features of the suspected disease. This last situation was eloquently expressed by Simon and Garfunkel, who sang, "A man hears what he wants to hear and disregards the rest." The practice of EM requires the processing of multiple complex data elements in a real-time, high-stakes environment. Patient re-evaluation ideally should occur at every step of the process as new data become available, and EPs should update posttest probabilities on the basis of new information.

A solution suggested by Croskerry and Sinclair[6] is the use of metacognition by experienced providers and education of medical students and residents on the use of this cognitive strategy. Metacognition involves stepping back and thinking about the cognitive process that goes into making a decision. Specifically, medical educators should focus on teaching the cognitive process to students and residents and should realize the limitations of medical data by using cognitive aids (such as computers and personal data assistants). They need to consciously step back and see the broader range of possibilities, reexamine decision making as new data become available, avoid overconfidence, and effectively select strategies to deal with problems in decision making.

Another potential solution is the use of cognitive forcing strategies. These can be categorized into universal, generic, and specific strategies. A universal cognitive forcing strategy is defined as a generalized understanding of the error theory and the appreciation and application of metacognition. A generic cognitive forcing strategy involves understanding the general heuristics in medical decision making and under what circumstances they fail. Understanding and recognizing confirmation bias is a subset of this process, and clinicians must be aware of such bias to adjust initial impressions on the basis of new objective data. *Search-satisficing,* or calling off a search once a positive result is found, is an example of a generic strategy that could be applied to stopping a search for coingestants in a toxic poisoning once a primary ingestant is found. A specific cognitive forcing strategy relates to known pitfalls in specific diagnostic workups.[6] There are many pitfalls in clinical EM; for example, failure to consider a closed-head injury in an inebriated patient. In this sense, the usage of a cognitive forcing strategy is the deliberate usage of a particular strategy in a specific situation that optimizes medical decision making and minimizes error.[6]

Still, urgent clinical decisions must be made in EDs without complete information. These strategies need to be balanced against the dangers of indecision in cases (and busy departments) in which delay can have adverse consequences to the patient or to those waiting to be treated. The nature of EM is, whether providers like it or not, tied to situations in which the ED is crowded and the demand for services is pushed to capacity. Simply acknowledging this brings providers no further toward a safer system for patients or providers.

Even the information that is available is imperfect: historical information and physical examination results both have high interrater variability, and diagnostic tests ordered in the ED each have an intrinsic error rate (sensitivity and specificity) that must be considered. Recognition and understanding that confirmation bias may exist may help clinicians to rethink the objective data when using a specific data point to guide medical decision making.

CONCLUSIONS

Stepping back and reconsidering the objective facts may guide clinicians to reconsider the initial impression and pursue a completely different diagnostic strategy. Using and teaching metacognition and an understanding of error theory and cognitive forcing strategies may be helpful in minimizing confirmation bias. When the initial clinical impression is not corroborated by objective data, EPs must be open to revisiting the possibility of an inaccurate diagnosis and may have to start again at diagnostic time zero or, alternatively, defer to an appropriate inpatient or outpatient workup.

References

1. Brennan TA. The Institute of Medicine report on medical error—could it do harm? N Engl J Med. 2000; 342:1123–5.
2. Famularo G, Salvini P, Terranova A, Gerace C. Clinical errors in emergency medicine: experience at the emergency department of an Italian teaching hospital. Acad Emerg Med. 2000; 7:1278–81.
3. Glick TH, Workman TP, Gaifberg SV. Suspected conversion disorder: foreseeable risks and avoidable errors. Acad Emerg Med. 2000; 7:1272–7.
4. Kuhn GJ. Diagnostic errors. Acad Emerg Med. 2002; 9:740–50.
5. Croskerry P, Sinclair D. Emergency medicine: a practice prone to error? CJEM. 2001; 3:271–6.
6. Croskerry P. Cognitive forcing strategies in clinical decision-making. Ann Emerg Med. 2003; 41:110–20.
7. U.S. General Accounting Office, the Ohio Hospital Association and the St. Paul (MN) Insurance Company. 1998 Data. Available at: http://hookman.com/mp9807.htm. Accessed Aug 6, 2004.
8. McQuade JS. The medical malpractice crisis—reflections on the alleged causes and proposed cures: discussion paper. J R Soc Med. 1991; 84:408–11.
9. Kronz J, Westra W. The role of second opinion pathology in the management of lesions of the head and neck. Curr Opin Otolaryngol Head Neck Surg. 2005; 13:81–4.
10. McCarthy BD, Beshansky JR, D'Agostino RB, Selker HP. Missed diagnoses of acute myocardial infarction in the emergency department; results from a multicenter study. Ann Emerg Med. 1993; 22:579–82.
11. Kovacs G, Croskerry P. Clinical decision making: an emergency medicine perspective. Acad Emerg Med. 1999; 6:947–52.
12. Croskerry P. The cognitive imperative: thinking about how we think. Acad Emerg Med. 2000; 7:1223–31.
13. Clarke JR, Spejewski B, Gertner AS, et al. An objective analysis of process errors in trauma resuscitations. Acad Emerg Med. 2000; 7:1303–10.
14. Bacon F. Novum Organum. New Organon, 1620.
15. Gilovich T. How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life. New York, NY: Free Press, 1993.
16. Mynatt CR, Doherty ME, Tweney RD. Consequences of confirmation and disconfirmation in a simulated research environment. Q J Exp Psychol. 1978; 30: 395–406.
17. Dave C, Wolfe KW. On Confirmation Bias and Deviations from Bayesian Updating. Available at: http://www.peel.pitt.edu/esa2003/papers/wolfe_confirmationbias.pdf. Accessed Apr 21, 2004.
18. Jonas E, Schulz-Hardt S, Frey D, Thelen N. Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information. J Pers Soc Psychol. 2001; 80:557–71.
19. Maddow CL, Shah MN, Olsen J. Efficient communication: assessment-oriented case presentation. Acad Emerg Med. 2003; 10:842–7.
20. Festinger L. A Theory of Cognitive Dissonance. Palo Alto, CA: Stanford University Press, 1957.
21. Mahoney M. Publication prejudices: an experimental study of confirmatory bias in the peer review system. Cog Ther Res. 1977; 1:161–75.
22. Mahoney MJ, DeMonbreun BG. Confirmatory bias in scientists and non-scientists. Cog Ther Res. 1977; 1:176–80.
23. Fesmire FM, Hughes AD, Fody EP, et al. The Erlanger chest pain evaluation protocol: a one-year experience with serial 12-lead ECG monitoring, two-hour delta serum marker measurements, and selective nuclear stress testing to identify and exclude acute coronary syndromes. Ann Emerg Med. 2002; 40:584–94.
24. Muthard EK, Wickens CD. Factors that Mediate Flight Plan Monitoring and Errors in Plan Revision: Planning under Automated and High Workload Conditions. Presented at the 12th International Symposium on Aviation Psychology. 2003. Available at: http://www.aviation.uiuc.edu/UnitsHFD/conference/Dayton03/mutwic.pdf. Accessed April 21, 2004.
25. Bates DW. Using information technology to reduce rates of medication errors in hospitals. Br Med J. 2000; 7237:788–91.
26. Schermer M. Smart people believe weird things. Sci Am. 2002.
27. El-Gamal MA, Grether DM. Are people bayesian? Uncovering behavioral strategies. J Am Stat Assoc. 1995; 90:1137–45.

The American Journal of Medicine (2008) Vol 121 (5A), S2–S23

ELSEVIER

THE AMERICAN JOURNAL *of* MEDICINE®

Overconfidence as a Cause of Diagnostic Error in Medicine

Eta S. Berner, EdD,[a] and Mark L. Graber, MD[b]

[a]Department of Health Services Administration, School of Health Professions, University of Alabama at Birmingham, Birmingham, Alabama, USA; and [b]VA Medical Center, Northport, New York and Department of Medicine, State University of New York at Stony Brook, Stony Brook, New York, USA

ABSTRACT

The great majority of medical diagnoses are made using automatic, efficient cognitive processes, and these diagnoses are correct most of the time. This analytic review concerns the exceptions: the times when these cognitive processes fail and the final diagnosis is missed or wrong. We argue that physicians in general underappreciate the likelihood that their diagnoses are wrong and that this tendency to overconfidence is related to both intrinsic and systemically reinforced factors. We present a comprehensive review of the available literature and current thinking related to these issues. The review covers the incidence and impact of diagnostic error, data on physician overconfidence as a contributing cause of errors, strategies to improve the accuracy of diagnostic decision making, and recommendations for future research.

KEYWORDS: Cognition; Decision making; Diagnosis; Diagnosis, computer-assisted; Diagnostic errors; Feedback

Not only are they wrong but physicians are "walking . . . in a fog of misplaced optimism" with regard to their confidence.

—Fran Lowry[1]

Mongerson[2] describes in poignant detail the impact of a diagnostic error on the individual patient. Large-scale surveys of patients have shown that patients and their physicians perceive that medical errors in general, and diagnostic errors in particular, are common and of concern. For instance, Blendon and colleagues[3] surveyed patients and physicians on the extent to which they or a member of their family had experienced medical errors, defined as mistakes that "result in serious harm, such as death, disability, or additional or prolonged treatment." They found that 35% of physicians and 42% of patients reported such errors.

A more recent survey of 2,201 adults in the United States commissioned by a company that markets a diagnostic decision-support tool found similar results.[4] In that survey, 35% experienced a medical mistake in the past 5 years involving themselves, their family, or friends; half of the mistakes were described as diagnostic errors. Of these, 35% resulted in permanent harm or death. Interestingly, 55% of respondents listed misdiagnosis as the greatest concern when seeing a physician in the outpatient setting, while 23% listed it as the error of most concern in the hospital setting. Concerns about medical errors also were reported by 38% of patients who had recently visited an emergency department; of these, the most common worry was misdiagnosis (22%).[5]

These surveys show that patients report frequent experience with diagnostic errors and/or that these errors are of significant concern for them in their encounters with the healthcare system. However, as pointed out in an editorial by Tierney,[6] patients may not always interpret adverse events accurately, or may differ with their physicians as to the reason for the adverse event. For this reason, we have reviewed the scientific literature on the incidence and impact of diagnostic error and have examined the literature on overconfidence as a contributing cause of diagnostic errors. In the latter portion of this article we review the literature on the effectiveness of potential strategies to reduce diagnostic error and recommend future directions for research.

This research was supported through the Paul Mongerson Foundation within the Raymond James Charitable Endowment Fund (ESB) and the National Patient Safety Foundation (MLG).

Statement of author disclosures: Please see the Author Disclosures section at the end of this article.

Requests for reprints should be addressed to Eta S. Berner, EdD, Department of Health Services Administration, School of Health Professions, University of Alabama at Birmingham, 1675 University Boulevard, Room 544, Birmingham, Alabama 35294-3361.

E-mail address: eberner@uab.edu.

0002-9343/$ -see front matter
doi:10.1016/j.amjmed.2008.01.001

INCIDENCE AND IMPACT OF DIAGNOSTIC ERROR

We reviewed the scientific literature with several questions in mind: (1) What is the extent of incorrect diagnosis? (2) What percentage of documented adverse events can be attributed to diagnostic errors and, conversely, how often do diagnostic errors lead to adverse events? (3) Has the rate of diagnostic errors decreased over time?

What is the Extent of Incorrect Diagnosis?

Diagnostic errors are encountered in every specialty, and are generally lowest for the 2 perceptual specialties, radiology and pathology, which rely heavily on visual interpretation. An extensive knowledge base and expertise in visual pattern recognition serve as the cornerstones of diagnosis for radiologists and pathologists.[7] The error rates in clinical radiology and anatomic pathology probably range from 2% to 5%,[8–10] although much higher rates have been reported in certain circumstances.[9,11] The typically low error rates in these specialties should not be expected in those practices and institutions that allow x-rays to be read by frontline clinicians who are not trained radiologists. For example, in a study of x-rays interpreted by emergency department physicians because a staff radiologist was unavailable, up to 16% of plain films and 35% of cranial computed tomography (CT) studies were misread.[12]

Error rates in the clinical specialties are higher than in perceptual specialties, consistent with the added demands of data gathering and synthesis. A study of admissions to British hospitals reported that 6% of the admitting diagnoses were incorrect.[13] The emergency department requires complex decision making in settings of above-average uncertainty and stress. The rate of diagnostic error in this arena ranges from 0.6% to 12%.[14,15]

Based on his lifelong experience studying diagnostic decision making, Elstein[16] estimated that the rate of diagnostic error in clinical medicine was approximately 15%. In this section, we review data from a wide variety of sources that suggest this estimate is reasonably correct.

Second Opinions and Reviews. Several studies have examined changes in diagnosis after a second opinion. Kedar and associates,[17] using telemedicine consultations with specialists in a variety of fields, found a 5% change in diagnosis. There is a wealth of information in the perceptual specialties using second opinions to judge the rate of diagnostic error. These studies report a variable rate of discordance, some of which represents true error, and some is disagreement in interpretation or nonstandard defining criteria. It is important to emphasize that only a fraction of the discordance in these studies was found to cause harm.

Dermatology. Most studies focused on the diagnosis of pigmented lesions (e.g., ruling out melanoma). For example, in a study of 5,136 biopsies, a major change in diagnosis was encountered in 11% on second review. Roughly 1% of diagnoses were changed from benign to malignant, roughly 1% were downgraded from malignant to benign, and in roughly 8% the tumor grade was changed enough to alter treatment.[18]

Anatomic Pathology. There have been several attempts to determine the true extent of diagnostic error in anatomic pathology, although the standards used to define an error in this field are still evolving.[19] In 2000, The American Society of Clinical Pathologists convened a consensus conference to review second opinions in anatomic pathology.[20] In 1 such study, the pathology department at the Johns Hopkins Hospital required a second opinion on each of the 6,171 specimens obtained over an 18-month period; discordance resulting in a major change of treatment or prognosis was found in just 1.4 % of these cases.[10] A similar study at Hershey Medical Center in Pennsylvania identified a 5.8% incidence of clinically significant changes.[20] Disease-specific incidences ranged from 1.3% in prostate samples to 5% in tissues from the female reproductive tract and 10% in cancer patients. Certain tissues are notoriously difficult; for example, discordance rates range from 20% to 25% for lymphomas and sarcomas.[21,22]

Radiology. Second readings in radiology typically disclose discordance rates in the range of 2% to 20% for most general radiology imaging formats, although higher rates have been found in some studies.[23,24] The discordance rate in practice seems to be $<5\%$ in most cases.[25,26]

Mammography has attracted the most attention in regard to diagnostic error in radiology. There is substantial variability from one radiologist to another in the ability to accurately detect breast cancer, and it is estimated that 10% to 30% of breast cancers are missed on mammography.[27,28] A recent study of breast cancer found that the diagnosis was inappropriately delayed in 9%, and a third of these reflected misreading of the mammogram.[29] In addition to missing cancer known to be present, mammographers can be overly aggressive in reading studies, frequently recommending biopsies for what turn out to be benign lesions. Given the differences regarding insurance coverage and the medical malpractice systems between the United States and the United Kingdom, it is not surprising that women in the United States are twice as likely as women in the United Kingdom to have a negative biopsy.[30]

Studies of Specific Conditions. Table 1 is a sampling of studies[18,27,31–46] that have measured the rate of diagnostic error in specific conditions. An unsettling consistency emerges: the frequency of diagnostic error is disappointingly high. This is true for both relatively benign conditions and disorders where rapid and accurate diagnosis is essential, such as myocardial infarction, pulmonary embolism, and dissecting or ruptured aortic aneurysms.

Table 1 Sampling of Diagnostic Error Rates in Specific Conditions

Study	Conditions	Findings
Shojania et al (2002)[32]	Pulmonary TB	Review of autopsy studies that have specifically focused on the diagnosis of pulmonary TB; ~50% of these diagnoses were not suspected antemortem
Pidenda et al (2001)[33]	Pulmonary embolism	Review of fatal embolism over a 5-yr period at a single institution. Of 67 patients who died of pulmonary embolism, the diagnosis was not suspected clinically in 37 (55%)
Lederle et al (1994),[34] von Kodolitsch et al (2000)[35]	Ruptured aortic aneurysm	Review of all cases at a single medical center over a 7-yr period. Of 23 cases involving abdominal aneurysms, diagnosis of ruptured aneurysm was initially missed in 14 (61%); in patients presenting with chest pain, diagnosis of dissecting aneurysm of the proximal aorta was missed in 35% of cases
Edlow (2005)[36]	Subarachnoid hemorrhage	Updated review of published studies on subarachnoid hemorrhage: ~30% are misdiagnosed on initial evaluation
Burton et al (1998)[37]	Cancer detection	Autopsy study at a single hospital: of the 250 malignant neoplasms found at autopsy, 111 were either misdiagnosed or undiagnosed, and in 57 of the cases the cause of death was judged to be related to the cancer
Beam et al (1996)[27]	Breast cancer	50 accredited centers agreed to review mammograms of 79 women, 45 of whom had breast cancer; the cancer would have been missed in 21%
McGinnis et al (2002)[18]	Melanoma	Second review of 5,136 biopsy samples; diagnosis changed in 11% (1.1% from benign to malignant, 1.2% from malignant to benign, and 8% had a change in tumor grade)
Perlis (2005)[38]	Bipolar disorder	The initial diagnosis was wrong in 69% of patients with bipolar disorder and delays in establishing the correct diagnosis were common
Graff et al (2000)[39]	Appendicitis	Retrospective study at 12 hospitals of patients with abdominal pain and operations for appendicitis. Of 1,026 patients who had surgery, there was no appendicitis in 110 (10.5%); of 916 patients with a final diagnosis of appendicitis, the diagnosis was missed or wrong in 170 (18.6%)
Raab et al (2005)[40]	Cancer pathology	The frequency of errors in diagnosing cancer was measured at 4 hospitals over a 1-yr period. The error rate of pathologic diagnosis was 2%–9% for gynecology cases and 5%–12% for nongynecology cases; errors represented sampling deficiencies, preparation problems, and mistakes in histologic interpretation
Buchweitz et al (2005)[41]	Endometriosis	Digital videotapes of laparoscopies were shown to 108 gynecologic surgeons; the interobserver agreement regarding the number of lesions was low (18%)
Gorter et al (2002)[42]	Psoriatic arthritis	1 of 2 SPs with psoriatic arthritis visited 23 rheumatologists; the diagnosis was missed or wrong in 9 visits (39%)
Bogun et al (2004)[43]	Atrial fibrillation	Review of automated ECG interpretations read as showing atrial fibrillation; 35% of the patients were misdiagnosed by the machine, and the error was detected by the reviewing clinician only 76% of the time
Arnon et al (2006)[44]	Infant botulism	Study of 129 infants in California suspected of having botulism during a 5-yr period; only 50% of the cases were suspected at the time of admission
Edelman (2002)[45]	Diabetes mellitus	Retrospective review of 1,426 patients with laboratory evidence of diabetes mellitus (glucose >200 mg/dL* or hemoglobin A_{1c} >7%); there was no mention of diabetes in the medical record of 18% of patients
Russell et al (1988)[46]	Chest x-rays in the ED	One third of x-rays were incorrectly interpreted by the ED staff compared with the final readings by radiologists

ECG = electrocardiograph; ED = emergency department; SP = standardized patient; TB = tuberculosis.

*1 mg/dL = 0.05551 mmol/L.

Adapted from *Advances in Patient Safety: From Research to Implementation*.[31]

Autopsy Studies. The autopsy has been described as "the most powerful tool in the history of medicine"[47] and the "gold standard" for detecting diagnostic errors. Richard Cabot correlated case records with autopsy findings in several thousand patients at Massachusetts General Hospital, concluding in 1912 that the clinical diagnosis was wrong 40% of the time.[48,49] Similar discrepancies between clinical and autopsy diagnoses were found in a more recent study of geriatric patients in the Netherlands.[50] On average, 10% of autopsies revealed that the clinical diagnosis was wrong, and 25% revealed a new problem that had not been suspected clinically. Although a fraction of these discrepancies reflected incidental findings of no clinical significance, major unexpected discrepancies that potentially could have changed the outcome were found in approximately 10% of all autopsies.[32,51]

Shojania and colleagues[32] point out that autopsy studies only provide the error rate in patients who die. Because the diagnostic error rate is almost certainly lower among patients with the condition who are still alive, error rates measured solely from autopsy data may be distorted. That is, clinicians are attempting to make the diagnosis among living patients before death, so the more relevant statistic in this setting is the sensitivity of clinical diagnosis. For example, whereas autopsy studies suggest that fatal pulmonary embolism is misdiagnosed approximately 55% of the time (see Table 1), the misdiagnosis rate for all cases of pulmonary embolism is only 4%. Shojania and associates[32] argue that a large discrepancy also exists regarding the misdiagnosis rate for myocardial infarction: although autopsy data suggest roughly 20% of these events are missed, data from the clinical setting (patients presenting with chest pain or other relevant symptoms) indicate that only 2% to 4% are missed.

Studies Using Standardized Cases. One method of testing diagnostic accuracy is to control for variations in case presentation by using standardized cases that can enable comparisons of performance across physicians. One such approach is to incorporate what are termed *standardized patients* (SPs). Usually, SPs are lay individuals trained to portray a specific case or are individuals with certain clinical conditions trained to be study subjects.[52,53] Diagnostic errors are inevitably detected when physicians are tested with SPs or standardized case scenarios.[42,54] For example, when asked to evaluate SPs with common conditions in a clinic setting, internists missed the correct diagnosis 13% of the time.[55] Other studies using different types of standardized cases have found that not only is there variation between providers who analyze the same case[27,56] but that physicians can even disagree with themselves when presented again with a case they have previously diagnosed.[57]

What Percentage of Adverse Events is Attributable to Diagnostic Errors and What Percentage of Diagnostic Errors Leads to Adverse Events?

