

California Law Review

Vol. 90

January 2002

No. 1

Copyright © 2002 by California Law Review, Inc.

The *Daubert/Kumho* Implications of Observer Effects in Forensic Science: Hidden Problems of Expectation and Suggestion

D. Michael Risinger,[†] Michael J. Saks,[‡]

William C. Thompson,^{††} & Robert Rosenthal^{‡‡}

Table of Contents

Introduction: The Requirements of <i>Kumho Tire Co. v. Carmichael</i>	3
I. Observer Effects.....	6
A. Evolution of the Awareness of Observer Effects.....	6
B. What This Article Is Not About (Honesty and Observer Effects).....	10
C. The Psychology of Observer Effects	12

Copyright © 2002 California Law Review, Inc. California Law Review, Inc. (CLR) is a California nonprofit corporation. CLR and the authors are solely responsible for the content of their publications.

[†] Professor of Law, Seton Hall University School of Law, B.A., Yale University, 1966, J.D., Harvard University, 1969.

[‡] Professor of Law and Professor of Psychology, Arizona State University, B.A., B.S., Pennsylvania State University, 1969, M.A., Ohio State University, 1972, Ph.D., Ohio State University, 1975, M.S.L., Yale University, 1989.

^{††} Professor, Department of Criminology, Law and Society, University of California at Irvine, J.D., University of California at Berkeley, 1982, Ph.D., Stanford University, 1984.

^{‡‡} Distinguished Professor of Psychology, University of California at Riverside, and Edgar Pierce Professor of Psychology, Emeritus, Harvard University, Ph.D., U.C.L.A., 1956.

1. In General.....	12
2. Observer Effects and Decision Thresholds	16
3. Anchoring Effects	17
4. Role Effects	18
5. Conformity Effects.....	19
6. Experimenter Effects.....	20
D. The Pervasiveness of Observer Effects.....	22
E. The Enhancement of Observer Effects by Desire and Motivation.....	24
F. The Lack of Linkage Among Confidence, Accuracy, and Amount of Information.....	26
II. Observer Effects in Forensic Science.....	27
A. Proper and Improper Information in the Forensic Science Practice	27
B. Improper Information Contamination in Forensic Science Practice	31
C. Specific Sources of Induced Observer Error in Forensic Science Practice	35
1. Direct Communication Between Investigators and Examiners.....	35
2. Revision of Findings in Light of New Test-Irrelevant Information.....	38
3. Selective Re-examination of Evidence.....	39
III. Minimizing Observer Effects in Forensic Science: Conclusions and Recommendations.....	42
A. Preventing Distortions Due to Expectation and Suggestion: Blind Testing	45
B. Preventing Distortions Due to Assumed Base Rates of Inculcation: Evidence Lineups	47
C. Likely Objections to the Recommendations.....	50
IV. Observer Effects and Admissibility Under Federal Rule of Evidence 702.....	53

The *Daubert/Kumho* Implications of Observer Effects in Forensic Science: Hidden Problems of Expectation and Suggestion

D. Michael Risinger, Michael J. Saks,
William C. Thompson & Robert Rosenthal

One must not equate ignorance of error with the lack of error. The lack of demonstration of error in certain fields of inquiry often derives from the nonexistence of methodological research into the problem and merely denotes a less advanced stage of that profession.¹

This is a criminal investigation, sir. You are asking about bias controls, which refers to research.²

Introduction

The Requirements of *Kumho Tire Co. v. Carmichael*

In *Kumho Tire Co. v. Carmichael*³ the United States Supreme Court put forward two important principles for the control of expert evidence. The first is that the judge's gatekeeping responsibility to insure minimum reliability of expert testimony pursuant to Federal Rule of Evidence 702⁴

1. Herbert H. Hyman et al., *Interviewing in Social Research* 4 (1954).

2. Robert Hazelwood of the FBI Behavioral Sciences Unit, responding to a question from Representative Nicholas Mavroules concerning the conduct of the investigation into the alleged responsibility of Clayton Hartwig for the 1989 explosion of the center gun turret on the battleship U.S.S. Iowa. H. Paul Vettters, *Who Killed Precious 183* (1991).

3. 526 U.S. 137 (1999).