Data from large-scale, retrospective, chart-review studies of adverse events have shown a high percentage of diagnostic errors. In the Harvard Medical Practice Study of 30,195 hospital records, diagnostic errors accounted for 17% of adverse events.[58,59] A more recent follow-up study of 15,000 records from Colorado and Utah reported that diagnostic errors contributed to 6.9% of the adverse events.[60] Using the same methodology, the Canadian Adverse Events Study found that 10.5% of adverse events were related to diagnostic procedures.[61] The Quality in Australian Health Care Study identified 2,351 adverse events related to hospitalization, of which 20% represented delays in diagnosis or treatment and 15.8% reflected failure to "synthesize/decide/act on" information.[62] A large study in New Zealand examined 6,579 inpatient medical records from admissions in 1998 and found that diagnostic errors accounted for 8% of adverse events; 11.4% of those were judged to be preventable.[63]

Error Databases. Although of limited use in quantifying the absolute incidence of diagnostic errors, voluntary error-reporting systems provide insight into the relative incidence of diagnostic errors compared with medication errors, treatment errors, and other major categories. Out of 805 voluntary reports of medical errors from 324 Australian physicians, there were 275 diagnostic errors (34%) submitted over a 20-month period.[64] Compared with medication and treatment errors, diagnostic errors were judged to have caused the most harm, but were the least preventable. A smaller study reported a 14% relative incidence of diagnostic errors from Australian physicians and 12% from physicians of other countries.[65] Mandatory error-reporting systems that rely on self-reporting typically yield fewer error reports than are found using other methodologies. For example, only 9 diagnostic errors were reported out of almost 1 million ambulatory visits over a 5.5-year period in a large healthcare system.[66]

Diagnostic errors are the most common adverse event reported by medical trainees.[67,68] Notably, of the 29 diagnostic errors reported voluntarily by trainees in 1 study, none of these were detected by the hospital's traditional incident-reporting mechanisms.[68]

Malpractice Claims. Diagnostic errors are typically the leading or the second-leading cause of malpractice claims in the United States and abroad.[69–72] Surprisingly, the vast majority of claims filed reflect a very small subset of diagnoses. For example, 93% of claims in the Australian registry reflect just 6 scenarios (failure to diagnose cancer, injuries after trauma, surgical problems, infections, heart attacks, and venous thromboembolic disease).[73] In a recent study of malpractice claims,[74] diagnostic errors were equally preva-

lent in successful and unsuccessful claims and represented 30% of all claims.

The percentage of diagnostic errors that leads to adverse events is the most difficult to determine, in that the prospective tracking needed for these studies is rarely done. As Schiff,[75] Redelmeier,[76] and Gandhi and colleagues[77] advocate, much better methods for tracking and follow-up of patients are needed. For some authors, diagnostic errors that do not result in serious harm are not even considered misdiagnoses.[78] This is little consolation, however, for the patients who suffer the consequences of these mistakes. The increasing adoption of electronic medical records, especially in ambulatory practices, will lead to better data for answering this question; research should be conducted to address this deficiency.

Has the Diagnostic Error Rate Changed Over Time?

Autopsy data provide us the opportunity to see whether the rate of diagnostic errors has decreased over time, reflecting the many advances in medical imaging and diagnostic testing. Only 3 major studies have examined this question. Goldman and colleagues[79] analyzed 100 randomly selected autopsies from the years 1960, 1970, and 1980 at a single institution in Boston and found that the rate of misdiagnosis was stable over time. A more recent study in Germany used a similar approach to study autopsies over a range of 4 decades, from 1959 to 1989. Although the autopsy rate decreased over these years from 88% to 36%, the misdiagnosis rate was stable.[78]

Shojania and colleagues[80] propose that the near-constant rate of misdiagnosis found at autopsy over the years probably reflects 2 factors that offset each other: diagnostic accuracy actually has improved over time (more knowledge, better tests, more skills), but as the autopsy rate declines, there is a tendency to select only the more challenging clinical cases for autopsy, which then have a higher likelihood of diagnostic error. A longitudinal study of autopsies in Switzerland (constant 90% autopsy rate) supports that the absolute rate of diagnostic errors is, as suggested, decreasing over time.[81]

Summary

In aggregate, studies consistently demonstrate a rate of diagnostic error that ranges from <5% in the perceptual specialties (pathology, radiology, dermatology) up to 10% to 15% in most other fields.

It should be noted that the accuracy of clinical diagnosis in practice may differ from that suggested by most studies assessing error rates. Some of the variability in the estimates of diagnostic errors described may be attributed to whether researchers first evaluated diagnostic errors (not all of which will lead to an adverse event) or adverse events (which will miss diagnostic errors that do not cause significant injury or disability). In addition, basing conclusions about the extent of misdiagnosis on the patients who died and had an autopsy, or who filed malpractice claims, or even who had a serious disease leads to overestimates of the extent of errors, because such samples are not representative of the vast majority of patients seen by most clinicians. On the other hand, given the fragmentation of care in the outpatient setting, the difficulty of tracking patients, and the amount of time it often takes for a clear picture of the disease to emerge, these data may actually underestimate the extent of error, especially in ambulatory settings.[82] Although the exact frequency may be difficult to determine precisely, it is clear that an extensive and ever-growing literature confirms that diagnostic errors exist at nontrivial and sometimes alarming rates. These studies span every specialty and virtually every dimension of both inpatient and outpatient care.

PHYSICIAN OVERCONFIDENCE

". . . what discourages autopsies is medicine's twenty-first century, tall-in-the-saddle confidence."
"When someone dies, we already know why. We don't need an autopsy to find out. Or so I thought."
—Atul Gawande[83]

"He who knows best knows how little he knows."
—attributed to Thomas Jefferson[84]

"Doctors think a lot of patients are cured who have simply quit in disgust."
—attributed to Don Herold[85]

As Kirch and Schafii[78] note, autopsies not only document the presence of diagnostic errors, they also provide an opportunity to learn from one's errors (*errando discimus*) if one takes advantage of the information. The rate of autopsy in the United States is not measured any more, but is widely assumed to be significantly <10%. To the extent that this important feedback mechanism is no longer a realistic option, clinicians have an increasingly distorted view of their own error rates. In addition to the lack of autopsies, as the above quote by Gawande indicates, physician overconfidence may prevent them from taking advantage of these important lessons. In this section, we review studies related to physician overconfidence and explore the possibility that this is a major factor contributing to diagnostic error.[86] Overconfidence may have both attitudinal as well as cognitive components and should be distinguished from complacency.

There are several reasons for separating the various aspects of overconfidence and complacency: (1) Some areas have undergone more research than others. (2) The strategies for addressing these 2 qualities may be different. (3) Some aspects are more amenable to being addressed than others. (4) Some may be a more frequent cause of misdiagnoses than others.

Attitudinal Aspects of Overconfidence

This aspect (i.e., "I know all I need to know") is reflected within the more pervasive attitude of *arrogance*, an outlook

that expresses disinterest in any decision support or feedback, regardless of the specific situation.

Comments like those quoted at the beginning of this section reflect the perception that physicians are arrogant and pervasively overconfident about their abilities; however, the data on this point are mostly indirect. For example, the evidence discussed above—that autopsies are on the decline despite their providing useful data—inferentially provides support for the conclusion that physicians do not think they need diagnostic assistance. Substantially more data are available on a similar line of evidence, namely, the general tendency on the part of physicians to disregard, or fail to use, decision-support resources.

Knowledge-Seeking Behavior. Research shows that physicians admit to having many questions that could be important at the point of care, but which they do not pursue.[87–89] Even when information resources are automated and easily accessible at the point of care with a computer, Rosenbloom and colleagues[90] found that a tiny fraction of the resources were actually used. Although the method of accessing resources affected the degree to which they were used, even when an indication flashed on the screen that relevant information was available, physicians rarely reviewed it.

Response to Guidelines and Decision-Support Tools. A second area related to the attitudinal aspect is research on physician response to clinical guidelines and to output from computerized decision-support systems, often in the form of guidelines, alerts, and reminders. A comprehensive review of medical practice in the United States found that the care provided deviated from recommended best practices half of the time.[91] For many conditions, consensus exists on the best treatments and the recommended goals; nevertheless, these national clinical guidelines have a high rate of noncompliance.[92,93] The treatment of high cholesterol is a good example: although 95% of physicians were aware of lipid treatment guidelines from a recent study, they followed these guidelines only 18% of the time.[94] Decision-support tools have the potential to improve care and decrease variations in care delivery, but, unfortunately, clinicians disregard them, even in areas where care is known to be suboptimal and the support tool is well integrated into their workflow.[95–99]

In part, this disregard reflects the inherent belief on the part of many physicians that their practice conforms to consensus recommendations, when in fact it does not. For example, Steinman and colleagues[100] were unable to find a significant correlation between perceived and actual adherence to hypertension treatment guidelines in a large group of primary care physicians.

Similarly, because treatment guidelines are frequently dependent on accurate diagnoses, if the clinician does not recognize the diagnosis, the guideline may not be invoked. For instance, Tierney and associates[101] implemented computer-based guidelines for asthma that did not work successfully, in part because physicians did not consider certain cases to be asthma even though they met identified clinical criteria for the condition.

Timmermans and Mauck[102] suggest that the high rate of noncompliance with clinical guidelines relates to the sociology of what it means to be a professional. Being a professional connotes possessing expert knowledge in an area and functioning relatively autonomously. In a similar vein, Tanenbaum[103] worries that evidence-based medicine will decrease the "professionalism" of the physician. van der Sijs and colleagues[104] suggest that the frequent overriding of computerized alerts may have a positive side in that it shows clinicians are not becoming overly dependent on an imperfect system. Although these authors focus on the positive side to professionalism, the converse, a pervasive attitude of overconfidence, is certainly a possible explanation for the frequent overrides. At the very least, as Katz[105] noted many years ago, the discomfort in admitting uncertainty to patients that many physicians feel can mask inherent uncertainties in clinical practice even to the physicians themselves. Physicians do not tolerate uncertainty well, nor do their patients.

Cognitive Aspects of Overconfidence

The cognitive aspect (i.e., "not knowing what you don't know") is situation specific, that is, in a particular instance, the clinician thinks he/she has the correct diagnosis, but is wrong. Rarely, the reason for not knowing may be lack of knowledge per se, such as seeing a patient with a disease that the physician has never encountered before. More commonly, cognitive errors reflect problems gathering data, such as failing to elicit complete and accurate information from the patient; failure to recognize the significance of data, such as misinterpreting test results; or most commonly, failure to synthesize or "put it all together."[106] This typically includes a breakdown in clinical reasoning, including using faulty heuristics or "cognitive dispositions to respond," as described by Croskerry.[107] In general, the cognitive component also includes a failure of metacognition (the willingness and ability to reflect on one's own thinking processes and to critically examine one's own assumptions, beliefs, and conclusions).

Direct Evidence of Overconfidence. A direct approach to studying overconfidence is to simply ask physicians how confident they are in their diagnoses. Studies examining the cognitive aspects of overconfidence generally have examined physicians' expressed confidence in specific diagnoses, usually in controlled "laboratory" settings rather than studies in actual practice settings. For instance, Friedman and colleagues[108] used case scenarios to examine the accuracy of physicians', residents', and medical students' actual diagnoses compared with how confident they were that their diagnoses were correct. The researchers found that residents had the greatest mismatch. That is, medical students were

both least accurate and least confident, whereas attending physicians were the most accurate and highly confident. Residents, on the other hand, were more confident about the correctness of their diagnoses, but they were less accurate than the attending physicians.

Berner and colleagues,[99] while not directly assessing confidence, found that residents often stayed wedded to an incorrect diagnosis even when a diagnostic decision support system suggested the correct diagnosis. Similarly, experienced dermatologists were confident in diagnosing melanoma in >50% of test cases, but were wrong in 30% of these decisions.[109] In test settings, physicians are also overconfident in treatment decisions.[110] These studies were done with simulated clinical cases in a formal research setting and, although suggestive, it is not clear that the results would be the same with cases seen in actual practice.

Concrete and definite evidence of overconfidence in medical practice has been demonstrated at least twice, using autopsy findings as the gold standard. Podbregar and colleagues[111] studied 126 patients who died in the ICU and underwent autopsy. Physicians were asked to provide the clinical diagnosis and also their level of uncertainty: level 1 represented complete certainty, level 2 indicated minor uncertainty, and level 3 designated major uncertainty. The rates at which the autopsy showed significant discrepancies between the clinical and postmortem diagnosis were essentially identical in all 3 of these groups. Specifically, clinicians who were "completely certain" of the diagnosis antemorten were wrong 40% of the time.[111] Similar findings were reported by Landefeld and coworkers[112]: the level of physician confidence showed no correlation with their ability to predict the accuracy of their clinical diagnosis. Additional direct evidence of overconfidence has been demonstrated in studies of radiologists given sets of "unknown" films to classify as normal or abnormal. Potchen[113] found that diagnostic accuracy varied among a cohort of 95 board-certified radiologists: The top 20 had an aggregate accuracy rate of 95%, compared with 75% for the bottom 20. Yet, the confidence level of the worst performers was actually higher than that of the top performers.

Causes of Cognitive Error. Retrospective studies of the accuracy of diagnoses in actual practice, as well as the autopsy and other studies described previously,[77,106,114,115] have attempted to determine reasons for misdiagnosis. Most of the cognitive errors in diagnosis occur during the "synthesis" step, as the physician integrates his/her medical knowledge with the patient's history and findings.[106] This process is largely subconscious and automatic.

Heuristics. Research on these automatic responses has revealed a wide variety of *heuristics* (subconscious rules of thumb) that clinicians use to solve diagnostic puzzles.[116] Croskerry[107] calls these responses our "cognitive predispositions to respond." These heuristics are powerful clinical tools that allow problems to be solved quickly and, typically, correctly. For example, a clinician seeing a weekend gardener with linear streaks of intensely itchy vesicles on the legs easily diagnoses the patient as having a contact sensitivity to poison ivy using the *availability heuristic*. He or she has seen many such reactions because this is a common problem, and it is the first thing to come to mind. The *representativeness heuristic* would be used to diagnose a patient presenting with chest pain if the pain radiates to the back, varies with posture, and is associated with a cardiac friction rub. This patient has pericarditis, an extremely uncommon reason for chest pain, but a condition with a characteristic clinical presentation.

Unfortunately, the unconscious use of heuristics can also predispose to diagnostic errors. If a problem is solved using the availability heuristic, for example, it is unlikely that the clinician considers a comprehensive differential diagnosis, because the diagnosis is so immediately obvious, or so it appears. Similarly, using the representativeness heuristic predisposes to base rate errors. That is, by just matching the patient's clinical presentation to the prototypical case, the clinician may not adequately take into account that other diseases may be much more common and may sometimes present similarly.

Additional cognitive errors are described below. Of these, premature closure and the context errors are the most common causes of cognitive error in internal medicine.[86]

Premature Closure. Premature closure is narrowing the choice of diagnostic hypotheses too early in the process, such that the correct diagnosis is never seriously considered.[117–119] This is the medical equivalent of Herbert Simon's concept of "satisficing."[120] Once our minds find an adequate solution to whatever problem we are facing, we tend to stop thinking of additional, potentially better solutions.

Confirmation Bias and Related Biases. These biases reflect the tendency to seek out data that confirm one's original idea rather than to seek out disconfirming data.[115]

Context Errors. Very early in clinical problem solving, healthcare practitioners start to characterize a problem in terms of the organ system involved, or the type of abnormality that might be responsible. For example, in the instance of a patient with new shortness of breath and a past history of cardiac problems, many clinicians quickly jump to a diagnosis of congestive heart failure, without consideration of other causes of the shortness of breath. Similarly, a patient with abdominal pain is likely to be diagnosed as having a gastrointestinal problem, although sometimes organs in the chest can present in this fashion. In these situations, clinicians are biased by the history, a previously established diagnosis, or other factors, and the case is formulated in the wrong context.

Clinical Cognition. Relevant research has been conducted on how physicians make diagnoses in the first place. Early

work by Elstein and associates,[121] and Barrows and colleagues[122–124] showed that when faced with what is perceived as a difficult diagnostic problem, physicians gather some initial data and very quickly often within seconds, develop diagnostic hypotheses. They then gather more data to evaluate these hypotheses and finally reach a diagnostic conclusion. This approach has been referred to as a hypothetico-deductive mode of diagnostic reasoning and is similar to the traditional descriptions of the scientific method.[121] It is during this evaluation process that the problems of confirmation bias and premature closure are likely to occur.

Although hypothetico-deductive models may be followed for situations perceived as diagnostic challenges, there is also evidence that as physicians gain experience and expertise, most problems are solved by some sort of pattern-recognition process, either by recalling prior similar cases, attending to prototypical features, or other similar strategies.[125–129] As Eva and Norman[130] and Klein[128] have emphasized, most of the time this pattern recognition serves the clinician well. However, it is during the times when it does not work, whether because of lack of knowledge or because of the inherent shortcomings of heuristic problem solving, that overconfidence may occur.

There is substantial evidence that overconfidence— that is, miscalibration of one's own sense of accuracy and actual accuracy—is ubiquitous and simply part of human nature. Miscalibration can be easily demonstrated in experimental settings, almost always in the direction of overconfidence.[84,131–133] A striking example derives from surveys of academic professionals, 94% of whom rate themselves in the top half of their profession.[134] Similarly, only 1% of drivers rate their skills below that of the average driver.[135] Although some attribute the results to statistical artifacts, and the degree of overconfidence can vary with the task, the inability of humans to accurately judge what they know (in terms of accuracy of judgment or even thinking that they know or do not know something) is found in many areas and in many types of tasks.

Most of the research that has examined expert decision making in natural environments, however, has concluded that rapid and accurate pattern recognition is characteristic of experts. Klein,[128] Gladwell,[127] and others have examined how experts in fields other than medicine diagnose a situation and find that they routinely rapidly and accurately assess the situation and often cannot even describe how they do it. Klein[128] refers to this process as "recognition primed" decision making, referring to the extensive experience of the expert with previous similar cases. Gigerenzer and Goldstein[136] similarly support the concept that most real-world decisions are made using automatic skills, with "fast and frugal" heuristics that lead to the correct decisions with surprising frequency.

Again, when experts recognize that the pattern is incorrect they may revert back to a hypothesis testing mode or may run through alternative scripts of the situation. Expertise is characterized by the ability to recognize when one's initial impression is wrong and to having back-up strategies readily available when the initial strategy does not work.

Hamm[137] has suggested that what is known as the cognitive continuum theory can explain some of the contradictions as to whether experts follow a hypothetico-deductive or a pattern-recognition approach. The cognitive continuum theory suggests that clinical judgment can appropriately range from more intuitive to more analytic, depending on the task. Intuitive judgment, as Hamm conceives it, is not some vague sense of intuition, but is really the rapid pattern acteristic of experts in many situations. Although intuitive judgment may be most appropriate in the uncertain, fast-paced field environment where Klein observed his subjects, other strategies might best suit the laboratory environment that others use to study decision making. In addition, forcing research subjects to verbally explain their strategies, as done in most experimental studies of physician problem solving, may lead to the hypothetico-deductive description. In contrast, Klein,[128] who studied experts in field situations, found his subjects had a very difficult time articulating their strategies.

Even if we accept that a pattern-recognition strategy is appropriate under some circumstances and for certain types of tasks, we are still left with the question as to whether overconfidence is in fact a significant problem. Gigerenzer[138] (like Klein) feels that most of the formal studies of cognition leading to the conclusion of overconfidence use tasks that are not representative of decision making in the real world, either in content or in difficulty. As an example, to study diagnostic problem solving, most researchers of necessity use "diagnostically challenging cases,"[139] which are clearly not typical of the range of cases seen in clinical practice. The zebra adage (i.e., when you hear hoofbeats think of horses, not zebras) may for the most part be adaptive in the clinicians' natural environment, where zebras are much rarer than horses. However, in experimental studies of clinician diagnostic decision making, the reverse is true. The challenges of studying clinicians' diagnostic accuracy in the natural environment are compounded by the fact that most initial diagnoses are made in ambulatory settings, which are notoriously difficult to assess.[82]

Complacency Aspect of Overconfidence

Complacency (i.e., "nobody's perfect") reflects a combination of underestimation of the amount of error, tolerance of error, and the belief that errors are inevitable. Complacency may show up as thinking that misdiagnoses are more infrequent than they actually are, that the problem exists but not in the physician's own practice, that other problems are more important to address, or that nothing can be done to minimize diagnostic errors.

Given the overwhelming evidence that diagnostic error exists at nontrivial rates, one might assume that physicians would appreciate that such error is a serious problem. Yet this is not the case. In 1 study, family physicians asked to recall memorable errors were able to recall very few.[140]

However, 60% of those recalled were diagnostic errors. When giving talks to groups of physicians on diagnostic errors, Dr. Graber (coauthor of this article) frequently asks whether they have made a diagnostic error in the past year. Typically, only 1% admit to having made a diagnostic error. The concept that they, personally, could err at a significant rate is inconceivable to most physicians.

While arguing that clinicians grossly underestimate their own error rates, we accept that they are generally aware of the problem of medical error, especially in the context of medical malpractice. Indeed, 93% of physicians in formal surveys reported that they practice "defensive medicine," including ordering unnecessary lab tests, imaging studies, and consultations.[141] The cost of defensive medicine is estimated to consume 5% to 9% of healthcare expenditures in the United States.[142] We conclude that physicians acknowledge the possibility of error, but believe that mistakes are made by others.

The remarkable discrepancy between the known prevalence of error and physician perception of their own error rate has not been formally quantified and is only indirectly discussed in the medical literature, but lies at the crux of the diagnostic error puzzle, and explains in part why so little attention has been devoted to this problem. Physicians tend to be overconfident of their diagnoses and are largely unaware of this tendency at any conscious level. This may reflect either inherent or learned behaviors of self-deception. Self-deception is thought to be an everyday occurrence, serving to emphasize to others our positive qualities and minimize our negative ones.[143] From the physician's perspective, such self-deception can have positive effects. For example, it can help foster the patient's perception of the physician as an all-knowing healer, thus promoting trust, adherence to the physician's advice, and an effective patient-physician relationship.