4. When *Kumho Tire* was decided, Rule 702 read as follows: "If scientific, technical or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise." *Kumho Tire*, 526 U.S. at 147. Since *Kumho Tire* was decided, an amendment to Rule 702 has become effective that reinforces the case's task-specific approach. The new rule requires that "the testimony is the product of reliable principles and methods,"

applies to all proffered expert testimony, not just the explicit products of “science.”⁵ The second, less explicit but no less important, is that this judgment must be made concerning the “task at hand,”⁶ instead of globally in regard to the average dependability of a broadly defined area of expertise.⁷ In other words, reliability cannot be judged “as drafted,” but must be judged only specifically “as applied.” The Court repeatedly made this clear in *Kumho Tire*,⁸ perhaps best when it said:

contrary to respondents’ suggestion, the specific issue before the court was not the reasonableness in general of a tire expert’s use of a visual and tactile inspection Rather, it was the reasonableness of using such an approach, along with [the expert’s] particular method of analyzing the data thereby obtained, to draw a conclusion regarding *the particular matter to which the expert testimony was directly relevant* The relevant issue was whether the expert could reliably determine the cause of this tire’s separation.⁹

As the Court further stated, “Rule 702 grants the district court the discretionary authority . . . to determine reliability in light of the particular facts and circumstances of the particular case.”¹⁰

As a result of *Kumho Tire*, courts will be called upon to develop criteria for the proper delineation of both the “task at hand” and the particular circumstances affecting its reliability.¹¹ The development of such criteria is not a trivial task, both because individual cases may present complicated situations, as *Kumho Tire* illustrates, and because not all considerations that may bear on the reliability of an expert assertion should be taken into

Fed. R. Evid. 702(2), and that these principles and methods have been applied “reliably to the facts of the case.” Fed. R. Evid. 702(3). See *infra* text accompanying notes 232-234.

5. *Kumho Tire*, 526 U.S. at 141.

6. This phrase originally appeared offhandedly in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 597 (1993), but was quoted at the beginning of the *Kumho Tire* opinion, 526 U.S. at 141, and appropriately captures the particularized methodology of *Kumho Tire*.

7. The first principle is currently more widely perceived, but the second is no less central to the decision and potentially more important in the actual determination of cases. See generally D. Michael Risinger, *Defining the “Task at Hand”: Non-Science Forensic Science after Kumho Tire v. Carmichael*, 57 Wash. & Lee L. Rev. 767 (2000) (hereinafter Risinger, *Defining the “Task at Hand”*). Interestingly, in hindsight, one can see the *Kumho Tire* task-at-hand approach prefigured in the following language from Justice Rehnquist’s opinion in *General Electric v. Joiner*: “Of course, whether animal studies can ever be a proper foundation for an expert’s opinion was not the issue. The issue was whether these experts’ opinions were sufficiently supported by the animal studies on which they purported to rely.” 522 U.S. 136, 144 (1997).

8. All of the textual passages in the *Kumho Tire* opinion describing the task-at-hand approach and illustrating its application are analyzed in Risinger, *Defining the “Task at Hand,” supra* note 7, at 773-75.

9. *Kumho Tire*, 526 U.S. at 153-54.

10. *Id.* at 158.

11. *Id.* at 153.

account in a Rule 702 determination.¹² For example, it seems that at a minimum, expert veracity and sincerity are not proper Rule 702 factors and, for good or ill, are to be left to the evaluation of the trier of fact as they are in regard to fact witnesses.¹³ In addition, it seems inappropriate for a court to exclude relevant and reliable expert testimony simply because the judge had concluded based on other evidence in the case that the expert was simply wrong. Beyond this, however, after *Kumho Tire* it appears both appropriate and necessary for the judge to consider any factor that could be shown to affect the reliability of an expert's testimony under the "particular circumstances of the particular case."¹⁴ Because *Kumho Tire* obligates a trial court to make a reliability determination under Rule 702 where any proposed expert testimony's "factual basis, data, principles, methods, or their application are called sufficiently into question,"¹⁵ it would seem incumbent upon judges and lawyers to inform themselves concerning the status of knowledge bearing on such factors.

It is the aim of this Article to aid in this process. Specifically, we will show that there are certain factors which, when present, undermine to some degree the reliability of virtually any form of expertise. Further, we will show that the extent to which reliability is undermined depends not only on the presence of such factors, but on the characteristics of the expertise at issue, most particularly the degree to which it depends on subjective human judgment. Moreover, we will show that there is an entire established constellation of expertise, celebrated in popular culture and heretofore generally admissible, in which such factors form a rampant and uncontrolled part of normal practice. We will then put forward some practical proposals for reform of internal practice, and some suggestions about the proper legal response to an admissibility challenge under Rule 702.