Other evidence for complacency can be seen in data from the review by van der Sijs and colleagues.[104] The authors cite several studies that examined the outcomes of the overrides of automated alerts, reminders, and guidelines. In many cases, the overrides were considered clinically justified, and when they were not, there were very few ($\leq$3%) adverse events as a result. While it may be argued that even those few adverse events could have been averted, such contentions may not be convincing to a clinician who can point to adverse events that occur even with adherence to guidelines or alerts. Both types of adverse events may appear to be unavoidable and thus reinforce the physician's complacency.

Gigerenzer,[138] like Eva and Norman[130] and Klein,[128] suggests that many strategies used in diagnostic decision making are adaptive and work well most of the time. For instance, physicians are likely to use data on patients' health outcome as a basis for judging their own diagnostic acumen. That is, the physician is unconsciously evaluating the number of clinical encounters in which patients improve compared with the overall number of visits in a given period of time, or more likely, over years of practice. The denominator that the clinician uses is clearly not the number of adverse events, which some studies of diagnostic errors have used. Nor is it a selected sample of challenging cases, as others have cited. Because most visits are not diagnostically challenging, the physician not only is going to diagnose most of these cases appropriately but he/she also is likely to get accurate feedback to that effect, in that most patients (1) do not wind up in the hospital, (2) appear to be satisfied when next seen, or (3) do not return for the particular complaint because they are cured or treated appropriately.

Causes of inadequate feedback include patients leaving the practice, getting better despite the wrong diagnosis, or returning when symptoms are more pronounced and thus eventually getting diagnosed correctly. Because immediate feedback is not even expected, feedback that is delayed or absent may not be recognized for what it is, and the perception that "misdiagnosis is not a big problem" remains unchallenged. That is, in the absence of information that the diagnosis is wrong, it is assumed to be correct ("no news is good news"). This phenomenom is illustrated in epigraph above from Herold, "Doctors think a lot of patients are cured who have simply quit in disgust."[85] The perception that misdiagnosis is not a major problem, while not necessarily correct, may indeed reflect arrogance, "tall in the saddle confidence,"[83] or "omniscience."[144] Alternatively, it may simply reflect that over all the patient encounters a physician has, the number of diagnostic errors *of which he or she is aware* is very low.

Thus, despite the evidence that misdiagnoses do occur more frequently than often presumed by clinicians, and despite the fact that recognizing that they do occur is the first step to correcting the problem, the assumption that misdiagnoses are made only a very small percentage of the time can be seen as a rational conclusion given the current healthcare environment where feedback is limited and only selective outcome data are available for physicians to accurately calibrate the extent of their own misdiagnoses.

Summary

Pulling together the research described above, we can see why there may be complacency and why it is difficult to address. First, physicians generate hypotheses almost immediately upon hearing a patient's initial symptom presentation and in many cases these hypotheses suggest a familiar pattern. Second, even if more exploration is needed, the most likely information sought is that which confirms the initial hypothesis; often, a decision is reached without full exploration of a large number of other possibilities. In the great majority of cases, this approach leads to the correct diagnosis and a positive outcome. The patient's diagnosis is made quickly and correctly, treatment is initiated, and both the patient and physician feel better. This explains why this approach is used, and why it is so difficult to change. In addition, in many of the cases where the diagnosis is incorrect, the physician never knows it. If the diagnostic process

routinely led to errors that the physician recognized, they could get corrected. Additionally, the physician might be humbled by the frequent oversights and become inclined to adopt a more deliberate, contemplative approach or develop strategies to better identify and prevent the misdiagnoses.

STRATEGIES TO IMPROVE THE ACCURACY OF DIAGNOSTIC DECISION MAKING

"Ignorance more frequently begets confidence than does knowledge."

—Charles Darwin, 1871[145]

We believe that strategies to reduce misdiagnoses should focus on physician calibration, i.e., improving the match between the physician's self-assessment of errors and actual errors. Klein[128] has shown that experts use their intuition on a routine basis, but rethink their strategies when that does not work. Physicians also rethink their diagnoses when it is obvious that they are wrong. In fact, it is in these situations that diagnostic decision-support tools are most likely to be used.[146]

The challenge becomes how to increase physicians' awareness of the possibility of error. In fact, it could be argued that their awareness needs to be increased for a select type of case: that in which the healthcare provider thinks he/she is correct and does not receive any timely feedback to the contrary, but where he/she is, in fact, mistaken. Typically, most of the clinician's cases are diagnosed correctly; these do not pose a problem. For the few cases where the clinician is consciously puzzled about the diagnosis, it is likely that an extended workup, consultation, and research into possible diagnoses occurs. It is for the cases that fall between these types, where miscalibration is present but unrecognized, that we need to focus on strategies for increasing physician awareness and correction.

If overconfidence, or more specifically, miscalibration, is a problem, what is the solution? We examine 2 broad categories of solutions: strategies that focus on the individual and system approaches directed at the healthcare environment in which diagnosis takes place. The individual approaches assume that the physician's cognition needs improvement and focus on making the clinician smarter, a better thinker, less subject to biases, and more cognizant of what he or she knows and does not know. System approaches assume that the individual physician's cognition is adequate for the diagnostic and metacognitive tasks, but that he/she needs more, and better, data to improve diagnostic accuracy. Thus, the system approaches focus on changing the healthcare environment so that the data on the patients, the potential diagnoses, and any additional information are more accurate and accessible. These 2 approaches are not mutually exclusive and the major aim of both is to improve the physician's calibration between his/her perception of the case and the actual case. Theorectically, if improved calibration occurs, overconfidence should decrease, including the attitudinal components of arrogance and complacency.

In the discussion about individually focused solutions, we review the effectiveness of clinical education and practice, development of metacognitive skills, and training in reflective practice. In the section on systems-focused solutions, we examine the effectiveness of providing performance feedback, the related area of improving follow-up of patients and their health outcomes, and using automation—such as providing general knowledge resources at the point of care and specific diagnostic decision-support programs.

Strategies that Focus on the Individual

Education, Training and Practice. By definition, experts are smarter, e.g., more knowledgeable than novices. A fascinating (albeit frightening) observation is the general tendency of novices to overrate their skills.[84,108,132] Exactly the same tendency is seen in testing of medical trainees in regard to skills such as communicating with patients.[147] In a typical experiment a cohort with varying degrees of expertise are asked to undertake a skilled task. At the completion of the task, the test subjects are asked to grade their own performance. When their self-rated scores are compared with the scores assigned by experts, the individuals with the lowest skill levels predictably overestimate their performance.

Data from a study conducted by Friedman and colleagues[108] showed similar results: residents in training performed worse than faculty physicians, but were more confident in the correctness of their diagnoses. A systematic review of studies assessing the accuracy of physicians' self-assessment of knowledge compared with an external measure of competence showed very little correlation between self-assessment and objective data.[148] The authors also found that those physicians who were least expert tended to be most overconfident in their self-assessments.

These observations suggest a possible solution to overconfidence: make physicians more expert. The expert is better calibrated (i.e. better assesses his/her own accuracy), and excels at distinguishing cases that are easily diagnosed from those that require more deliberation. In addition to their enhanced ability to make this distinction, experts are likely to make the correct diagnosis more often in both recognized as well as unrecognized cases. Moreover, experts carry out these functions automatically, more efficiently, and with less resource consumption than nonexperts.[127,128]

The question, of course, is how to develop that expertise. Presumably, thorough medical training and continuing education for physicians would be useful; however, data show that the effects on actual practice of many continuing education programs are minimal.[149–151] Another approach is to advocate the development of expertise in a narrow domain. This strategy has implications for both individual clinicians and healthcare systems. At the level of the individual clinician, the mandate to become a true expert would drive more trainees into subspecialty training and emphasize development of a comprehensive knowledge base.

Another mechanism for gaining knowledge is to gain more extensive practice and experience with actual clinical

cases. Both Bordage[152] and Norman[151,153] champion this approach, arguing that "practice is the best predictor of performance." Having a large repertoire of mentally stored exemplars is also the key requirement for Gigerenzer's "fast and frugal"[136,138] and Klein's[128] "recognition-primed" decision making. Extensive practice with simulated cases may supplement, although not supplant, experience with real ones. The key requirements in regard to clinical practice are extensive, i.e., necessitating more than just a few cases and occasional feedback.

Metacognitive Training and Reflective Practice. In addition to strategies that aim to increase the overall level of clinicians' knowledge, other educational approaches focus on increasing physicians' self-awareness so that they can recognize when additional information is needed or the wrong diagnostic path is taken. One such approach is to increase what has been called "situational awareness,'" the lack of which has been found to lie behind errors in aviation.[154] Singh and colleagues[154] advocate this strategy; their definition of types of situational awareness is similar to what others have called metacognitive skills. Croskerry[115,155] and Hall[156] champion the idea that metacognitive training can reduce diagnostic errors, especially those involving subconscious processing. The logic behind this approach is appealing: Because much of intuitive medical decision making involves the use of cognitive dispositions to respond, the assumption is if trainees or clinicians were educated about the inherent biases involved in the use of these strategies, they would be less susceptible to decision errors.

Croskerry[157] has outlined the use of what he refers to as "cognitive forcing strategies" to counteract the tendency to cognitive error. These would orient clinicians to the general concepts of metacognition (a universal forcing strategy), familiarize them with the various heuristics they use intuitively and their associated biases (generic forcing strategies), and train them to recognize any specific pitfalls that apply to the types of patients they see most commonly (specific forcing strategies).

Another noteworthy approach developed by the military, which suggests focusing on a comprehensive conscious view of the proposed diagnosis and how this was derived, is the technique of *prospective hindsight*.[158] Once the initial diagnosis is made, the clinician figuratively gazes into a crystal ball to see the future, sees that the initial diagnosis is not correct, and is thus forced to consider what else it could it be. A related technique, which is taught in every medical school, is to construct a comprehensive differential diagnosis on each case before planning an appropriate workup. Although students and residents excel at this exercise, they rarely use it outside the classroom or teaching rounds. As we discussed earlier, with more experience, clinicians begin to use a pattern-recognition approach rather than an exhaustive differential diagnosis. Other examples of cognitive forcing strategies include advice to always "consider the opposite," or ask "what diagnosis can I not afford to miss?"[76] Evidence that metacognitive training can decrease the rate of diagnostic errors is not yet available, although preliminary results are encouraging.[156]

Reflective practice is an approach defined as the ability of physicians to critically consider their own reasoning and decisions during professional activities.[159] This incorporates the principles of metacognition and 4 additional attributes: (1) the tendency to search for alternative hypotheses when considering a complex, unfamiliar problem; (2) the ability to explore the consequences of these alternatives; (3) a willingness to test any related predictions against the known facts; and (4) openness toward reflection that would allow for better toleration of uncertainty.[160] Experimental studies show that reflective practice enhances diagnostic accuracy in complex situations.[161] However, even advocates of this approach recognize that it is an untested assumption in terms of whether lessons learned in educational settings can transfer to the practice setting.[162]

System Approaches

One could argue that effectively incorporating the education and training described above would require system-level change. For instance, at the level of healthcare systems, in addition to the development of required training and education, a concerted effort to increase the level of expertise of the individual would require changes in staffing policies and access to specialists.

If they are designed to teach the clinician, or at least function as an adjunct to the clinician's expertise, some decision-support tools also serve as systems-level interventions that have the potential to increase the total expertise available. If used correctly, these products are designed to allow the less expert clinician to function like a more expert clinician. Computer- or web-based information sources also may serve this function. These resources may not be very different from traditional knowledge resources (e.g., medical books and journals), but by making them more accessible at the point of care they are likely to be used more frequently (assuming the clinician has the metacognitive skills to recognize when they are needed).

The systems approaches described below are based on the assumption that both the knowledge and metacognitive skills of the healthcare provider are generally adequate. These approaches focus on providing better and more accurate information to the clinician primarily to improve calibration. James Reason's ideas on systems approaches for reducing medical errors have formed the background of the patient safety movement, although they have not been applied specifically to diagnostic errors.[163] Nolan[164] advocates 3 main strategies based on a systems approach: prevention, making error visible, and mitigating the effects of error. Most of the cognitive strategies described above fall into the category of prevention.

The systems approaches described below fall chiefly into the latter two of Nolan's strategies. One approach is to provide expert consultation to the physician. Usually this is done by calling in a consultant or seeking a second opinion.

A second approach is to use automated methods to provide diagnostic suggestions. Usually a diagnostic decision-support system is used once the error is visible (e.g., the clinician is obviously puzzled by the clinical situation). Using the system may prevent an initial misdiagnosis and may also mitigate possible sequelae.

Computer-based Diagnostic Decision Support. A variety of diagnostic decision-support systems were developed out of early expert system research. Berner and colleagues[139] performed a systematic evaluation of 4 of these systems; in 1994, Miller[165] described these and other systems. In a review article. Miller's overall conclusions were that while the niche systems for well-defined specific areas were clearly effective, the perceived usefulness of the more general systems such as Quick Medical Reference (QMR), DXplain, Iliad, Meditel was less certain, despite evidence that they could suggest diagnoses that even expert physicians had not considered. The title, "A Report Card on Computer-Assisted Diagnosis—The Grade Is C," of Kassirer's editorial[166] that accompanied the article by Berner and associates[139] is illustrative of an overall negative attitude toward these systems. In a subsequent study, Berner and colleagues[167] found that less experienced physicians were more likely than more experienced physicians to find QMR useful; some researchers have suggested that these systems may be more useful in educational settings.[168] Lincoln and colleagues[169–171] have shown the effectiveness of the Iliad system in educational settings. Arene and associates[172] showed that QMR was effective in improving residents' diagnoses, but then concluded that it took too much time to learn to use the system.

A similar response was found more recently in a randomized controlled trial of another decision-support system (Problem-Knowledge Couplers (PKC), Burlington, Vt).[173] Users felt that the information provided by PKC was useful, but that it took too much time to use. More disturbing was that use of the system actually increased costs, perhaps by suggesting more diagnoses to rule out. What is interesting about PKC is that in this system the patient rather than the physician enters all the data, so the complaint that the system required too much time most likely reflected physician time to review and discuss the results rather than data entry.

One of the more recent entries into the diagnostic decision-support system arena is Isabel (Isabel Healthcare, Inc., Reston, VA; Isabel Healthcare, Ltd., Haslemere, UK.) which was initially begun as a pediatric system and now is also available for use in adults.[174–178] The available studies using Isabel show that it provides diagnoses that are considered both accurate and relevant by physicians. Both Miller[179] and Berner[180] have reviewed the challenges in evaluating medical diagnostic programs. Basically, it is difficult to determine the gold standard against which the systems should be evaluated, but both investigators advocate that the criterion should be how well the clinician using the computer compares with use of only his/her own cognition.[179,180] Virtually all of the published studies have evaluated these systems only in artificial situations and many of them have been performed by the developers themselves.

The history of these systems is reflective of the overall problem we have demonstrated in other domains: despite evidence that these systems can be helpful, and despite studies showing users are satisfied with their results when they do use them, many physicians are simply reluctant to use decision-support tools in practice.[181] Meditel, QMR, and Iliad are no longer commercially available. DXplain, PKC, and Isabel are still available commercially, but although there may be data on the extent of use, there are no data on how often they are used compared with how often they could/should have been used. The study by Rosenbloom and colleagues,[90] which used a well-integrated, easy-to-access system, showed that clinicians very rarely take advantage of the available opportunities for decision support. Because diagnostic tools require the user to enter the data into the programs, it is likely that their usage would be even lower or that the data entry may be incomplete.

An additional concern is that the output of most of these decision-support programs requires subsequent mental filtering, because what is usually displayed is a (sometimes lengthy) list of diagnostic considerations. As we have discussed previously, not only does such filtering take time,[173] but the user must be able to distinguish likely from unlikely diagnoses, and data show that such recognition can be difficult.[99] Also, as Teich and colleagues[182] noted with other decision-support tools, physicians accept reminders about things they intend to do, but are less willing to accept advice that forces them to change their plans. It is likely that if physicians already have a work-up strategy in mind, or are sure of their diagnoses, they would be less willing to consult such a system. For many clinicians, these factors may make the perceived utility of these systems not worth the cost and effort to use them. That does not mean that they are not potentially useful, but the limited interest in them has made several commercial ventures unsustainable.

In summary, the data on diagnostic decision-support systems in reducing diagnostic errors shows that they can provide what are perceived as useful diagnostic suggestions. Every commercial system also has what amounts to testimonials about its usefulness in real life—stories of how the system helped the clinician recognize a rare disease[146]—but to date their use in actual clinical situations has been limited to those times that the physician is puzzled by a diagnostic problem. Because such puzzles occur rarely, there is not enough use of the systems in real practice situations to truly evaluate their effectiveness.

Feedback and Calibration. A second general category of a systems approach is to design systems to provide feedback to the clinician. Overconfidence represents a mismatch between perceived and actual performance. It is a state of miscalibration that, according to existing paradigms of cognitive psychology, should be correctable by providing feedback. Feedback in general can serve to make the diagnostic

error visible, and timely feedback can mitigate the harm that the initial misdiagnosis might have caused. Accurate feedback can improve the basis on which the clinicians are judging the frequency of events, which may improve calibration.

Feedback is an essential element in developing expertise. It confirms strengths and identifies weaknesses, guiding the way to improved performance. In this framework, a possible approach to reducing diagnostic error, overconfidence, and error-related complacency is to enhance feedback with the goal of improving calibration.[183]

Experiments confirm that feedback can improve performance,[184] especially if the feedback includes cognitive information (for example, why a certain diagnosis is favored) as opposed to simple feedback on whether the diagnosis was correct or not.[185,186] A recent investigation by Sieck and Arkes,[131] however, emphasizes that overconfidence is highly ingrained and often resistant to amelioration by simple feedback interventions.

The timing of feedback is important. Immediate feedback is effective, delayed feedback less so.[187] This is particularly problematic for diagnostic feedback in real clinical settings, outside of contrived experiments, because such feedback often is not available at all, much less immediately or soon after the diagnosis is made. In fact, the gold standard for feedback regarding clinical judgment is the autopsy, which of course can only provide retrospective, not real-time, diagnostic feedback.

Radiology and pathology are the only fields of medicine where feedback has been specifically considered, and in some cases adopted, as a method of improving performance and calibration.

Radiology. The accuracy of radiologic diagnosis is most sharply focused in the area of mammography, where both false-positive and false-negative reports have substantial clinical impact. Of note, a recent study called attention to an interesting difference between radiologists in the United States and their counterparts in the United Kingdom: US radiologists suggested follow-up studies (more radiologic testing, biopsy, or close clinical follow-up) twice as often as UK radiologists, and US patients had twice as many normal biopsies, whereas the cancer detection rates in the 2 countries were comparable.[30] In considering the reasons for this difference in performance, the authors point out that 85% of mammographers in the United Kingdom voluntarily participate in "PERFORMS," an organized calibration process, and 90% of programs perform double readings of mammograms. In contrast, there are no organized calibration exercises in the United States and few programs require "double reads." An additional difference is the expectation for accreditation: US radiologists must read 480 mammograms annually to meet expectations of the Mammography Quality Standards Act, whereas the comparable expectation for UK mammographers is 5,000 mammograms per year.[30]

As an initial step toward performance improvement by providing organized feedback, the American College of Radiology (ACR) recently developed and launched the "RADPEER" process.[188] In this program, radiologists keep track of their agreement with any prior imaging studies they re-review while they are evaluating a current study, and the ACR provides a mechanism to track these scores. Participation is voluntary; it will be interesting to see how many programs enroll in this effort.

Pathology. In response to a *Wall Street Journal* exposé on the problem of false-negative Pap smears, the US Congress enacted the Clinical Laboratory Improvement Act of 1988. This act mandated more rigorous quality measures in regard to cytopathology, including proficiency testing and mandatory reviews of negative smears.[189] Even with these measures in place, however, rescreening of randomly selected smears discloses a discordance rate in the range of 10% to 30%, although only a fraction of these discordances have major clinical impact.[190]

There are no comparable proficiency requirements for anatomic pathology, other than the voluntary "Q-Probes" and "Q-Tracks" programs offered by the College of American Pathologists (CAP). Q-Probes are highly focused reviews that examine individual aspects of diagnostic testing, including preanalytical, analytical, and postanalytical errors. The CAP has sponsored hundreds of these probes. Recent examples include evaluating the appropriateness of testing for β-natriuretic peptides, determining the rate of urine sediment examinations, and assessing the accuracy of send-out tests. Q-Tracks are monitors that "reach beyond the testing phase to evaluate the processes both within and beyond the laboratory that can impact test and patient outcomes."[191] Participating labs can track their own data and see comparisons with all other participating labs. Several monitors evaluate the accuracy of diagnosis by clinical pathologists and cytopathologists. For example, participating centers can track the frequency of discrepancies between diagnoses suggested from Pap smears compared with results obtained from biopsy or surgical specimens. However, a recent review estimated that $<1\%$ of US programs participate in these monitors.[192]

Pathology and radiology are 2 specialties that have pioneered the development of computerized second opinions. Computer programs to overread mammograms and Pap smears have been available commercially for a number of years. These programs point out for the radiologists and cytopathologists suspicious areas that might have been overlooked. After some early studies with positive results that led to approval by the US Food and Drug Administration (FDA), these programs have been commercially available. Now that they have been in use for awhile, however, recently published, large-scale, randomized trials of both programs have raised doubts about their performance in practice.[193–195] A recently completed randomized trial of Pap smear results showed a very slight advantage of the computer programs over unaided cytopathologists,[194] but earlier reports of the trial before completion did not show

any differences.[193] The authors suggest that it may take time for optimal quality to be achieved with a new technique.

In the area of computer-assisted mammography interpretation, a randomized trial showed no difference in cancer detection but an increase in false-positives with the use of the software compared with unaided interpretation by radiologists.[195] It is certainly possible that technical improvements have made later systems better than earlier ones, and, as suggested by Nieminen and colleagues[194] about the Pap smear program, and Hall[196] about the mammography programs, it may take time, perhaps years, for the users to learn how to properly interpret and work with the software. These results highlight that realizing the potential advantages of second opinions (human or automated) may be a challenge.

Autopsy. Sir William Osler championed the belief that medicine should be learned from patients, at the bedside and in the autopsy suite. This approach was espoused by Richard Cabot and many others, a tradition that continues today in the "Clinical Pathological Correlation" (CPC) exercises published weekly in *The New England Journal of Medicine*. Autopsies and CPCs teach more than just the specific medical content; they also illustrate the uncertainty that is inherent in the practice of medicine and effectively convey the concepts of fallibility and diagnostic error.

Unfortunately, as discussed above, autopsies in the United States have largely disappeared. Federal tracking of autopsy rates was suspended a decade ago, at which point the autopsy rate had already fallen to $<7\%$. Most trainees in medicine today will never see an autopsy. Patient safety advocates have pleaded to resurrect the autopsy as an effective tool to improve calibration and reduce overconfidence, but so far to no avail.[144,197]

If autopsies are not generally available, has any other process emerged to provide a comparable feedback experience? An innovative candidate is the "Morbidity and Mortality (M & M) Rounds on the Web" program sponsored by the Agency for Healthcare Research and Quality (AHRQ).[198] This site features a quarterly set of 4 cases, each involving a medical error. Each case includes a comprehensive, well-referenced discussion by a safety expert. These cases are attractive, capsulized gems that, like an autopsy, have the potential to educate clinicians regarding medical error, including diagnostic error. The unknown factor regarding this endeavor is whether these lessons will provide the same impact as an autopsy, which teaches by the principle of learning from one's own mistakes.[78] Local "morbidity and mortality" rounds have the same potential to alert providers to the possibility of error, and the impact of these exercises increases if the patient sustains harm.[199]

A final option to provide feedback in the absence of a formal autopsy involves detailed postmortem magnetic resonance imaging scanning. This option obviates many of the traditional objections to an autopsy, and has the potential to reveal many important diagnostic discrepancies.[200]

Feedback in Other Field Settings (The Questec Experiment). A fascinating experiment is underway that could substantially clarify the power of feedback to improve calibration and performance. This is the Questec experiment sponsored by Major League Baseball to improve the consistency of umpires in calling balls and strikes. Questec is a company that installs cameras in selected stadiums that track the ball path across home plate. At the end of the game, the umpire is provided a recording that replays every pitch, and gives him the opportunity to compare the called balls and strikes with the true ball path.[201] Umpires have vigorously objected to this project, including a planned civil lawsuit to stop the experiment. The results from this study have yet to be released, but they will certainly shed light on the question of whether a skeptical cohort of professionals can improve their performance through directed feedback.

Follow-up. A systems approach recommended by Redelmeier[76] and Gandhi et al[77] is to promote the use of follow-up. Schiff[31,75] also has long advocated the importance of follow-up and tracking to improve diagnoses. Planned follow-up after the initial diagnosis allows time for other thoughts to emerge, and time for the clinician to apply more conscious problem-solving strategies (such as decision-support tools) to the problem. A very appealing aspect of planned follow-up is that a patient's problems will evolve over the intervening period, and these changes will either support the original diagnostic possibilities, or point toward alternatives. If the follow-up were done soon enough, this approach might also mitigate the potential harm of diagnostic error, even without solving the problem of how to prevent cognitive error in the first place.

ANALYSIS OF STRATEGIES TO REDUCE OVERCONFIDENCE

The strategies suggested above, even if they are successful in addressing the problem of overconfidence or miscalibration, have limitations that must be acknowledged. One involves the trade-offs of time, cost, and accuracy. We can be more certain, but at a price.[202] A second problem is unanticipated negative effects of the intervention.

Tradeoffs in Time, Cost, and Accuracy

As clinicians improve their diagnostic competency from beginning level skills to expert status, reliability and accuracy improve with decreased cost and effort. However, using the strategies discussed earlier to move nonexperts into the realm of experts will involve some expense. In any given case, we can improve diagnostic accuracy but with increased cost, time, or effort.

Several of the interventions entail direct costs. For instance, expenditures may be in the form of payment for consultation or purchasing diagnostic decision-support systems. Less tangible costs relate to clinician time. Attending training programs involves time, effort, and money. Even

strategies that do not have direct expenses may still be costly in terms of physician time. Most medical decision making takes place in the "adaptive subconscious." The application of expert knowledge, pattern and script recognition, and heuristic synthesis takes place essentially instantaneously for the vast majority of medical problems. The process is effortless. If we now ask physicians to reflect on how they arrived at a diagnosis, the extra time and effort required may be just enough to discourage this undertaking.

Applying conscious review of subconscious processing hopefully uncovers at least some of the hidden biases that affect subconscious decisions. The hope is that these events outnumber the new errors that may evolve as we second-guess ourselves. However, it is not clear that conscious articulation of the reasoning process is an accurate picture of what really occurs in expert decision making. As discussed above, even reviewing the suggestions from a decision-support system (which would facilitate reflection) is perceived as taking too long, even though the information is viewed as useful.[173] Although these arguments may not be persuasive to the individual patient,[2] it is clear that the time involved is a barrier to physician use of decision aids. Thus, in deciding to use methods to increase reflection, decisions must be made as to: (1) whether the marginal improvements in accuracy are worth the time and effort and, given the extra time involved, (2) how to ensure that clinicians will routinely make the effort.

Unintended Consequences

Innovations made in the name of improving safety sometimes create new opportunities to fail, or have unintended consequences that decrease the expected benefit. In this framework, we should carefully examine the possibility that some of the interventions being considered might actually increase the risk of diagnostic error.

As an example, consider the interventions we have grouped under the general heading of "reflective practice." Most of the education and feedback efforts, and even the consultation strategies, are aimed at increasing such reflection. Imagine a physician who has just interviewed and examined an elderly patient with crampy abdominal pain, and who has concluded that the most likely explanation is constipation. What is the downside of consciously reconsidering this diagnosis before taking action?

It Takes More Time. The extra time the reflective process takes not only affects the physician but may have an impact on the patient as well. The extra time devoted to this activity may actually delay the diagnosis for one patient and may be time subtracted from another.

It Can Lead to Extra Testing. As other possibilities are envisioned, additional tests and imaging may be ordered. Our patient with simple constipation now requires an abdominal CT scan. This greatly increases the chances of discovering incidental findings and the risk of inducing *cascade effects*, where one thing leads to another, all of them extraneous to the original problem.[203] Not only might these pose additional risks to the patient, such testing is also likely to increase costs.[173] The risk of changing a "right" diagnosis to a "wrong" one will necessarily increase as the number of options enlarges; research has found that this sometimes occurs in experimental settings.[99,168]

It May Change the Patient-Physician Dynamic. Like physicians, most patients much prefer certainty over ambiguity. Patients want to believe that their healthcare providers know exactly what their disorder is, and what to do about it. An approach that lays out all the uncertainties involved and the probabilistic nature of medical decisions is unlikely to be warmly received by patients unless they are highly sophisticated. A patient who is reassured that he or she most likely has constipation will probably sleep a lot better than the one who is told that the abdominal CT scan is needed to rule out more serious concerns.

The Risk of Diagnostic Error May Actually Increase. The quality of automatic decision making may be degraded if subjected to conscious inspection. As pointed out in *Blink*,[127] we can all easily envision Marilyn Monroe, but would be completely stymied in attempting to describe her well enough for a stranger to recognize her from a set of pictures. There is, in fact, evidence that complex decisions are solved best without conscious attention.[204] A complementary observation is that the quality of conscious decision making degrades as the number of options to be considered increases.[205]

Increased Reliance on Consultative Systems May Result in "Deskilling." Although currently the diagnostic decision-support systems claim that they are only providing suggestions, not "the definitive diagnosis,"[206]there is a tendency on the part of users to believe the computer. Tsai and colleagues[207] found that residents reading electrocardiograms improved their interpretations when the computer interpretation was correct, but were worse when it was incorrect. A study by Galletta and associates[208] using the spell-checker in a word-processing program found similar results. There is a risk that, as the automated programs get more accurate, users will rely on them and lose the ability to tell when the systems are incorrect.

A summary of the strategies, their assumptions, which may not always be accurate, and the tradeoffs in implementing them is shown in Table 2.

RECOMMENDATIONS FOR FUTURE RESEARCH

"Happy families are all alike; every unhappy family is unhappy in its own way."

—Leo Tolstoy, *Anna Karenina*[209]

We are left with the challenge of trying to consider solutions based on our current understanding of the research

Table 2 Strategies to Reduce Diagnostic Errors

Strategy	Purpose	Timing	Focus	Underlying Assumptions	Tradeoffs
Education and training					
Training in reflective practice and avoidance of biases	Provide metacognitive skills	Not tied to specific patient cases	Individual, prevention	Transfer from educational to practice setting will occur; clinician will recognize when thinking is incorrect	Not tied to action: expensive and time consuming except in defined educational settings
Increase expertise	Provide knowledge and experience	Not tied to specific patient cases	Individual, prevention	Transfer across cases will occur; errors are a result of lack of knowledge or experience	Expensive and time consuming except in defined educational settings
Consultation					
Computer-based general knowledge resources	Validate or correct initial diagnosis; suggest alternatives	At the point-of-care while considering diagnosis	Individual, prevention	Users will recognize the need for information and will use the feedback provided	Delay in action; most sources still need better indexing to improve speed of accessing information
Second opinions/ consult with experts	Validate or correct initial diagnosis	Before treatment of specific patient	System, prevention/ mitigation	Expert is correct and/or agreement would mean diagnosis is correct	Delay in action; expense, bottlenecks, may need 3rd opinion if there is disagreement; if not mandatory would be only used for cases where physician is puzzled
DDSS	Validate or correct initial diagnosis	Before definitive diagnosis of specific patient	System, prevention	DDSS suggestions would include correct diagnosis; physician will recognize correct diagnosis when DDSS suggests it	Delay in action, cost of system; if not mandatory for all cases would be only used for cases where physician is puzzled
Feedback					
Increase number of autopsies/M&M	Prevent future errors	After an adverse event or death has occurred	System, prevention in future	Clinician will learn from errors and will not make them again; feedback will improve calibration	Cannot change action, too late for specific patient, expensive
Audit and feedback	Prevent future errors	At regular intervals covering multiple patients seen over a given period	System, prevention in future	Clinician will learn from errors and will not make them again; feedback will improve calibration	Cannot change action, too late for specific patient, expensive
Rapid follow-up	Prevent future errors and mitigate harm from errors for specific patient	At specified intervals unique to specific patients shortly after diagnosis or treatment	System, mitigation	Error may not be preventable, but harm in selected cases may be mitigated; feedback will improve calibration	Expense, change in workflow, MD time in considering problem areas

DDSS = diagnostic decision-support system; MD = medical doctor; M&M = morbidity and mortality.

on overconfidence and the strategies to overcome it. Studies show that experts seem to know what to do in a given situation and what they know works well most of the time. What this means is that diagnoses are correct most of the time. However, as advocated in the Institute of Medicine (IOM) reports, the engineering principle of "design for the usual, but plan for the unusual" should apply to this situation.[210] As Gladwell[211] discussed in an article in *The New Yorker* on homelessness, however, the solutions to address the "unusual" (or the "unhappy families" referenced in the epigraph above) may be very different from those that work for the vast majority of cases. So while we are not advocating complacency in the face of error, we are assuming that some errors will escape our prevention. For these situations, we must have contingency plans in place for reducing the harm ensuing from them.

If we look at the aspects of overconfidence discussed in this review, the cognitive and systemic factors appear to be more easily addressed than the attitudinal issues and those related to complacency. However, the latter two may be affected by addressing the former ones. If physicians were better calibrated, i.e., knew accurately when they were correct or incorrect, arrogance and complacency would not be a problem.

Our review demonstrates that while all of the methods to reduce diagnostic error can potentially reduce misdiagnosis, none of the educational approaches are systematically used outside the initial educational setting and when automated devices operate in the background they are not used uniformly. Our review also shows that on some level, physicians' overconfidence in their own diagnoses and complacency in the face of diagnostic error can account for the lack of use. That is, given information and incentives to examine and modify one's initial diagnoses, physicians choose not to undertake the effort. Given that physicians in general are reasonable individuals, the only feasible explanation is that they believe that their initial diagnoses are correct (even when they are not) and there is no reason for change. We return to the problem that prompted this literature review, but with a more focused research agenda to address the areas listed below.

Overconfidence

Because most studies actually addressed overconfidence indirectly and usually in laboratory as opposed to real-life settings, we still do not know the prevalence of overconfidence in practice, whether it is the same across specialties, and what its direct role is in misdiagnosis.

Preventability of Diagnostic Error

One of the glaring issues that is unresolved in the research to date is the extent to which diagnostic errors are preventable. The answer to this question will influence error-reduction strategies.

Mitigating Harm

More research and evaluation of strategies that focus on mitigating the harm from the errors is needed. The research approach should include what Nolan has called "making the error visible."[164] Because these errors are likely the ones that have traditionally been unrecognized, focusing research on them can provide better data on how extensively they occur in routine practice. Most strategies for addressing diagnostic errors have focused on prevention; it is in the area of mitigation where the strategies are sorely lacking.

Debiasing

Is instruction on cognitive error and cognitive forcing strategies effective at improving diagnosis? What is the best stage of medical education to introduce this training? Does it transfer from the training to the practice setting?

Feedback

How much feedback do physicians get and how much do they need? What mechanisms can be constructed to get them more feedback on their own cases? What are the most effective ways to learn from the mistakes of others?

Follow-up

How can planned follow-up of patient outcomes be encouraged and what approaches can be used for rapid follow-up to provide more timely feedback on diagnoses?

Minimizing the Downside

Does conscious attention decrease the chances of diagnostic error or increase it? Can we think of ways to minimize the possibility that conscious attention to diagnosis may actually make things worse?

CONCLUSIONS

Diagnostic error exists at an appreciable rate, ranging from $<5\%$ in the perceptual specialties up to 15% in most other areas of medicine. In this review, we have examined the possibility that overconfidence contributes to diagnostic error. Our review of the literature leads us to 2 main conclusions.

Physicians Overestimate the Accuracy of Their Diagnoses

Overconfidence exists and is probably a trait of human nature—we all tend to overestimate our skills and abilities. Physicians' overconfidence in their decision making may simply reflect this tendency. Physicians come to trust the fast and frugal decision strategies they typically use. These strategies succeed so reliably that physicians can become complacent; the failure rate is minimal and errors may not come to their attention for a variety of reasons. Physicians acknowledge that diagnostic error exists, but seem to believe that the likelihood of error is less than it really is. They

believe that they personally are unlikely to make a mistake. Indirect evidence of overconfidence emerges from the routine disregard that physicians show for tools that might be helpful. They rarely seek out feedback, such as autopsies, that would clarify their tendency to err, and they tend not to participate in other exercises that would provide independent information on their diagnostic accuracy. They disregard guidelines for diagnosis and treatment. They tend to ignore decision-support tools, even when these are readily accessible and known to be valuable when used.

Overconfidence Contributes to Diagnostic Error

Physicians in general have well-developed metacognitive skills, and when they are uncertain about a case they typically devote extra time and attention to the problem and often request consultation from specialty experts. We believe many or most cognitive errors in diagnosis arise from the cases where they *are* certain. These are the cases where the problem appears to be routine and resembles similar cases that the clinician has seen in the past. In these situations, the metacognitive angst that exists in more challenging cases may not arise. Physicians may simply stop thinking about the case, predisposing them to all of the pitfalls that result from our cognitive "dispositions to respond." They fail to consider other contexts or other diagnostic possibilities, and they fail to recognize the many inherent shortcomings that derive from heuristic thinking.

In summary, improving patient safety will ultimately require strategies that take into account the data from this review—why diagnostic errors occur, how they can be prevented, and how the harm that results can be reduced.

ACKNOWLEDGMENTS

We are grateful to Paul Mongerson for encouragement and financial support of this research. The authors also appreciate the insightful comments of Arthur S. Elstein, PhD, on an earlier draft of this manuscript. We also appreciate the assistance of Muzna Mirza, MBBS, MSHI, Grace Garey, and Mary Lou Glazer in compiling the bibliography.

AUTHOR DISCLOSURES

The authors report the following conflicts of interest with the sponsor of this supplement article or products discussed in this article:

Eta S. Berner, EdD, has no financial arrangement or affiliation with a corporate organization or manufacturer of a product discussed in this article.

Mark L. Graber, MD, has no financial arrangement or affiliation with a corporate organization or manufacturer of a product discussed in this article.

References

1. Lowry F. Failure to perform autopsies means some MDs "walking in a fog of misplaced optimism." *CMAJ.* 1995;153:811–814.
2. Mongerson P. A patient's perspective of medical informatics. *J Am Med Inform Assoc.* 1995;2:79–84.
3. Blendon RJ, DesRoches CM, Brodie M, et al. Views of practicing physicians and the public on medical errors. *N Engl J Med.* 2002; 347:1933–1940.
4. YouGov survey of medical misdiagnosis. Isabel Healthcare–Clinical Decision Support System, 2005. Available at: http://www.isabel-healthcare.com. Accessed April 3, 2006.
5. Burroughs TE, Waterman AD, Gallagher TH, et al. Patient concerns about medical errors in emergency departments. *Acad Emerg Med.* 2005;23:57–64.
6. Tierney WM. Adverse outpatient drug events—a problem and an opportunity. *N Engl J Med.* 2003;348:1587–1589.
7. Norman GR, Coblentz CL, Brooks LR, Babcook CJ. Expertise in visual diagnosis: a review of the literature. *Acad Med.* 1992; 67(suppl):S78–S83.
8. Foucar E, Foucar MK. Medical error. In: Foucar MK, ed. *Bone Marrow Patholog,* 2nd ed. Chicago: ASCP Press, 2001:76–82.
9. Fitzgerald R. Error in radiology. *Clin Radiol.* 2001;56:938–946.
10. Kronz JD, Westra WH, Epstein JI. Mandatory second opinion surgical pathology at a large referral hospital. *Cancer.* 1999;86:2426–2435.
11. Berlin L, Hendrix RW. Perceptual errors and negligence. *Am J Radiol.* 1998;170:863–867.
12. Kripalani S, Williams MV, Rask K. Reducing errors in the interpretation of plain radiographs and computed tomography scans. In: Shojania KG, Duncan BW, McDonald KM, Wachter RM, eds. *Making Health Care Safer. A Critical Analysis of Patient Safety Practices*. Rockville, MD: Agency for Healthcare Research and Quality, 2001.
13. Neale G, Woloschynowych J, Vincent C. Exploring the causes of adverse events in NHS hospital practice. *J R Soc Med.* 2001;94:322–330.
14. O'Connor PM, Dowey KE, Bell PM, Irwin ST, Dearden CH. Unnecessary delays in accident and emergency departments: do medical and surgical senior house officers need to vet admissions? *Acad Emerg Med.* 1995;12:251–254.
15. Chellis M, Olson JE, Augustine J, Hamilton GC. Evaluation of missed diagnoses for patients admitted from the emergency department. *Acad Emerg Med.* 2001;8:125–130.
16. Elstein AS. Clinical reasoning in medicine. In: Higgs JJM, ed. *Clinical Reasoning in the Health Professions*. Oxford, England: Butterworth-Heinemann Ltd, 1995:49–59.
17. Kedar I, Ternullo JL, Weinrib CE, Kelleher KM, Brandling-Bennett H, Kvedar JC. Internet based consultations to transfer knowledge for patients requiring specialised care: retrospective case review. *BMJ.* 2003;326:696–699.
18. McGinnis KS, Lessin SR, Elder DE. Pathology review of cases presenting to a multidisciplinary pigmented lesion clinic. *Arch Dermatol.* 2002;138:617–621.
19. Zarbo RJ, Meier FA, Raab SS. Error detection in anatomic pathology. *Arch Pathol Lab Med.* 2005;129:1237–1245.
20. Tomaszewski JE, Bear HD, Connally JA, et al. Consensus conference on second opinions in diagnostic anatomic pathology: who, what, and when. *Am J Clin Pathol.* 2000;114:329–335.
21. Harris M, Hartley AL, Blair V, et al. Sarcomas in north west England. I. Histopathological peer review. *Br J Cancer.* 1991;64:315–320.
22. Kim J, Zelman RJ, Fox MA, et al. Pathology Panel for Lymphoma Clinical Studies: a comprehensive analysis of cases accumulated since its inception. *J Natl Cancer Inst.* 1982;68:43–67.
23. Goddard P, Leslie A, Jones A, Wakeley C, Kabala J. Error in radiology. *Br J Radiol.* 2001;74:949–951.
24. Berlin L. Defending the "missed" radiographic diagnosis. *Am J Radiol.* 2001;176:317–322.

25. Espinosa JA, Nolan TW. Reducing errors made by emergency physicians in interpreting radiographs: longitudinal study. *BMJ.* 2000; 320:737–740.
26. Arenson RL. The wet read. AHRQ [Agency for Heathcare Research and Quality] Web M&M, March 2006. Available at: http://webmm.ahrq.gov/printview.aspx?caseID=121. Accessed November 28, 2007.
27. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample. *Arch Intern Med.* 1996;156:209–213.
28. Majid AS, de Paredes ES, Doherty RD, Sharma NR, Salvador X. Missed breast carcinoma: pitfalls and pearls. *Radiographics.* 2003; 23:881–895.
29. Goodson WH III, Moore DH II. Causes of physician delay in the diagnosis of breast cancer. *Arch Intern Med.* 2002;162:1343–1348.
30. Smith-Bindman R, Chu PW, Miglioretti DL, et al. Comparison of screening mammography in the United States and the United Kingdom. *JAMA.* 2003;290:2129–2137.
31. Schiff GD, Kim S, Abrams R, et al. Diagnosing diagnosis errors: lessons from a multi-institutional collaborative project. In: *Advances in Patient Safety: From Research to Implementation*, vol 2. Rockville: MD Agency for Healthcare Research and Quality, February 2005. AHRQ Publication No. 050021. Available at: http://www.ahrq.gov/downloads/pub/advances/vol2/schiff.pdf./. Accessed December 3, 2007.
32. Shojania K, Burton E, McDonald K, et al. The autopsy as an outcome and performance measure: evidence report/technology assessment #58. Rockville, MD: Agency for Healthcare Research and Quality, October 2002. AHRQ Publication No. 03-E002.
33. Pidenda LA, Hathwar VS, Grand BJ. Clinical suspicion of fatal pulmonary embolism. *Chest.* 2001;120:791–795.
34. Lederle FA, Parenti CM, Chute EP. Ruptured abdominal aortic aneurysm: the internist as diagnostician. *Am J Med.* 1994;96:163–167.
35. von Kodolitsch Y, Schwartz AG, Nienaber CA. Clinical prediction of acute aortic dissection. *Arch Intern Med.* 2000;160:2977–2982.
36. Edlow JA. Diagnosis of subarachnoid hemorrhage. *Neurocrit Care.* 2005;2:99–109.
37. Burton EC, Troxclair DA, Newman WP III. Autopsy diagnoses of malignant neoplasms: how often are clinical diagnoses incorrect? *JAMA.* 1998;280:1245–1248.
38. Perlis RH. Misdiagnosis of bipolar disorder. *Am J Manag Care*. 2005;11(suppl):S271–S274.
39. Graff L, Russell J, Seashore J, et al. False-negative and false-positive errors in abdominal pain evaluation: failure to diagnose acute appendicitis and unneccessary surgery. *Acad Emerg Med.* 2000;7:1244–1255.
40. Raab SS, Grzybicki DM, Janosky JE, et al. Clinical impact and frequency of anatomic pathology errors in cancer diagnoses. *Cancer.* 2005;104:2205–2213.
41. Buchweitz O, Wulfing P, Malik E. Interobserver variability in the diagnosis of minimal and mild endometriosis. *Eur J Obstet Gynecol Reprod Biol.* 2005;122:213–217.
42. Gorter S, van der Heijde DM, van der Linden S, et al. Psoriatic arthritis: performance of rheumatologists in daily practice. *Ann Rheum Dis.* 2002;61:219–224.
43. Bogun F, Anh D, Kalahasty G, et al. Misdiagnosis of atrial fibrillation and its clinical consequences. *Am J Med.* 2004;117:636–642.
44. Arnon SS, Schecter R, Maslanka SE, Jewell NP, Hatheway CL. Human botulism immune globulin for the treatment of infant botulism. *N Engl J Med.* 2006;354:462–472.
45. Edelman D. Outpatient diagnostic errors: unrecognized hyperglycemia. *Eff Clin Pract*. 2002;5:11–16.
46. Russell NJ, Pantin CF, Emerson PA, Crichton NJ. The role of chest radiography in patients presenting with anterior chest pain to the Accident & Emergency Department. *J R Soc Med.* 1988;81:626–628.
47. Dobbs D. Buried answers. *New York Times Magazine.* April 24, 2005:40–45.
48. Cabot RC. Diagnostic pitfalls identified during a study of three thousand autopsies. *JAMA.* 1912;59:2295–2298.
49. Cabot RC. A study of mistaken diagnosis: based on the analysis of 1000 autopsies and a comparison with the clinical findings. *JAMA.* 1910;55:1343–1350.
50. Aalten CM, Samsom MM, Jansen PA. Diagnostic errors: the need to have autopsies. *Neth J Med.* 2006;64:186–190.
51. Shojania KG. Autopsy revelation. AHRQ [Agency for Heathcare Research and Quality] Web M&M, March 2004. Available at: http://webmm.ahrq.gov/case.aspx?caseID=54&searchStr=shojania. Accessed November 28, 2007.
52. Tamblyn RM. Use of standardized patients in the assessment of medical practice. *CMAJ.* 1998;158:205–207.
53. Berner ES, Houston TK, Ray MN, et al. Improving ambulatory prescribing safety with a handheld decision support system: a randomized controlled trial. *J Am Med Inform Assoc*. 2006;13:171–179.
54. Christensen-Szalinski JJ, Bushyhead JB. Physician's use of probabalistic information in a real clinical setting. *J Exp Psychol Hum Percept Perform.* 1981;7:928–935.
55. Peabody JW, Luck J, Jain S, Bertenthal D, Glassman P. Assessing the accuracy of administrative data in health information systems. *Med Care.* 2004;42:1066–1072.
56. Margo CE. A pilot study in ophthalmology of inter-rater reliability in classifying diagnostic errors: an underinvestigated area of medical error. *Qual Saf Health Care.* 2003;12:416–420.
57. Hoffman PJ, Slovic P, Rorer LG. An analysis-of-variance model for the assessment of configural cue utilization in clinical judgment. *Psychol Bull.* 1968;69:338–349.
58. Kohn L, Corrigan JM, Donaldson M. *To Err Is Human: Building a Safer Health System.* Washington, DC: National Academy Press, 1999.
59. Leape L, Brennan TA, Laird N, et al. The nature of adverse events in hospitalized patients: results of the Harvard Medical Practice Study II. *N Engl J Med.* 1991;324:377–384.
60. Thomas EJ, Studdert DM, Burstin HR, et al. Incidence and types of adverse events and negligent care in Utah and Colorado. *Med Care.* 2000;38:261–271.
61. Baker GR, Norton PG, Flintoft V, et al. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. *CMAJ.* 2004;170:1678–1686.
62. Wilson RM, Harrison BT, Gibberd RW, Hamilton JD. An analysis of the causes of adverse events from the Quality in Australian Health Care Study. *Med J Aust.* 1999;170:411–415.
63. Davis P, Lay-Yee R, Briant R, Ali W, Scott A, Schug S. Adverse events in New Zealand public hospitals II: preventability and clinical context. *N Z Med J.* 2003;116:U624.
64. Bhasale A, Miller G, Reid S, Britt HC. Analyzing potential harm in Australian general practice: an incident-monitoring study. *Med J Aust.* 1998;169:73–76.
65. Makeham M, Dovey S, County M, Kidd MR. An international taxonomy for errors in general practice: a pilot study. *Med J Aust.* 2002;177:68–72.
66. Fischer G, Fetters MD, Munro AP, Goldman EB. Adverse events in primary care identified from a risk-management database. *J Fam Pract.* 1997;45:40–46.
67. Wu AW, Folkman S, McPhee SJ, Lo B. Do house officers learn from their mistakes? *JAMA.* 1991;265:2089–2094.
68. Weingart S, Ship A, Aronson M. Confidential clinical-reported surveillance of adverse events among medical inpatients. *J Gen Intern Med.* 2000;15:470–477.
69. Balsamo RR, Brown MD. Risk management. In: Sanbar SS, Gibofsky A, Firestone MH, LeBlang TR, eds. *Legal Medicin*e, 4th ed. St Louis, MO: Mosby, 1998:223–244.
70. Failure to diagnose. Midiagnosis of conditions and diseaes. Medical Malpractice Lawyers and Attorneys Online, 2006. Available at: http://www.medical-malpractice-attorneys-lawsuits.com/pages/failure-to-diagnose.html. Accessed November 28, 2007.

71. *General and Family Practice Claim Summary.* Physician Insurers Association of America, Rockville, MD, 2002.
72. Berlin L. Fear of cancer. *AJR Am J Roentgenol.* 2004;183:267–272.
73. Missed or failed diagnosis: what the UNITED claims history can tell us. United GP Registrar's Toolkit, 2005. Available at: http://www.unitedmp.com.au/0/0.13/0.13.4/Missed_diagnosis.pdf. Accessed November 28, 2007.
74. Studdert DM, Mello MM, Gawande AA, et al. Claims, errors, and compensation payments in medical malpractice litigation. *N Engl J Med.* 2006;354:2024–2033.
75. Schiff GD. Commentary: diagnosis tracking and health reform. *Am J Med Qual.* 1994;9:149–152.
76. Redelmeier DA. Improving patient care: the cognitive psychology of missed diagnoses. *Ann Intern Med.* 2005;142:115–120.
77. Gandhi TK, Kachalia A, Thomas EJ, et al. Missed and delayed diagnoses in the ambulatory setting: a study of closed malpractice claims. *Ann Intern Med.* 2006;145:488–496.
78. Kirch W, Schafii C. Misdiagnosis at a university hospital in 4 medical eras. *Medicine (Baltimore).* 1996;75:29–40.
79. Goldman L, Sayson R, Robbins S, Cohn LH, Bettmann M, Weisberg M. The value of the autopsy in three different eras. *N Engl J Med.* 1983;308:1000–1005.
80. Shojania KG, Burton EC, McDonald KM, Goldman L. Changes in rates of autopsy-detected diagnostic errors over time: a systematic review. *JAMA.* 2003;289:2849–2856.
81. Sonderegger-Iseli K, Burger S, Muntwyler J, Salomon F. Diagnostic errors in three medical eras: a necropsy study. *Lancet.* 2000;355:2027–2031.
82. Berner ES, Miller RA, Graber ML. Missed and delayed diagnoses in the ambulatory setting. *Ann Intern Med.* 2007;146:470–471.
83. Gawande A. Final cut. Medical arrogance and the decline of the autopsy. *The New Yorker.* March 19, 2001:94–99.
84. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J Pers Soc Psychol.* 1999;77:1121–1134.
85. LaFee S. Well news: all the news that's fit. *The San Diego Union-Tribune.* March 7, 2006. Available at: http://www.quotegarden.com/medical.html. Accessed February 6, 2008.
86. Graber ML. Diagnostic error in medicine: a case of neglect. *Jt Comm J Qual Patient Saf.* 2005;31:112–119.
87. Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? *Ann Intern Med.* 1985;103:596–599.
88. Gorman PN, Helfand M. Information seeking in primary care: how physicians choose which clinical questions to pursue and which to leave unanswered. *Med Decis Making.* 1995;15:113–119.
89. Osheroff JA, Bankowitz RA. Physicians' use of computer software in answering clinical questions. *Bull Med Libr Assoc.* 1993;81:11–19.
90. Rosenbloom ST, Geissbuhler AJ, Dupont WD, et al. Effect of CPOE user interface design on user-initiated access to educational and patient information during clinical care. *J Am Med Inform Assoc.* 2005;12:458–473.
91. McGlynn EA, Asch SM, Adams J, et al. The quality of health care delivered to adults in the United States. *N Engl J Med.* 2003;348:2635–2645.
92. Cabana MD, Rand CS, Powe NR, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. *JAMA.* 1999;282:1458–1465.
93. Eccles MP, Grimshaw JM. Selecting, presenting and delivering clinical guidelines: are there any "magic bullets"? *Med J Aust.* 2004;180(suppl):S52–S54.
94. Pearson TA, Laurora I, Chu H, Kafonek S. The lipid treatment assessment project (L-TAP): a multicenter survey to evaluate the percentages of dyslipidemic patients receiving lipid-lowering therapy and achieving low-density lipoprotein cholesterol goals. *Arch Intern Med.* 2000;160:459–467.
95. Eccles M, McColl E, Steen N, et al. Effect of computerised evidence based guidelines on management of asthma and angina in adults in primary care: cluster randomised controlled trial [primary care]. *BMJ.* 2002;325:941.
96. Smith WR. Evidence for the effectiveness of techniques to change physician behavior. *Chest.* 2000;118:8S–17S.
97. Militello L, Patterson ES, Tripp-Reimer T, et al. Clinical reminders: why don't people use them? In: *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting*, New Orleans LA, 2004:1651–1655.
98. Patterson ES, Doebbeling BN, Fung CH, Militello L, Anders S, Asch SM. Identifying barriers to the effective use of clinical reminders: bootstrapping multiple methods. *J Biomed Inform.* 2005;38:189–199.
99. Berner ES, Maisiak RS, Heudebert GR, Young KR Jr. Clinician performance and prominence of diagnoses displayed by a clinical diagnostic decision support system. *AMIA Annu Symp Proc.* 2003;2003:76–80.
100. Steinman MA, Fischer MA, Shlipak MG, et al. Clinician awareness of adherence to hypertension guidelines. *Am J Med.* 2004;117:747–754.
101. Tierney WM, Overhage JM, Murray MD, et al. Can computer-generated evidence-based care suggestions enhance evidence-based management of asthma and chronic obstructive pulmonary disease? A randomized, controlled trial. *Health Serv Res.* 2005;40:477–497.
102. Timmermans S, Mauck A. The promises and pitfalls of evidence-based medicine. *Health Aff (Millwood).* 2005;24:18–28.
103. Tanenbaum SJ. Evidence and expertise: the challenge of the outcomes movement to medical professionalism. *Acad Med.* 1999;74:757–763.
104. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. *J Am Med Inform Assoc.* 2006;13:138–147.
105. Katz J. Why doctors don't disclose uncertainty. *Hastings Cent Rep.* 1984;14:35–44.
106. Graber ML, Franklin N, Gordon RR. Diagnostic error in internal medicine. *Arch Intern Med.* 2005;165:1493–1499.
107. Croskerry P. Achieving quality in clinical decision making: cognitive strategies and detection of bias. *Acad Emerg Med.* 2002;9:1184–1204.
108. Friedman CP, Gatti GG, Franz TM, et al. Do physicians know when their diagnoses are correct? *J Gen Intern Med.* 2005;20:334–339.
109. Dreiseitl S, Binder M. Do physicians value decision support? A look at the effect of decision support systems on physician opinion. *Artif Intell Med.* 2005;33:25–30.
110. Baumann AO, Deber RB, Thompson GG. Overconfidence among physicians and nurses: the 'micro-certainty, macro-uncertainty' phenomenon. *Soc Sci Med.* 1991;32:167–174.
111. Podbregar M, Voga G, Krivec B, Skale R, Pareznik R, Gabrscek L. Should we confirm our clinical diagnostic certainty by autopsies? *Intensive Care Med.* 2001;27:1750–1755.
112. Landefeld CS, Chren MM, Myers A, Geller R, Robbins S, Goldman L. Diagnostic yield of the autopsy in a university hospital and a community hospital. *N Engl J Med.* 1988;318:1249–1254.
113. Potchen EJ. Measuring observer performance in chest radiology: some experiences. *J Am Coll Radiol.* 2006;3:423–432.
114. Kachalia A, Gandhi TK, Puopolo AL, et al. Missed and delayed diagnoses in the emergency department: a study of closed malpractice claims from 4 liability insurers. *Ann Emerg Med.* 2007;49:196–205.
115. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad Med.* 2003;78:775–780.
116. Bornstein BH, Emler AC. Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *J Eval Clin Pract.* 2001;7:97–107.
117. McSherry D. Avoiding premature closure in sequential diagnosis. *Artif Intell Med.* 1997;10:269–283.
118. Dubeau CE, Voytovich AE, Rippey RM. Premature conclusions in the diagnosis of iron-deficiency anemia: cause and effect. *Med Decis Making.* 1986;6:169–173.
119. Voytovich AE, Rippey RM, Suffredini A. Premature conclusions in diagnostic reasoning. *J Med Educn.* 1985;60:302–307.

120. Simon HA. *The Sciences of the Artificial,* 3rd ed. Cambridge, MA: MIT Press, 1996.
121. Elstein AS, Shulman LS, Sprafka SA. *Medical Problem Solving. An Analysis of Clinical Reasoning.* Cambridge, MA: Harvard University Press, 1978.
122. Barrows HS, Norman GR, Neufeld VR, Feightner JW. The clinical reasoning of randomly selected physicians in general medical practice. *Clin Invest Med.* 1982;5:49–55.
123. Barrows HS, Feltovich PJ. The clinical reasoning process. *Med Educ.* 1987;21:86–91.
124. Neufeld VR, Norman GR, Feightner JW, Barrows HS. Clinical problem-solving by medical students: a cross-sectional and longitudinal analysis. *Med Educ.* 1981;15:315–322.
125. Norman GR. The epistemology of clinical reasoning: perspectives from philosophy, psychology, and neuroscience. *Acad Med.* 2000; 75(suppl):S127–S135.
126. Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise: theory and implications. *Acad Med.* 1990;65: 611–621.
127. Gladwell M. *Blink: The Power of Thinking Without Thinking*. Boston: Little Brown and Company, 2005.
128. Klein G. *Sources of Power: How People Make Decisions*. Cambridge, MA: MIT Press, 1998.
129. Rosch E, Mervis CB. Family resemblances: studies in the internal structure of categories. *Cognit Psychol.* 1975;7:573–605.
130. Eva KW, Norman GR. Heuristics and biases—a biased perspective on clinical reasoning. *Med Educ.* 2005;39:870–872.
131. Sieck WR, Arkes HR. The recalcitrance of overconfidence and its contribution to decision aid neglect. *J Behav Decis Making.* 2005; 18:29–53.
132. Kruger J, Dunning D. Unskilled and unaware—but why? A reply to Krueger and Mueller (2002). *J Pers Soc Psychol.* 2002;82:189–192.
133. Krueger J, Mueller RA. Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *J Pers Soc Psychol.* 2002;82:180–188.
134. Mele AR. Real self-deception. *Behav Brain Sci.* 1997;20:91–102.
135. Reason JT, Manstead ASR, Stradling SG. Errors and violation on the roads: a real distinction? *Ergonomics.* 1990;33:1315–1332.
136. Gigerenzer G, Goldstein DG. Reasoning the fast and frugal way: models of bounded rationality. *Psychol Rev.* 1996;103:650–669.
137. Hamm RM. Clinical intuition and clinical analysis: expertise and the cognitive continuum. In: Elstein A, Dowie J, eds. *Professional Judgment: A Reader in Clinical Decision Making*. Cambridge, UK: Cambridge University Press, 1988:78–105.
138. Gigerenzer G. *Adaptive Thinking.* New York: Oxford University Press, 2000.
139. Berner ES, Webster GD, Shugerman AA, et al. Performance of four computer-based diagnostic systems. *N Engl J Med.* 1994;330:1792–1796.
140. Ely JW, Levinson W, Elder NC, Mainous AG III, Vinson DC. Perceived causes of family physicians' errors. *J Fam Pract.* 1995; 40:337–344.
141. Studdert DM, Mello MM, Sage WM, et al. Defensive medicine among high-risk specialist physicians in a volatile malpractice environment. *JAMA.* 2005;293:2609–2617.
142. Anderson RE. Billions for defense: the pervasive nature of defensive medicine. *Arch Intern Med.* 1999;159:2399–2402.
143. Trivers R. The elements of a scientific theory of self-deception. *Ann N Y Acad Sci.* 2000;907:114–131.
144. Lundberg GD. Low-tech autopsies in the era of high-tech medicine: continued value for quality assurance and patient safety. *JAMA.* 1998;280:1273–1274.
145. Darwin C. *The Descent of Man.* Project Gutenberg, August 1, 2000. Available at: http://www.gutenberg.org/etext/2300. Accessed November 28, 2007.
146. Leonhardt D. Why doctors so often get it wrong. *The New York Times.* February 22, 2006 [published correction appears in *The New York Times*, February 28, 2006]. Available at: http://www.nytimes.com/2006/02/22/business/22leonhardt.html?ex=1298264400&en=c2d9f1d654850c17&ei=5088&partner=rssnyt&emc=rss. Accessed November 28, 2007.
147. Hodges B, Regehr G, Martin D. Difficulties in recognizing one's own incompetence: novice physicians who are unskilled and unaware of it. *Acad Med.* 2001;76(suppl):S87–S89.
148. Davis DA, Mazmanian PE, Fordis M, Van HR, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA.* 2006;296:1094–1102.
149. Davis D, O'Brien MA, Freemantle N, Wolf FM, Mazmanian P, Taylor-Vaisey A. Impact of formal continuing medical education: do conferences, workshops, rounds, and other traditional continuing education activities change physician behavior or health care outcomes? *JAMA.* 1999;282:867–874.
150. Bowen JL. Educational strategies to promote clinical diagnostic reasoning. *N Engl J Med.* 2006;355:2217–2225.
151. Norman G. Building on experience—the development of clinical reasoning. *N Engl J Med.* 2006;355:2251–2252.
152. Bordage G. Why did I miss the diagnosis? Some cognitive explanations and educational implications. *Acad Med.* 1999;74(suppl):S128–S143.
153. Norman G. Research in clinical reasoning: past history and current trends. *Med Educ.* 2005;39:418–427.
154. Singh H, Petersen LA, Thomas EJ. Understanding diagnostic errors in medicine: a lesson from aviation. *Qual Saf Health Care.* 2006;15: 159–164.
155. Croskerry P. When diagnoses fail: new insights, old thinking. *Canadian Journal of CME.* 2003;Nov:79–87.
156. Hall KH. Reviewing intuitive decision making and uncertainty: the implications for medical education. *Med Educ.* 2002;36:216–224.
157. Croskerry P. Cognitive forcing strategies in clinical decision making. *Ann Emerg Med.* 2003;41:110–120.
158. Mitchell DJ, Russo JE, Pennington N. Back to the future: temporal perspective in the explanation of events. *J Behav Decis Making.* 1989;2:25–38.
159. Schon DA. *Educating the Reflective Practitioner*. San Francisco: Jossey-Bass, 1987.
160. Mamede S, Schmidt HG. The structure of reflective practice in medicine. *Med Educ.* 2004;38:1302–1308.
161. Soares SMS. Reflective practice in medicine (PhD thesis). Erasmus Universiteit, Rotterdam, Rotterdam, the Netherlands, 2006. 30552B 6000.
162. Mamede S, Schmidt HG, Rikers R. Diagnostic errors and reflective practice in medicine. *J Eval Clin Pract.* 2007;13:138–145.
163. Reason J. Human error: models and management. *BMJ.* 2000;320: 768–770.
164. Nolan TW. System changes to improve patient safety. *BMJ.* 2000; 320:771–773.
165. Miller RA. Medical diagnostic decision support systems—past, present, and future: a threaded bibliography and brief commentary. *J Am Med Inform Assoc.* 1994;1:8–27.
166. Kassirer JP. A report card on computer-assisted diagnosis—the grade: C. *N Engl J Med.* 1994;330:1824–1825.
167. Berner ES, Maisiak RS. Influence of case and physician characteristics on perceptions of decision support systems. *J Am Med Inform Assoc.* 1999;6:428–434.
168. Friedman CP, Elstein AS, Wolf FM, et al. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. *JAMA.* 1999;282:1851–1856.
169. Lincoln MJ, Turner CW, Haug PJ, et al. Iliad's role in the generalization of learning across a medical domain. *Proc Annu Symp Comput Appl Med Care.* 1992; 174–178.
170. Lincoln MJ, Turner CW, Haug PJ, et al. Iliad training enhances medical students' diagnostic skills. *J Med Syst.* 1991;15:93–110.
171. Turner CW, Lincoln MJ, Haug P, et al. Iliad training effects: a cognitive model and empirical findings. *Proc Annu Symp Comput Appl Med Care.* 1991:68–72.

172. Arene I, Ahmed W, Fox M, Barr CE, Fisher K. Evaluation of quick medical reference (QMR) as a teaching tool. *MD Comput.* 1998;15: 323–326.
173. Apkon M, Mattera JA, Lin Z, et al. A randomized outpatient trial of a decision-support information technology tool. *Arch Intern Med.* 2005;165:2388–2394.
174. Ramnarayan P, Roberts GC, Coren M, et al. Assessment of the potential impact of a reminder system on the reduction of diagnostic errors: a quasi-experimental study. *BMC Med Inform Decis Mak.* 2006;6:22.
175. Maffei FA, Nazarian EB, Ramnarayan P, Thomas NJ, Rubenstein JS. Use of a web-based tool to enhance medical student learning in the pediatric ICU and inpatient wards. *Pediatr Crit Care Med.* 2005;6: 109.
176. Ramnarayan P, Tomlinson A, Kularni G, Rao A, Britto J. A novel diagnostic aid (ISABEL): development and preliminary evaluation of clinical performance. *Medinfo.* 2004;11:1091–1095.
177. Ramnarayan P, Kapoor RR, Coren J, et al. Measuring the impact of diagnostic decision support on the quality of clinical decision making: development of a reliable and valid composite score. *J Am Med Inform Assoc.* 2003;10:563–572.
178. Ramnarayan P, Tomlinson A, Rao A, Coren M, Winrow A, Britto J. ISABEL: a web-based differential diagnosis aid for paediatrics: results from an initial performance evaluation. *Arch Dis Child.* 2003; 88:408–413.
179. Miller RA. Evaluating evaluations of medical diagnostic systems. *J Am Med Inform Assoc.* 1996;3:429–431.
180. Berner ES. Diagnostic decision support systems: how to determine the gold standard? *J Am Med Inform Assoc.* 2003;10:608–610.
181. Bauer BA, Lee M, Bergstrom L, et al. Internal medicine resident satisfaction with a diagnostic decision support system (DXplain) introduced on a teaching hospital service. *Proc AMIA Symp.* 2002; 31–35.
182. Teich JM, Merchia PR, Schmiz JL, Kuperman GJ, Spurr CD, Bates DW. Effects of computerized physician order entry on prescribing practices. *Arch Intern Med.* 2000;160:2741–2747.
183. Croskerry P. The feedback sanction. *Acad Emerg Med.* 2000;7:1232–1238.
184. Jamtvedt G, Young JM, Kristoffersen DT, O'Brien MA, Oxman AD. Does telling people what they have been doing change what they do? A systematic review of the effects of audit and feedback. *Qual Saf Health Care.* 2006;15:433–436.
185. Papa FJ, Aldrich D, Schumacker RE. The effects of immediate online feedback upon diagnostic performance. *Acad Med.* 1999;74(suppl): S16–S18.
186. Stone ER, Opel RB. Training to improve calibration and discrimination: the effects of performance and environment feedback. *Organ Behav Hum Decis Process.* 2000;83:282–309.
187. Duffy FD, Holmboe ES. Self-assessment in lifelong learning and improving performance in practice: physician know thyself. *JAMA.* 2006;296:1137–1139.
188. Borgstede JP, Zinninger MD. Radiology and patient safety. *Acad Radiol.* 2004;11:322–332.
189. Frable WJ. "Litigation cells" in the Papanicolaou smear: extramural review of smears by "experts." *Arch Pathol Lab Med.* 1997;121:292–295.
190. Wilbur DC. False negatives in focused rescreening of Papanicolaou smears: how frequently are "abnormal" cells detected in retrospective review of smears preceding cancer or high grade intraepithelial neoplasia? *Arch Pathol Lab Med.* 1997;121:273–276.
191. College of American Pathologists. Available at: http://www.cap.org/apps/cap.portal.
192. Raab SS. Improving patient safety by examining pathology errors. *Clin Lab Med.* 2004;24:863.
193. Nieminen P, Kotaniemi L, Hakama M, et al. A randomised public-health trial on automation-assisted screening for cervical cancer in Finland: performance with 470,000 invitations. *Int J Cancer.* 2005; 115:307–311.
194. Nieminen P, Kotaniemi-Talonen L, Hakama M, et al. Randomized evaluation trial on automation-assisted screening for cervical cancer: results after 777,000 invitations. *J Med Screen.* 2007;14:23–28.
195. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med.* 2007;356:1399–1409.
196. Hall FM. Breast imaging and computer-aided detection. *N Engl J Med.* 2007;356:1464–1466.
197. Hill RB, Anderson RE. *Autopsy: Medical Practice and Public Policy*. Boston: Butterworth-Heinemann, 1988.
198. Cases and commentaries. AHRQ Web M&M, October 2007. Available at: http://www.webmm.ahrq.gov/index.aspx. Accessed December 12, 2007.
199. Fischer MA, Mazor KM, Baril J, Alper E, DeMarco D, Pugnaire M. Learning from mistakes: factors that influence how students and residents learn from medical errors. *J Gen Intern Med.* 2006;21:419–423.
200. Patriquin L, Kassarjian A, Barish M, et al. Postmortem whole-body magnetic resonance imaging as an adjunct to autopsy: preliminary clinical experience. *J Magn Reson Imaging*. 2001;13:277–287.
201. Umpire Information System (UIS). Available at: http://www.questec.com/q2001/prod_uis.htm. Accessed April 10, 2008.
202. Graber ML, Franklin N, Gordon R. Reducing diagnostic error in medicine: what's the goal? *Acad Med.* 2002;77:981–992.
203. Deyo RA. Cascade effects of medical technology. *Annu Rev Public Health.* 2002;23:23–44.
204. Dijksterhuis A, Bos MW, Nordgren LF, van Baaren RB. On making the right choice: the deliberation-without-attention effect. *Science.* 2006;311:1005–1007.
205. Redelmeier DA, Shafir E. Medical decision making in situations that offer multiple alternatives. *JAMA.* 1995;273:302–305.
206. Miller RA, Masarie FE Jr. The demise of the "Greek Oracle" model for medical diagnostic systems. *Methods Inf Med.* 1990;29:1–2.
207. Tsai TL, Fridsma DB, Gatti G. Computer decision support as a source of interpretation error: the case of electrocardiograms. *J Am Med Inform Assoc*. 2003;10:478–483.
208. Galletta DF, Durcikova A, Everard A, Jones BM. Does spell-checking software need a warning label? *Communications of the ACM.* 2005;48:82–86.
209. Tolstoy L. *Anna Karenina*. Project Gutenberg, July 1, 1998. Available at: http://www.gutenberg.org/etext/1399. Accessed December 11, 2007.
210. Committee on Quality of Health Care in America, Institute of Medicine Report. Washington, DC: The National Academy Press, 2001.
211. Gladwell M. Million-dollar Murray. *The New Yorker.* February 13, 2006:96–107.

The American Journal of Medicine (2008) Vol 121 (5A), S38–S42

ELSEVIER

THE AMERICAN JOURNAL *of* MEDICINE®

Minimizing Diagnostic Error: The Importance of Follow-up and Feedback

An open-loop system (also called a "nonfeedback controlled" system) is one that makes decisions based solely on preprogrammed criteria and the preexisting model of the system. This approach does not use feedback to calibrate its output or determine if the desired goal is achieved. Because open-loop systems do not observe the output of the processes they are controlling, they cannot engage in learning. They are unable to correct any errors they make or compensate for any disturbances to the process. A commonly cited example of the open-loop system is a lawn sprinkler that goes on automatically at a certain hour each day, regardless of whether it is raining or the grass is already flooded.[1]

To an unacceptably large extent, clinical diagnosis is an open-loop system. Typically, clinicians learn about their diagnostic successes or failures in various ad hoc ways (e.g., a knock on the door from a server with a malpractice subpoena; a medical resident learning, upon bumping into a surgical resident in the hospital hallway that a patient he/she cared for has been readmitted; a radiologist accidentally stumbling upon an earlier chest x-ray of a patient with lung cancer and noticing a nodule that had been overlooked). Physicians lack systematic methods for calibrating diagnostic decisions based on feedback from their outcomes. Worse yet, organizations have no way to learn about the thousands of collective diagnostic decisions that are made each day—information that could allow them to both improve overall performance as well as better hear the voices of the patients living with the outcomes.[2]

THE NEED FOR SYSTEMATIC FEEDBACK

In this commentary, I consider the issues raised in the review by Drs. Berner and Graber[3] and take the discussion further in contemplating the need for systematic feedback to improve diagnosis. Whereas their emphasis centers around the question of physician overconfidence regarding their own cognitive abilities and diagnostic decisions, I suspect many physicians feel more beleaguered and distracted than overconfident and complacent. There simply is not enough time in their rushed outpatient encounters, and too much "noise" in the nonspecified undifferentiated complaints that patients bring to them, for physicians, particularly primary care physicians, to feel overly secure. Both physicians and patients know this. Thus, we hear frequent complaints from both parties about brief appointments lacking sufficient time for full and proper evaluation. We also hear physicians' confessions about excessive numbers of tests being done, "overordered" as a way to compensate for these constraints that often are conflated with and complicated by "defensive medicine"—usually tests and consults ordered solely to block malpractice attorneys.

The issue is not so much that physicians lack an awareness of the thin ice on which they often are skating, but that they have no consistent and reliable systems for obtaining feedback on diagnosis. The reasons for this deficiency are multifactorial. **Table 1** lists some of the factors that mitigate against more systematic feedback on diagnosis outcomes and error. These items invite us to explicitly recognize this problem and design approaches that will make diagnosis more of a closed rather than open-loop system.

Given the current emphasis on heuristics, cognition, and unconscious biases that has been stimulated by publications such as Kassier and Kopelman's classic book *Learning Clinical Reasoning*,[4] and *How Doctors Think*,[5] the recent bestseller by Dr. Jerome Groopman, it is important to keep in mind that good medicine is less about brilliant diagnoses being made or missed and more about mundane mechanisms to ensure adequate follow-up.[6] Although this assertion remains an untested empirical question, I suspect that the proportion of malpractice cases related to diagnosis error—the leading cause of malpractice suits, outnumbering claims from medication errors by a factor of 2:1—that concern failure to *consider* a particular diagnosis is less than imagined.[7,8] Despite popular imagery of a diagnosis being missed by a dozen previous physicians only to be eventually made correctly by a virtuoso thinker (such as that stimulated by the Groopman book and dramatic cases reported in the

Statement of Author Disclosure: Please see the Author Disclosures section at the end of this article.

Requests for reprints should be addressed to: Gordon D. Schiff, MD, Division of General Medicine, Brigham and Women's Hospital, 1620 Tremont, 3rd Floor, Boston, Massachusetts 02120.

E-mail address: gschiff@partners.org.

0002-9343/$ -see front matter
doi:10.1016/j.amjmed.2008.02.004

Table 1 Barriers to feedback and follow-up

- Physician lack of time and systematic approaches for obtaining follow-up
  - —Unrealistic to expect MDs to rely on memory or ad hoc methods
- Clinical practice often doesn't require a diagnosis to treat
  - —Blunts MDs interest in feedback/follow-up
  - —Legitimately seen as purely academic question
  - —Suggests it is not worth time for follow-up
- High frequency of symptoms for which no definite diagnosis is ever established
  - —Self-limited nature of many symptoms/diagnoses
  - —Nonspecific symptoms for which no "organic" etiology ever identified
- Threatening nature of critical feedback makes MDs defensive
  - —MDs pride themselves on being "good diagnosticians"
  - —Reluctance of colleagues to "criticize" peers and be critiqued by them
- Fragmentation and discontinuities of care
  - —Ultimate diagnoses are often made later, in different setting
  - —Patient seen in other ERs, by specialists, admitted to different hospital
  - —No organized system for feedback of findings across institutions
- Reliance on patient return for follow-up; fragile link
  - —Patients busy; inconvenient to return
  - —Cost barriers
    - ○ Out-of-pocket costs from first visit can inhibit return
    - ○ Perceived lack of "value" for return visit
      - —If improved, seems pointless
      - —If not improved, may also seem not worthwhile
  - —Patient satisfaction and convenience
    - ○ If not improved, disgruntled patient may seek care elsewhere
- Managed care barriers discourage access
  - —Prior approval often required for repeat visit
- "Information breakage" despite return to original setting/MD
  - —Original record or question(s) may be inaccessible or forgotten
  - —May see partner of MD or other member of team

ER = emergency room; MD = medical doctor.

press), I believe such cases are less common than those involving failure to definitively establish a diagnosis that was considered by one or more physicians earlier. Obvious examples include the case of a patient with chest pain being sent home from the emergency room (ER) with a missed myocardial infarction (MI) or that involving oversight of a subtle abnormality on mammogram. Every ER physician in the emergency considers MI in chest-pain patients, and why else is a mammogram performed other than for consideration of breast cancer?

EXPANDED PARADIGMS IN DIAGNOSIS

The true concern in routine clinical diagnosis is not whether unsuspected new diagnoses are made or missed as much as it is the complexities of weighing and pursuing diagnostic considerations that are either obvious, may have been previously considered, or simply represent "dropped balls" (e.g., failed follow-up on an abnormal test result).[9] Furthermore, other paradigms often turn out to be more important than simply affixing a label on a patient naming a specific diagnosis (**Table 2**). Central to each of these "expanded paradigms" is the role for follow-up: deciding when a patient is acutely ill and required hospitalization, versus relatively stable but in need of careful observation, watching for complications or response after a diagnosis is made and a treatment started, monitoring for future recurrences, or even simply revising the diagnosis as the syndrome evolves. It often is more important for an ER or primary care physician to accurately decide whether a patient is "sick" and needs to be hospitalized or sent home than it is to come up with the precisely correct diagnosis at that moment of first encounter.

RESPONSE OVER TIME: THE ULTIMATE TEST?

Although the traditional "test of time" is frequently invoked, it is rarely applied in a standardized or evidence-based fashion, and never in a way that involves systematic tracking and calculating of accuracy rates or formal use of data that evolves over time for recalibration. One key unanswered question is, To what extent can we judge the accuracy of diagnoses based on how patients do over time or respond to treatment? In other words, if a patient gets better and responds to recommended therapy, can we assume the treatment, and hence the diagnosis, was correct? Basing diagnosis accuracy and learning on capturing feedback on whether or not a patient successfully "responds" to treatment is fraught with nuances and complexities that are rarely explicitly considered or measured. A partial list of such complexities is shown in **Table 3**.

Despite these limitations, feedback on patient response is critical for knowing not just how the patient is doing but how we as clinicians are doing. Particularly if we are mindful of these pitfalls, and especially if we can build in rigor with quantitative data to better answer the above questions, feedback on response seems imperative to learning from and improving diagnosis.

VIEWING DIAGNOSIS AS A RELATIONSHIP RATHER THAN A LABEL

Feedback on how patients are doing embodies an important corollary to the entire paradigm of diagnosis tracking and feedback. To a certain extent, diagnosis has been "reified," i.e., taken as an abstraction—an artificially constructed label—and misconceived as a "fact of nature."[10,11] By turning complex dynamic relationships between patients and their social environments, and even relationships between physicians and their patients, into "things" that boil down to neat categories, we risk oversimplifying complicated interactions of factors that are, in practice, larger than an International Classification of Diseases, 9th Revision (ICD-9) or

Table 2 Limitations of using successful or failed "treatment response" as an indicator for diagnostic error

- Diagnosis of severity/acuity
  —Failure to recognize patient need to be hospitalized or sent to ICU
- Diagnosis of complication
  —Assessing sequelae of a disease, drug, or surgery
- Diagnosis of a recurrence
  —What follow-up surveillance is required and how to interpret results
- Diagnosis of cure or failure to respond
  —When can clinician feel secure vs worry if symptoms don't improve
  —When should "test-of-cure" be done routinely
- Diagnosis of a misdiagnosis
  —When should a previous diagnosis be questioned and revised

ICU = intensive care unit.

Table 3 Factors complicating assessment of treatment response

- Patients who respond to a nonspecific/nonselective drug (e.g., corticosteroids) despite a wrong diagnosis
- Patients who fail to respond to therapy despite the correct diagnosis
- Varying time intervals for expected response
  — When does a clinician decide a patient is/is not responding
- Interpretation of partial responses
- How to incorporate known variations in response
  —Timing
  —Degree
- Role of surrogate (e.g., lab test or x-ray improvement) vs actual clinical outcome
- Timing of repeat testing to check for patient response
  —When and how often to repeat an x-ray or blood test
- Role of mitigating factors
  —Self-limited illnesses
  —Placebo response
  —Naturally relapsing and remitting courses of disorders

Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV) label.[12]

Building dialogue into the clinical diagnostic process, whereby the patient tells the practitioner how he/she is doing, represents an important premise. At the most basic level, doing so demonstrates a degree of caring that extends the clinical encounter beyond the rushed 15-minute exam. It is impossible to exaggerate the amazement and appreciation of my patients when I call to ask how they are doing a day or a week after an appointment to follow up on a clinical problem (as opposed to them calling me to complain that they are not improving!). Such follow-up means acknowledging that patients are coproducers in diagnosis—that they have an extremely important role to play to ensure that our diagnoses are as accurate as possible.[13]

The concept of coproduction of diagnosis goes beyond patients going home and "googling" the diagnosis the physician has suggested in order to decide whether their symptoms are consistent with what they read on the Internet, although there is certainly a role for such searches. It also is about much more than patients obtaining a second opinion from a second physician to enhance and ensure the accuracy of the diagnosis they were given (although this also is happening all the time, and we lack good ways to learn from such error-checking activities). What coproduction of diagnosis really should mean is that the patient is a partner in thinking through and testing the diagnostic hypothesis and has various important roles to play, some of which are described below.

Confirming or refuting a diagnostic hypothesis based on temporal relationships. "Doc, I know you think this rash is from that drug, but I checked and the rash started a week before I began the medication," or "The fever started before I even went to Guatemala."

Noting relieving or exacerbating factors that otherwise might not have been considered. "I later noticed that every time I leaned forward it made my chest pain better." This is a possible clue for pericarditis.

Carefully assessing the response to treatment. "The medication seemed to help at first, but is no longer helping." This suggests that the diagnosis or treatment may be incorrect (see Table 3).

Feeding back the nuances of the comments of a specialist referral. "The cardiologist you sent me to didn't think the chest pain was related to the mitral valve problem but she wasn't sure."

Triggering other past historical clues. "After I went home and thought about it, I remembered that as a teenager I once had an injury to my left side and peed blood for a week," states a patient with an otherwise inexplicable nonfunctioning left kidney. "I remembered that I once did work in a factory that made batteries," offers a patient with a elevated lead level.

Should I, as the physician of each of the actual patients cited above, have "taken a better history" and uncovered each of these pieces of data myself on the initial visit? Each emerged only through subsequent follow-up. Shouldn't I have asked more detailed probing questions during my first encounter with the patient? Shouldn't I have asked follow-up questions during the initial encounter that more actively explored my differential diagnosis based on (what ideally should be) my extensive knowledge of various diseases? Realistically, this will never happen.

Hit-and-miss medicine needs to be replaced by pull systems, which are described by Najarian[14] as "going forward by moving backward." Communication fed back from downstream outcomes, like Japanese kanban cards, should reliably pull the physician back to the patient to adjust

his/her management as well as continuously redesign methods for approaching future patients.

AVOIDANCE OF TAMPERING

Carefully refined signals from downstream feedback represent an important antidote to a well-known cognitive bias, *anchoring*, i.e., fixing on a particular diagnosis despite cues and clues that such persistence is unwarranted. However, feedback can exacerbate another bias—*availability bias*,[15] i.e., overreacting to a recent or vividly recalled event. For example, upon learning that a patient with a headache that was initially dismissed as benign was found to have a brain tumor, the physician works up all subsequent headache patients with imaging studies, even those with trivial histories. Thus, potentially useful feedback on the patient with a missed brain tumor is given undue weight, thereby biasing future decisions and failing to properly account for the rarity of neoplasms as a cause of a mild or acute headache.

When the quality guru Dr. W. Edwards Deming came into a factory, one of the first ways he improved quality was to stop the well-intentioned workers from "tampering," i.e., fiddling with the "dials."[16] For example, at the Wausau Paper company, the variations in paper size decreased by simply halting repeated adjustments of the sizing dials, which Deming showed often represented chasing random variation. As he dramatically showed with his classic funnel experiment, in which subjects dropped marbles through a funnel over a bull's-eye target, the more the subject attempted to adjust the position to compensate for each drop (e.g., moving to the right when a marble fell to the left of the target), the more variation was introduced, resulting in fewer marbles hitting the target than if the funnel were held in a consistent position. By overreacting to this random variation each time the target was missed, the subjects worsened rather than improved their accuracy and thereby were even less likely to hit the target.

If each time a physician's discovery that his/her diagnostic assessment erred on the side of a making a common diagnosis (thus missing a rare disorder) led to overreactions regarding future patients, or conversely, if each time the physician learned of a fruitless negative workup for a rare diagnosis, he/she vowed never to order so many tests, our cherished continuous feedback loops merely could be adding to variations and exacerbating poor quality in diagnosis. Or to paraphrase the language of Berner and Graber[3] or Rudolph,[17] feedback that inappropriately leads to either shaking or bolstering the physician's confidence in future diagnostic decision making is perhaps doing more harm than good. The continuous quality improvement (CQI) notion of avoiding tampering can be seen as the counterpart to the cognitive availability bias. It suggests a critical need to develop methods to properly weigh feedback in order to better calibrate diagnostic decision making. Although some of the so-called "statistical process control" (SPC) rules can be adapted to ensure more quantitative rigor to recalibrating decisions, generally, physicians are unfamiliar with these techniques. Thus, developing easy ways to incorporate, weigh, and simplify feedback data needs to be a priority.

CONCLUSION

Learning and feedback are inseparable. The old tools—ad hoc fortuitous feedback, individual idiosyncratic systems to track patients, reliance on human memory, and patient adherence to or initiating of follow-up appointments—are too unreliable to be depended upon to ensure high quality in modern diagnosis. Individual efforts to become wiser from cumulative clinical experience, an uphill battle at best, lack the power to provide the intelligence needed to inform learning organizations. What is needed instead is a systematic approach, one that fully involves patients and possesses an infrastructure this is hard wired to capture and learn from patient outcomes. Nothing less than such a linking of disease natural history to learning organizations poised to hear and learn from patient experiences and physician practices will suffice.

Gordon D. Schiff, MD
Division of General Medicine
Brigham and Women's Hospital
Boston, Massachusetts, USA

AUTHOR DISCLOSURES

The author reports the following conflicts of interest with the sponsor of this supplement article or products discussed in this article:

Gordon D. Schiff, MD, has no financial arrangement or affiliation with a corporate organization or a manufacturer of a product discussed in this article.

References

1. Open-loop controller. Available at: http://www.en.wikipedia.org/wiki/Open-loop_controller. Accessed January 23, 2008.
2. Schiff GD, Kim S, Abrams R, et al. Diagnosing diagnostic errors: lessons from a multi-institutional collaborative project. In: *Advances in Patient Safety: From Research to Implementation,* vol 2. Rockville, MD: Agency for Healthcare Research & Quality [AHRQ], February 2005. AHRQ Publication No. 050021. Available at: http://www.ahrq.gov/qual/advnaces/. Accessed December 3, 2007.
3. Berner E, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *Am J Med.* 2008;121(suppl 5A):S2–S23.
4. Kassirer JP, Kopelman RI. *Learning Clinical Reasoning.* Baltimore, MD: Lippincott Williams & Wilkins, 1991.
5. Groopman J. *How Doctors Think.* New York: Houghton Mifflin, 2007.
6. Schiff GD. Commentary: diagnosis tracking and health reform. *Am J Med Qual.* 1994;9:149–152.
7. Phillips R, Bartholomew L, Dovey S, Fryer GE Jr, Miyoshi TJ, Green LA. Learning from malpractice claims about negligent, adverse events in primary care in the United States. *Qual Saf Health Care.* 2004;13:121–126.
8. Gandhi TK, Kachalia A, Thomas EJ, et al. Missed and delayed diagnoses in the ambulatory setting: a study of closed malpractice claims. *Ann Intern Med.* 2006;145:488–496.
9. Gandhi TK. Fumbled handoffs: one dropped ball after another. *Ann Intern Med.* 2005;142:352–358.
10. Gould SJ. *The Mismeasure of Man.* New York: Norton & Co, 1981.
11. Freeman A. Diagnosis as explanation. *Early Child Dev Care.* 1989; 44:61–72.

12. Mellsop G, Kumar S. Classification and diagnosis in psychiatry: the emperor's clothes provide illusory court comfort. *Psychiatry Psychol Law.* 2007;14:95–99.
13. Hart JT. *The Political Economy of Health Care. A Clinical Perspective.* Bristol, United Kingdom: The Policy Press, 2006.
14. Najarian G. The pull system mystery explained: drum, buffer and rope with a computer. The Manager.org. Available at: http://www.themanager.org/strategy/pull_system.htm. Accessed January 24, 2008.
15. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. *Science.* 1974;185:1124–1130.
16. Deming WE. *Out of the Crisis.* Cambridge, MA: MIT Press, 1982.
17. Rudolph JW. Confidence, error, and ingenuity in diagnostic problem solving: clarifying the role of exploration and exploitation. Presented at: Annual Meeting of the Healthcare Management Division of the Academy of Management. August 5–8, 2007; Philadelphia, PA.

BARRY C. SCHECK

Barry C. Scheck is a Professor of Law at Benjamin N. Cardozo School of Law and Co-Director of the Innocence Project, an independent non-profit entity affiliated with the law school. Professor Scheck earned a B.S. degree from Yale University and J.D. and M.C.P. degrees from the University of California at Berkeley.

For twenty years he has served as a Commissioner on New York's Forensic Science Commission, which oversees all the state's crime laboratories. He is a Past President of the National Association of Criminal Defense Lawyers and last year, the National Law Journal recognized him as one of the 100 most influential lawyers in America over the last decade. He was recently selected by the National Institute of Science and Technology (NIST) to be a member of its Legal Resources Committee, which will provide guidance throughout the Organization of Scientific Advisory Committees (OSAC) about the legal ramifications of forensic standards under development and provide input on the presentation of forensic science results to the legal system.

He is also a partner in the civil rights law firm of Neufeld, Scheck, & Brustin, and has represented clients in many noteworthy and highly publicized criminal and civil cases, including cases involving re-investigations and exonerations of individuals wrongfully convicted of serious crimes. Although perhaps remembered most for his defense of O.J. Simpson, he was also the lead defense attorney in the Massachusetts case of *State v. Elizabeth Woodward*, which is now universally known as “The Nanny Case.” A large part of Prof. Scheck’s presentation today will address the cognitive bias issues that can “infect” and “contaminate” the decision-making on the part of criminal investigators, medical personnel, and forensic pathologists tasked with determining the cause of death of an infant while in the custody of a parent or caretaker.

STATE OF NEW YORK
COUNTY COURT COUNTY OF MONROE

THE PEOPLE OF THE STATE OF NEW YORK **DECISION AND ORDER**

-vs- Ind. No.: 2001-0490

RENE BAILEY a/k/a RENEE BAILEY,

Defendant.

Appearances:

For the People: **SANDRA DOORLEY, ESQ., DISTRICT ATTORNEY**
Matthew Dunham, Esq., Assistant District Attorney
Andra Ackerman, Esq., Assistant District Attorney
47 South Fitzhugh Street
Rochester, New York 14614

For the Defendant: **ADELE BERNHARD, ESQ.**
Adjunct Professor and Supervising Attorney
New York Law School
Post-Conviction Innocence Clinic
New York Law School Legal Services, Inc.
185 West Broadway, S-928
New York, New York 10013

PIAMPIANO, J.

The Defendant, having been convicted upon a jury verdict of Murder in the Second Degree (Penal Law § 125.25 [4]), moved this Court for an order, pursuant to Criminal Procedure Law § 440.10 (1) (g) and (1) (h), vacating the judgment of conviction and sentence or, in the alternative, a hearing on the matter. The Defense request was premised, in large part, on the assertion that the Defendant was convicted on the basis of

uncorroborated evidence that is now widely disputed in the medical community. The Defense claimed that new medical and scientific research, relative to the existence and characteristics of Shaken Baby Syndrome, has undermined the reliability of the verdict.

In addition to new medical and scientific evidence, the Defense claimed the existence of new exculpatory evidence from a daycare provider about statements made by a child witness, who was interviewed by the police, but did not testify at trial. The Defense also asserted an ineffective assistance of counsel claim.

The People opposed the relief sought in the Defendant's application, on the grounds that additional medical and expert witness testimony about Shaken Baby Syndrome is not "new evidence" pursuant to CPL § 440.10 (1) (g); that the proposed, newly discovered evidence, some of which was available prior to the Defendant's trial, is cumulative; and that it is not probable that the admission of such evidence at a subsequent trial would result in an acquittal. The People further asserted that certain evidence which the Defense would offer at a subsequent trial constitutes inadmissible hearsay; that the Defendant did not act with due diligence in bringing her claims of newly discovered evidence; and that the Defendant did not establish the ineffective assistance of counsel.

Upon consideration of the parties' respective submissions and oral arguments, the Court granted the Defense request for a hearing with respect to, *inter alia*, the limited issues of whether the proffered expert witness testimony concerning head injuries in children, and whether the proffered testimony concerning Sandra Hennessy's observations

of Cameron Burnside's behavior, constitute"new evidence" as that term is contemplated by Criminal Procedure Law § 440.10 (1) (g).

The hearing commenced on April 17, 2014 and spanned three weeks, during which time both parties presented the testimony of numerous witnesses and offered a multitude of exhibits in support of their respective positions. Upon the close of proofs, the Court directed each party to submit proposed Findings of Fact and Conclusions of Law. The Court received written submissions on behalf of the respective parties.

Now, upon consideration of the credible evidence adduced at the hearing of this matter, the Court hereby makes the following Findings of Fact.

FINDINGS OF FACT

THE DEFENDANT'S TRIAL
December 2001

On the morning of June 6, 2001, two and a half year old Brittney Sheets was left in the care of the Defendant, who operated a daycare business at her home. Prosecution witnesses testified that Brittney did not exhibit any signs of injury prior to being dropped off at approximately 8:30 a.m. that day. At approximately 3:15 p.m., Brittney's father, David Sheets, received a telephone call from the Defendant, who said that Brittney had fallen off of a bench and bumped her head. The Defendant further advised Mr. Sheets that he needed to get to the daycare quickly. Mr. Sheets responded, and found Brittney to be unresponsive. The Defendant told him that, while she was in the bathroom, Brittney had fallen from a chair in the playroom.

Brittney's parents took her to the office of her pediatrician, Jack Finnell, M.D. Dr. Finnell called for an ambulance, and Brittney was taken to Strong Memorial Hospital. Although she received treatment in the pediatric intensive care unit, Brittney was pronounced dead the following day.

At the Trial, Dr. Finnell testified, as follows:

Q. Doctor, are you familiar with the term Shaken Baby Syndrome or Shaken Baby Impact Syndrome?

A. Yes.

Q. How are you familiar with that?

A. Throughout medical school, residency, reading about it in journals, experiencing a couple cases of it while in residency, mainly seeing those kids in the Intensive Care Unit after the fact.

Q. Doctor, based on your training and experience, did the injuries that you suspected that were ultimately borne out at the hospital, were those injuries consistent with a fall from a chair on to a carpeted floor?

A. No.

Q. Why not?

A. Again, I hark back to someone, one of the attendings when I was in medical school as well as reading it in different textbooks and different journals that it is rare and, in fact, never has been seen to have a child fall from less than ten feet or approximately a second story window result in a serious brain injury.

Q. And, Doctor, based on your training and experience, do you have an opinion as to whether or not the injuries that Brittney suffered were consistent with a shaking or a shaking impact?

A. I do.

Q. What is that?

A. My opinion is based on the fact that there was no external signs of trauma; based on what I know of the Medical Examiner's report that these injuries could not have been suffered any other way than a Shaken Child Syndrome.

At the trial, Frank Maffei, M.D., a pediatric intensive care physician at Strong Memorial Hospital who treated Brittney on June 7, 2001, testified that, "I believe this child suffered non-accidental brain injury and I believe the mechanism was from violent shaking." Dr. Maffei based his opinion on a "constellation" of findings that included the consideration of a "history." Dr. Maffei testified that he had noted in his medical chart that, in addition to shaking, there may have been an impact. As to the possibility that Brittney's injuries could have been caused by a fall, Dr. Maffei testified that the forces occurring in a fall "usually are not" or are ,"rarely, if ever" life threatening.

Upon examination of Brittney's eyes at approximately 7:30 a.m. on June 7, 2001, Dr. Maffei observed diffuse retinal hemorrhages in both eyes, with multiple areas of bleeding. Dr. Maffei further testified that an ophthalmologist later concurred with those findings. On cross-examination, Dr. Maffei testified that impact occurs along with shaking in the majority of cases, and that shaking with impact, generates greater forces than shaking alone. Dr. Maffei acknowledged that he was familiar with a study conducted by Dr. John Plunkett, published in 2001, in which Dr. Plunkett concluded that short falls can be fatal to children.

At the Trial, Ana Rubio, M.D., testified on behalf of the Prosecution regarding the autopsy that she had conducted on Brittney. Dr. Rubio noted bruises on Brittney's throat and abdomen, and the inside of Brittney's scalp. She "could see only one sign of external trauma in the back of the head on the right side and a little contusion of the cerebellum underneath that area . . . that would be clinically unsignificant [*sic*]." Dr. Rubio testified that Brittney had suffered a subarachnoid hemorrhage, as well as a subdural hemorrhage in the back and middle of her head. There was a contusion of the back portion of the brain itself, and there was blood in the space between the dura matter and the eye. Further examination revealed retinal hemorrhages in both eyes.

Dr. Rubio testified that Brittney had suffered the kind of massive trauma that one might see as a result of being in a car accident. She described Brittney's injuries as resulting from "Shaken Impact Baby Syndrome" and testified that the acceleration and deceleration forces in shaking a child will be greater if the child's head "is suddenly stopped by impact against a surface." Dr. Rubio explained Shaken Impact Baby Syndrome as, "a constellation of findings, pathologic findings that they are produced in the setting of a small child being shaken and shaken meaning not only shaking the baby, but maybe shaken plus sudden deceleration when the body is put against a soft object." Dr. Rubio testified that the cause of death in this case was, "multiple injuries to the central nervous system produced by rotational forces," and that, "the most likely explanation" was that Brittney died as a result of "shaking and impact."

When asked to state the constellation of injuries that she would need to find in order to make a diagnosis involving the Shaken Impact Baby Syndrome, Dr. Rubio testified:

> In the medical literature the three findings that are usually present are subdural hemorrhage which is the blood underneath the dura or subarachnoid hemorrhage or both, a regular [*sic* retinal] hemorrhage, and cerebral edema which is swelling of the brain.

Dr. Rubio found all of those injuries in Brittney's case, along with extensive hemorrhage in the nerves around the spinal cord.

Dr. Rubio further testified that, before making a diagnosis of Shaken Impact Baby Syndrome, she considered the history regarding the cause of the injuries including, in this case, the allegation that Brittney jumped or fell from a chair that was about 18 inches high. It was Dr. Rubio's opinion that Brittney's injuries were inconsistent with being caused by such a fall. On cross-examination, when Dr. Rubio was asked whether it was possible for Brittney to have sustained the injuries as a result of falling from a chair, Dr. Rubio indicated that it was "extremely unlikely."

The Defense called Robert Greendyke, M.D., as an expert witness at the Trial. Dr. Greendyke testified that, in preparation for his testimony, he reviewed Dr. Finnell's reports, reports of Brittney's visit to the emergency room, ambulance records, and the Medical Examiner's report. He also examined microscopic tissue sections that were prepared in conjunction with the autopsy by the Medical Examiner's Office, and he reviewed reports pertaining to CT and MRI scans performed on Brittney.

Dr. Greendyke testified that at least some of the blood observed on Brittney's brain during the autopsy was the result of post-mortem or peri-mortem bleeding, and that at least some of the retinal hemorrhaging also occurred post-mortem, or while Brittney was at the hospital. Dr. Greendyke further testified that some of the blood pigment on the brain was caused by a previous fall that Brittney allegedly had sustained in December of 2000, when she was in the Defendant's care.

According to Dr. Greendyke's testimony, there was an absence of axonal injuries, which indicated that Brittney's injuries were not the result of Shaken Baby Syndrome. Based upon the bleeding and bruising on the surface of the brain, Dr. Greendyke concluded that Brittney's death resulted from a "violent impact" to her head. He noted that there was "a bruise on the edge of the cerebellum which is a portion of the brain in the back lower central portion of the head," which was "evidence of the head while in motion having struck something," and that Brittney's injuries were consistent with a fall. On cross-examination, Dr. Greendyke testified that Brittney's injuries could have been the result of her falling from a height of 18 inches onto a carpeted floor, "without question."

THE POST-CONVICTION HEARING
April 2014

At the hearing, the Defense case commenced with the testimony of Peter Stephens, M.D. Based on the credible evidence adduced at the hearing, the Court finds Dr. Stephens to be an expert in the area of pathology. In that regard, the Court credits Dr. Stephens' testimony that the recognition of the danger of falling has changed since 2001. Dr. Stephens explained that, over the last ten years, there has been a progressive change in the

attitude toward pediatric head trauma in at least three areas. First, there now is general agreement that short distance falls can cause death. Second, since 2001, retinal hemorrhages have been shown to result from increased pressure inside of the skull, rather than any type of rotational injury. Third, there is a discrepancy in the classical shaken baby theory, with respect to the traditional thinking that shaking disrupted the bridging veins on the surface of the brain.

Based on the credible evidence established at the hearing, the Court finds Kenneth Monson, Ph.D., to be an expert in biomechanical engineering with respect to his testimony. In that regard, the Court credits Dr. Monson's testimony that the biomedical research and literature has developed significantly since 2001, on the issue of whether shaking, and not short falls, is likely to be the mechanism for the type of injury at issue. That is, shaking a child hard enough to cause brain injury also would cause neck injury, yet none was observed in this case. Further, even falls of just a few feet generate levels of force and velocity that exceed known thresholds for brain injury, which is far more force than an adult human can generate by shaking.

Based on Dr. Monson's knowledge, none of the modeling attempts made since 2001 were able to establish that the violent shaking of an infant or a toddler could cause the kind of subdural hematomas, retinal hemorrhages, brain injury, and death that were associated with this case. Rather, every biomedical investigation that has been performed continues to suggest that the accelerations associated with shaking are lower than what

would be expected as necessary to cause those injuries. Significantly, nothing before 2001 would contradict that finding.

Based on the credible evidence established at the hearing, the Court finds John Plunkett, M.D., to be an expert in the area of general and forensic pathology with respect to his testimony. Dr. Plunkett has extensively studied the dangers of short falls to children. A research paper published by Dr. Plunkett in 2001challenged the then-existing perception that short falls or low velocity impacts could not cause death, by proving that "it was wrong." The study documented cases in which children had died from falls, and it was discussed by expert witnesses at the Trial.

The Court further credits the testimony of Dr. Plunkett that the triad of subdural hematoma (subdural hemorrhage), retinal hemorrhage, and cerebral edema (swelling of the brain) was viewed as generally pathognomonic (i.e., distinctively characteristic), of Shaken Baby Syndrome prior to the time frame of 2001-2002. Similar testimony was given by several additional Defense witnesses: Peter Stephens. M.D., Patrick Lantz, M.D., and Patrick Barnes, M.D.

Dr. Plunkett described Brittney's injuries as including a small volume acute subdural hematoma, malignant rapid brain swelling, contusion at the base of her left temporal lobe, and brain herniation. Dr. Plunkett testified that the combination of the swelling and the herniation caused Brittney's death. Based on the injuries and the history provided, Dr. Plunkett concluded that the bruising on the back of Brittney's head was evidence of an impact injury. He determined that Brittney's head was in motion and

struck a solid object, and that her injuries were consistent with the alleged falling or jumping from an 18 inch high chair, and hitting her head on the floor.

Based on the credible evidence established at the hearing, the Court finds Michael Baden, M.D., to be an expert in the area of pathology with respect to his testimony. Dr. Baden is a retired pathologist who served as the Chief Medical Examiner of New York City, as well as the Director of the New York State Police Medico-Legal Investigations Unit. Dr. Baden testified that he reviewed the autopsy report in this case, photographs of the autopsy and the scene, the ambulance report, hospital records, and trial transcripts.

Referring to the autopsy report, Dr. Baden testified that Brittney's injuries, "were classically due to a fall." Dr. Baden disagreed with Dr. Rubio's conclusion that Brittney's injuries were consistent with rotational forces and opined that Brittney had a coup/contrecoup injury. Dr. Baden explained that a coup injury only occurs if the moving head strikes something. A contrecoup indicates an impact site on one area of the head and a bruise on the brain 180 degree opposite thereto. The presence of a coup/contrecoup injury signifies that the head had to be moving at the time that it struck the ground, which is typical of a fall.

In Dr. Baden's opinion, the opinions and conclusions of Dr. Rubio were inconsistent with some of her findings. By way of example, Dr. Baden noted that the cause of death listed in Brittney's autopsy report was multiple brain injuries due to rotational forces (Shaken Impact baby Syndrome) homicide. Rather than multiple brain injuries, however, Dr. Baden noted that there were no brain injuries other than the two that

he had previously referenced in the back of the brain and the front side of the brain, which "are classic for a fall and don't support the concept of shaken baby. And there is no subdural hemorrhage, which is part of the importance of the shaken baby that she describes."

The Prosecution did not deny that short falls can be fatal. Rather, they countered that fatal falls are so rare as to be inconsequential. Sandeep Narang, M.D., a pediatrician who appeared on behalf of the Prosecution, summarized the changes in short fall literature since 2001. He testified that short falls were better defined as five feet or less, and that there has been better biomechanical study of the forces involved in short falls, but the conclusion reached had been the same: that deaths from short falls are possible, but rare.

On cross-examination Dr. Narang testified that, "Yes," the epidemiology literature suggests that short falls can kill, although rarely. The Defense asked, "So, again, if the doctors at this trial testified that short falls cannot cause this, that's just wrong, isn't it?" Dr. Narang answered, "If they testified to that, yes, sir." Nevertheless, Dr. Narang conceded that in 1997, leading physicians in pediatrics were stating that the three findings which comprised the triad were virtually unique to Shaken Baby Syndrome.

Barbara Wolf, M.D., a pathologist who appeared on behalf of the Prosecution, opined that Brittney's death was caused by an impact to her head, with or without shaking. Furthermore, contrary to Dr. Rubio's Trial testimony, Dr. Wolf did not believe that Brittney was shaken and then placed against a soft surface, given that there was a bruise on Brittney's brain.

The credible evidence adduced at the hearing established, however, the falsity of the existing perception at the time of Trial, that short falls or low velocity impacts could not cause death. In that regard, Dr. Plunkett's 2001 paper included a number of cases that were either witnessed by non-family members or by a number of adults. One case was documented by a video which showed a low velocity impact that resulted in death. The child in that videotaped fall was 23 months old; she was about the same height, but "a little heavier" than Brittney.

The video depicting the fall was received, under seal, as a Court exhibit. The Court and the parties' respective attorneys viewed the video at the hearing. In that footage, the 23 month old girl and her older brother were playing on a plastic gym-type house in the garage of their parents' home. As the girl was straddling one of the rails, holding on with her hand, she lost her grip and fell sideways. The girl's head was about three and a half feet above the ground when she went into a free fall and struck the ground (carpet over a concrete garage floor). The girl first struck with her outstretched hands, and then with the right side of her forehead. She initially was conscious, but five or ten minutes later she had a seizure and was brought into a local hospital. The girl developed a large volume subdural hematoma, which was surgically evacuated, but she subsequently developed malignant cerebral edema and died. The fall occurred in 1993, but Dr. Plunkett did not become aware of it until 2000, when he accidentally came across the incident in the U.S. Consumer Product Safety Commission database. The Court credits Dr. Plunkett's testimony that the incident did not appear in the press, or anywhere else that he was aware.

Based on the credible evidence established at the hearing, the Court finds Patrick Lantz, M.D., to be an expert in the area of pathology with respect to his testimony. In that regard, the Court credits Dr. Lantz's testimony that, at the time of Trial, ophthalmologists believed that only the acceleration/deceleration forces generated by violent shaking could cause retinal hemorrhages. Subsequently, similar eye findings were made in relation to crush injuries, falls, and traffic accidents. Dr. Lantz testified that, based upon his review of the medical records and imaging studies in this case, Brittney had retinal hemorrhages that developed while she was in the hospital.

Brian Forbes, M.D., a pediatric ophthalmologist who appeared on behalf of the Prosecution, opined that retinal hemorrhages, including those seen in Brittney, are not consistent with a history from a short distance fall. On cross-examination, Dr. Forbes was asked to compare two writings of the American Academy of Ophthalmology (AAO), one of which was printed from the AAO web site and entitled, "Shaken Baby Resources." The following excerpt from that writing, which dated back to 2003, was read into the record by Defense counsel:

> When extensive retinal hemorrhage accompanied by perimacular folds and schisis cavities is found in association with intracranial hemorrhage or other evidence of trauma to the brain in an infant, *shaking injury can be diagnosed with confidence regardless of other circumstances* [emphasis added].

The second writing was entitled, "Information Statement, Abusive Head Trauma/Shaken Baby Syndrome. The following excerpt from that writing, which was dated June 2010, was read into the record by Defense counsel:

> When extensive retinal hemorrhage accompanied by perimacular folds and schisis cavities are found in association with intracranial hemorrhage or other evidence of trauma to the brain in an infant *without another clear explanation*, abusive head trauma can be diagnosed with confidence regardless of other circumstances [emphasis added].

Dr. Forbes admitted that the relevant medical community knows far more about retinal hemorrhages in 2014 than it did in 2001-2002.

Daniel Lindberg, M.D., an emergency room physician who was called to testify on behalf of the Prosecution, stated his belief that the phrase, "shaken child syndrome" was an unfortunate shorthand which could encompass impact. Dr. Lindberg also testified that retinal hemorrhages can be caused by many different types of trauma, but the severity of Brittney's retinal hemorrhages was inconsistent with a fall from a height of 18 inches. Contrary to Dr. Lindberg's testimony, the Court finds the testimony of Julie Mack, M.D. to be persuasive, in that the retinal hemorrhaging seen in Brittney's eyes was consistent with a short distance fall.

Based on the credible evidence established at the hearing, the Court finds Julie Mack, M.D., to be an expert in the area of radiology with respect to her testimony. In that regard, Dr. Mack testified that, while the idea that shaking caused bridging vein rupture was widely accepted at the time of the Defendant's Trial, it now is clear that shaking does not generate enough force to produce a bridging vein rupture. Furthermore, Dr. Mack reviewed Brittney's CT scans and MRI, and concluded that the radiology in this case was inconsistent with a bridging vein rupture.

The Court credits Dr. Mack's testimony that the swelling of Brittney's brain developed swiftly, and that it progressed while she was in the hospital. Brittney's first CT was performed approximately three hours after she allegedly fell. Shortly after Brittney's second CT, which was done around midnight, a monitor was placed into Brittney's brain to check the pressure; the pressure was high enough to significantly limit the amount of blood getting into the brain. That is, the brain was not being perfused even though the heart still was pumping. Given that the blood had to go elsewhere, there was distension in places where blood was not ordinarily seen.

The Court credits Dr. Mack's testimony that there was an altered blood flow pattern, which included blood going forward into the eye, including the retinas. The radiology was unequivocal that the hemorrhage progressed, becoming worse in the hospital.

In Brittney's case, there was the unusual occurrence of rapidly developing brain swelling. Blood seen at the autopsy might have represented a natural progression, rather than simply trauma. Dr. Mack summarized:

> So, really the summary of the imaging is we have severe, rapidly developing brain edema, progressing to brain death in the course of a day, associated with only a small amount of extra-axial hemorrhage. The contention is an alleged fall injury.

Dr. Mack acknowledged that radiology cannot distinguish between an injury that was accidental, versus an injury that was caused intentionally. Radiology can, however, provide the basis for findings that are consistent, or inconsistent with the provided clinical

history. In this case, the radiology was consistent with a small amount of bleeding and a contrecoup injury, such as a short distance fall from a chair.

Based on the credible evidence established at the hearing, the Court finds John Galaznik, M.D., to be an expert in the area of pediatrics with respect to his testimony. Dr. Galaznik testified regarding the change in opinion of the American Academy of Pediatrics, regarding head injury in children. At the hearing, Dr. Galaznik was questioned relative to a 2001 article published in *Pediatrics*, the official journal of the American Academy of Pediatrics, entitled, "Shaken Baby Syndrome: Rotational Injuries – Technical Report." Consistent with the contents of that article, Dr. Galaznik, as a pediatrician, understood that it was not possible for the constellation of injuries at issue to occur with a short fall.

In 2009, the American Academy of Pediatrics (AAP) published an article entitled, "Abusive Head Trauma in Infants and Children," in which it was acknowledged that injuries from accidental and abusive causes overlap. Further, the Academy removed the claim that short falls do not cause symptoms like those observed in Brittney Sheets. The abstract contained in the 2009 article states, in part:

> Shaken baby syndrome is a term used often by physicians and the public to describe abusive head trauma inflicted on infants and young children. Although the term is well known and has been used for a number of decades, advances in the understanding of the mechanisms and clinical spectrum of injury associated with abusive head trauma compel us to modify our terminology to keep pace with our understanding of pathologic mechanisms. Although shaking an infant has the potential to cause neurologic injury, blunt impact or a combination of shaking and blunt impact cause injury as well.

In 2010, the American Academy of Pediatrics published a clinical report, entitled, "The Eye Examination in the Evaluation of Child Abuse," which opened with the statement:

> Retinal hemorrhage is an important indicator of possible abusive head trauma, but it is also found in a number of other conditions.

The Court credits Dr. Galaznik's testimony that said statement represented a significant change from the AAP's 2001 position. That is, in 2001, retinal hemorrhages were presumed to indicate rotational head injury. By 2010, it was recognized that retinal hemorrhages could have multiple causes and be present in many situations. Therefore, retinal hemorrhages are non-specific.

Based on the credible evidence adduced at the hearing, the Court finds Patrick Barnes, M.D., to be an expert in the area of pediatric neuroradiology with respect to his testimony. Dr. Barnes reviewed current research and scientific literature regarding child abuse, Shaken Baby Syndrome, and the causes of head and brain injury in children. He outlined the research challenges, such as the circularity of many research designs.

Dr. Barnes stated his opinion, to a reasonable degree of medical certainty, that the findings in this case were more consistent with an impact injury than with a shaking mechanism without impact. Moreover, the findings were consistent with a fall from an 18 inch chair, according to the current, best available knowledge and science. The Court credits Dr. Barnes' testimony that the current, best available knowledge and science that led him to that conclusion was <u>not</u> available in 2001.

John Waldman, M.D., a pediatric neurosurgeon who testified on behalf of the Prosecution, also stated that there was no indication of a torn bridging vein in this case. Dr. Waldman explained, however, that a child who dies as a result of a short fall will suffer different injuries than those suffered by Brittney. The most common is an epidural hematoma, which acts as a space occupying lesion. When such an injury is sustained, there is slow bleeding between the dura and the skull, which expands and eventually crushes the brain until it dies. While an epidural hematoma is expanding, the victim may have a lucid interval. Dr. Waldman testified that epidural hematomas have been known to cause death, "as long as people have been falling down." However, without meaningful explanation, Dr. Waldman summarily concluded that Brittney Sheets did not suffer an epidural hematoma.

Two other injuries that Dr. Waldman testified might be suffered in a fatal, short fall are a subdural hematoma acting as a space occupying lesion and a carotid dissection. Dr. Plunkett testified that Brittney suffered a small volume acute subdural hematoma, in addition to malignant rapid brain swelling, contusion or bruising at the base of her left temporal lobe (a contrecoup contusion), and brain herniation.

On rebuttal, Dr. Plunkett explained the manner in which the terminology used by doctors has changed since 2002 when they describe injuries to a child's head, believed to be caused by abuse. Dr. Plunkett testified that forensic pathologists generally use the term "blunt head trauma" or "closed head trauma" to describe the results or cause of injury. Pediatricians tend to use the term "abusive head trauma" or "inflicted head trauma." When

asked his opinion as to the significance of that change in terminology, Dr. Plunkett testified, as follows:

> In terms of forensic pathologists, it's an acknowledgment that shaking is an unlikely, if not impossible mechanism for brain injury in an infant. In terms of pediatricians, I can only state what they said in 2008 or 2009, which is that they have changed the name from Shaken Baby Syndrome to Abusive Head Trauma because shaking was too narrow a definition of a mechanism of injury.

CONCLUSIONS OF LAW

The power to vacate a judgment of conviction upon the ground of newly discovered evidence and concomitantly grant a new trial rests within the discretion of the hearing court (*see People v McFarland*, 108 AD3d 1121 [4th Dept 2013]; *see also People v Tankleff*, 49 AD3d 160, 178 [2d Dept 2007], citing *People v Salemi*, 309 NY 208, 215 [1955], *cert denied* 350 US 950 [1956]). The court must make its final decision based upon the likely cumulative effect of the new evidence had it been presented at trial (*see* CPL § 440.10 [1] [g]; *see also People v McFarland*, 108 AD3d 1121 [4th Dept 2013]; *People v Bellamy*, 84 AD3d 1260 [2011], *lv denied* 17 NY3d 813 [2011]).

Criminal Procedure Law §440.10 (1) (g) states that the judgment may be vacated upon the ground that:

> New evidence has been discovered since the entry of a judgment based upon a verdict of guilty after trial, which could not have been produced by the defendant at the trial even with due diligence on his part and which is of such character as to create a probability that had such evidence been

> received at the trial the verdict would have been more favorable to the defendant; provided that a motion based upon such ground must be made with due diligence after the discovery of such alleged new evidence.

The new evidence may only be considered if it satisfies all the following criteria: (1) It must be such as will probably change the result if a new trial is granted; (2) It must have been discovered since the trial; (3) It must be such as could have not been discovered before the trial by the exercise of due diligence; (4) It must be material to the issue; (5) It must not be cumulative to the former issue; and (6) It must not be merely impeaching or contradicting the former evidence (CPL 440.10 [1] [g]; *People v Bryant*, 117 AD3d 1586 [4th Dept 2014]; *People v Hamilton*, 115 AD3d 12 [2d Dept 2014], citing *People v Salemi*, 309 NY at 216 [1955]; *People v Tankleff*, 49 AD3d 160, 178 [2d Dept 2007]). Implicit in the standard, set forth in CPL §440.10 (1) (g), is that the newly discovered evidence must be admissible (36A Carmody-Wait 2d §205:16).

Pursuant to CPL §440.30 (6), the Defendant bears the burden of proving, by a preponderance of the evidence, every fact essential to support the motion. The Defendant must overcome a presumption of validity attending the judgment of conviction and has the burden of going forward with allegations sufficient to create an issue of fact (36A Carmody-Wait 2d §205:85).

As to the probability of a different verdict, it is *not* sufficient that the Defendant demonstrate that there is a mere possibility that the jury would return a verdict more favorable to her, if presented with the newly discovered evidence. The proper standard is a probability that the result of the trial would be changed. (*See generally People v Jackson*,

238 AD2d 877 [4th Dept 1997], *lv denied* 90 NY2d 859 [1997].) By way of example, a motion to vacate on the grounds of newly discovered evidence would be denied where there was other, overwhelming evidence of the Defendant's guilt (34B NY Jur 2d Criminal Law: Procedure §3390).

The Court, with those considerations in mind, in conjunction with its Findings of Fact and the legal arguments of the respective parties, hereby concludes, as follows.

The credible and persuasive evidence presented by the Defense established, by a preponderance of the evidence, a significant change in medical science relating to head injuries in children, generally, and the Shaken Baby Syndrome hypothesis, in particular, since the time of the Trial in this matter. New research into the biomechanics of head injury reveals that the doctors who testified on behalf of the Prosecution at Trial misinterpreted the medical evidence to conclude that shaking, or shaking with impact, was the only mechanism capable of causing Brittney's injuries.

The People disputed the notion that the medical community did not accept the possibility that a short fall could be fatal until after January 2002. The Court determines, however, that the Defense established that the mainstream belief in 2001-2002, espoused by the Prosecution's expert witnesses at Trial, that children did not die from short falls, has been proven to be false. As more fully set forth in the Findings of Fact, the Court credited the testimony of the Defense experts that case studies have demonstrated that children have died from short falls, that biomechanical research has explained the force produced in falls, and that advances in imaging have undercut the theory that shaking causes fatal injury

through the tearing of bridging veins. The Court further determines that the availability of a text published in 2001 discussing the danger of falls does not undermine the Defendant's contention that there has been a sea change in medical belief regarding that danger.

The credible evidence adduced at the hearing also established that doctors view retinal hemorrhages very differently today than they did at the time of Trial. Even Dr. Forbes, a Prosecution witness, admitted that the relevant medical community knows more about retinal hemorrhages in 2014 than it did at the time of Trial. As more fully set forth in the Findings of Fact, at the time of Trial ophthalmologists believed that only the acceleration, and deceleration forces generated by violent shaking could cause retinal hemorrhages. At the hearing, Dr. Forbes agreed that doctors now know that other events, such as trauma, intracranial pressure, and many other events can cause retinal hemorrhages. Furthermore, Dr. Forbes conceded that the force generated by a single shake is similar to the force that would be caused by a fall.

Likewise, the credible evidence adduced at the hearing established changes in the field of pediatrics regarding head injury in children. In 2001, the American Academy of Pediatrics published an official paper stating that short falls do not cause the constellation of injuries, attributed at Trial to shaking. In 2009, the same organization published a new position paper acknowledging that injures from accidental and abusive causes overlap, and removing the claim that short falls do not cause symptoms like those observed in Brittney.

Similarly, in 2001, retinal hemorrhages were presumed to indicate rotational head injury, but by 2010, the American Academy of Pediatrics recognized that retinal hemorrhages can have many different causes.

Changes in the field of pediatric radiology concerning Shaken Baby Syndrome was reflected in the testimony of Dr. Baden and Dr. Barnes, who opined that it is impossible, relying on current medical knowledge and Evidence Based Medicine standards, to conclude that Brittney's injuries were inconsistent with a reported history of a fall from a chair onto a carpeted floor. Contrary to Dr. Rubio's determination at the time of Trial that Brittney was shaken, and her rejection of a short fall as an explanation for Brittney's death, Dr. Barnes' opinion was that Brittney's injuries were more consistent with a fall to the floor from an 18 inch chair than they were with shaking.

Although the Prosecution witnesses at the Hearing did not deny that short falls could be fatal, they countered that fatal, short falls are so rare as to be inconsequential, and that the injuries sustained by Brittney were not the kind of injuries caused by falls.

The People also challenged the relevance of the Defense argument that the "triad" is no longer pathognomonic for abuse, inasmuch as the Defense offered no proof that any medical expert for the Prosecution at Trial ever mentioned the "triad" or considered it to be dispositive in their diagnostic decision-making. Rather, the People contended that the medical experts at Trial testified that they considered the history provided by the Defendant, and Brittney's parents. In that regard, the Court finds the testimony of Dr. Barnes to be persuasive, such that in 2001-2002, when treating doctors observed the triad

of injuries, histories would be rejected unless a caretaker could provide an adequate explanation for the injuries, such as an automobile accident or a fall from two to three stories in height.

The Court is mindful of the Prosecution's argument that, although the Defense experts opined that the manner of Brittney's death was an accidental fall from a chair, those opinions do not constitute newly discovered evidence. Rather, they merely contradict former opinion testimony given at Trial by the Prosecution's expert witnesses. The People further argued that those differing opinions are not new, because the Defendant's medical expert at Trial testified that Brittney's injuries were consistent with a fall from a chair.

Nevertheless, the credible evidence adduced at the Hearing, which was supported by expert testimony from different disciplines and specialties – pediatrics, radiology, pathology, ophthalmology, and biomechanical engineering – established by a preponderance of the evidence that key medical propositions relied upon by the Prosecution at Trial were either demonstrably wrong, or are now subject to new debate.

The People argued that, even if the Court determined that there is new evidence regarding the lethality of short falls, it is not probable that a jury would acquit the Defendant at a new trial based on such evidence. The People posited that the jury would hear that it is exceedingly rare for a child to die from a short fall, and that the types of injuries a child suffers in a fatal short fall are different from the injuries Brittney suffered. Further, the jury would be left to examine that evidence in light of the Defendant's

arguably inconsistent version of events and the People's expert witnesses, who concluded that Brittney's injuries were inflicted, rather than caused by a fall.

The Court concludes, however, that in light of current information available to the medical and other scientific communities, it is unlikely that the Prosecution's experts at a new Trial would testify as adamantly, if at all, as they did in 2001, that Brittney's injuries were the type caused by shaking, and that they were not the type caused by a short fall (*see generally Cavazos v Smith*, __ US __, __; 132 S Ct 2, 21, [2011, Ginsburg, J., dissenting]). The credible evidence adduced at the hearing established that recent medical and scientific opinion significantly, and substantially, undermines that 2001 Trial testimony.

The newly discovered evidence in this case thus shows that there has been a compelling and consequential shift in mainstream medical opinion since the time of the Defendant's Trial as to the causes of the types of trauma that Brittney exhibited. Moreover, the Defense presented evidence that was not discovered until after the entry of judgment, in the form of expert medical testimony, that a significant and legitimate debate in the medical community has developed in the past 13 years, over whether young children can be fatally injured by means of shaking, particularly in consideration of the injuries suffered by Brittney at her age. Thus, the Court concludes that the evidence is of such character as to create a probability that it would change the result if a new Trial was granted. (*See generally State v Edmunds*, 2008 WI App 33 [2008]).

The Court notes that the due diligence requirement is measured against the Defendant's available resources and the practicalities of the particular situation (*People v*

Bryant, 117 AD3d 1586 [4th Dept 2014]). Accordingly, the Court concludes that the Defendant could not have produced such evidence at Trial, even with due diligence, as the credible evidence adduced at the Hearing demonstrated that the bulk of the medical research and literature supporting the Defense position, and the emergence of the Defense position in the medical community, only emerged in the 13 years following her Trial (*see generally State v Edmunds*, 2008 WI App 33 [2008]).

Further, the Court determines that the new medical testimony presents an alternate theory for the source of Brittney's injuries, and such evidence differs in substance and quality from the Defense evidence at Trial. The new evidence is material to the issue, and it is not cumulative , merely impeaching, or contradicting of the former evidence.

The Court thus concludes that the proffered expert witness testimony, concerning head injuries in children, does constitute "new evidence" as that term is contemplated by CPL § 440.10 (1) (g). The Defendant's alternative request to amend the motion to add a claim pursuant to CPL § 440.10 (1) (c), and to reinstate her claim pursuant to CPL § 440.10 (1) (h), is therefore rendered moot.

In light of the foregoing, the Court need not address whether the proffered testimony, concerning Sandra Hennessy's observations of Cameron Burnside's behavior, constitutes "new evidence" as that term is contemplated by Criminal Procedure Law § 440.10 (1) (g) or, whether it should be considered in support of this motion. The Defense argued that, unbeknownst to the parties at the time of Trial, for at least two years after leaving the Defendant's care, young Cameron was seen by Ms. Hennessy, his subsequent

daycare provider, to engage in a specific role playing game a couple of times per week. Cameron would speak to an imaginary friend named, "Brittney," whom he told to jump and encouraged her to "do it." Cameron then would use a particular stuffed animal to comfort his imaginary friend. Nevertheless, upon consideration of all of the testimony offered at the hearing from Sandra Hennessy, the Court finds that such testimony was credible, and compelling, but this Court is not considering that testimony upon reaching its decision in this matter.

Accordingly, it is hereby,

ORDERED, ADJUDGED AND DECREED, that the Defendant's motion for an order, pursuant to Criminal Procedure Law § 440.10 (1) (g), vacating the judgment of conviction and sentence in this matter, is hereby GRANTED; and it is hereby

ORDERED, ADJUDGED AND DECREED, that the Defendant is hereby GRANTED a new Trial, on the charge set forth in the above-referenced Indictment, on a date to be determined by the Court; and it is hereby

ORDERED, ADJUDGED AND DECREED, that the Defendant be brought back before the Court, forthwith, to schedule a new Trial.

The above constitutes the Decision and Order of this Court.

Dated: December 16, 2014
Rochester, New York

S/HON. JAMES J. PIAMPIANO
HON. JAMES J. PIAMPIANO
COUNTY COURT JUDGE

E N T E R