The factors we refer to are primarily expectation and suggestion, which drive much of what is globally labeled "observer effects" in social psychology and research methodology. And the constellation of expertise

12. This was the thrust of the well-known line in *Daubert v. Merrell Dow Pharmaceuticals*: "[t]he focus, of course, must be solely on principles and methodology, not on the conclusions that they generate." 509 U.S. 579, 595 (1993). The meaning of this line was never very clear, and the Court's declaration in *General Electric v. Joiner*, 522 U.S. 136 (1997), that "conclusions and methodology are not entirely distinct from one another" at least partly undermined the vitality of its dichotomy between methodology and conclusions. *Id.* at 146. For a discussion of the potential continuing validity of the distinction between methodology and conclusions, see Michael J. Saks, *The Aftermath of Daubert: An Evolving Jurisprudence of Expert Evidence*, 40 *Jurimetrics* 229, 235-36 (2000).

13. No principle is more embedded in general evidence jurisprudence than that such a normal veracity-based judgment of "credibility" is for the trier of fact: "Even the trial court, which has heard the testimony of witnesses firsthand, is not to . . . assess the credibility of witnesses when it judges the merits of a motion for acquittal." *Burks v. United States*, 437 U.S. 1, 16 (1978). It has never been suggested that Rule 702 alters this in regard to expert witnesses.

14. 526 U.S. at 150.

15. *Id.* at 149.

is “forensic science” in general, and especially those forensic science practices utilizing subjective human judgment as their primary instrumentality,¹⁶ and not based on techniques derived from normal science methodology.¹⁷

I

Observer Effects

A. *Evolution of the Awareness of Observer Effects*

An elementary principle of modern psychology is that the desires and expectations people possess influence their perceptions and interpretations of what they observe. In other words, the results of observation depend upon the state of the observer as well as the thing observed. This insight is not new; long before cognitive scientists began formally studying the psychological foundations of such effects, the phenomenon was noticed and commented upon. Julius Caesar, for instance, noted that “men generally believe quite freely that which they want to be true.”¹⁸

Sensitivity to the problems of observer effects has become integral to the modern scientific method. Soon after Renaissance natural philosophers began creating the scientific method, they began paying specific attention to the problem of observer effects. The writings of Sir Francis Bacon in 1620, for example, recognized the problem. Bacon suggested that “[t]he human understanding, when any proposition has once been laid down . . . forces everything else to add fresh support and confirmation; and although . . . instances may exist to the contrary, yet [the understanding] either does not observe or despises them”¹⁹ Bacon also posited that “it is the peculiar and perpetual error of the human understanding to be more moved and excited by affirmatives than negatives, whereas it ought duly to be impartial; nay, in establishing any true axiom, the negative instance is the most powerful.”²⁰ In the first passage, Bacon anticipated what modern

16. These forensic sciences include such fields as handwriting identification, bitemark identification, toolmark examination, and so forth.

17. Gas chromatography and scanning electron microscopy are two examples of techniques derived from normal science methodology. Even those areas with good scientific antecedents, such as DNA identification, can have surprising problems under some circumstances. See generally William C. Thompson, *Subjective Interpretation, Laboratory Error and the Value of Forensic DNA Evidence: Three Case Studies*, 96 *Genetica* 153 (1995). For instance, the DQa/polymarker DNA test can present highly ambiguous results when mixed samples are involved, which require the same kinds of subjective human interpretation as, say, toolmark or bitemark identification. See William C. Thompson, *Accepting Lower Standards: The National Research Council's Second Report on Forensic DNA Evidence*, 37 *Jurimetrics* 405, 414 n.24 (1997).

18. “(H)omines fere credunt libentur id quod volunt.” G. Julius Caesar, *Caesar's Commentaries on the Gallic War* 155 (51 B.C.E.) (Frederick Holland Dewey ed., Translation Publishing Co. 1918).

19. Francis Bacon, *Novum Organum*, Book I, 109, point 46 (1620), *reprinted in* 30 *Great Books of the Western World* 110 (Robert M. Hutchins ed., 1952).

20. *Id.*

research has shown to be the cognitive phenomenon of selective attention: the tendency of observers to seek out some information and avoid other information.²¹ In both passages, Bacon anticipated what modern cognitive scientists refer to as confirmation bias, the tendency to test a hypothesis by looking for instances that confirm it rather than by searching for potentially falsifying instances, even though most scientists and philosophers of science today agree with Bacon that the best scientific method is to proceed by doing the latter.²² Bacon adds that “[t]he human understanding resembles not a dry light, but admits a tincture of the will and passions, which generate their own system accordingly, for man always believes more readily that which he prefers.”²³ Like Caesar before him, Bacon took a step beyond cognition and raised the issue of motivational or attitudinal effects on what a person thinks he or she has observed.

Perhaps the first recorded instance of a scientist recognizing that the attributes of an observer were influencing the accuracy of particular observations occurred more than 200 years ago. In 1795, Nevil Maskelyne, Astronomer Royal at the Greenwich Observatory, realized that he and his assistant were obtaining different results for the times of stellar transits, even though they were using identical methods.²⁴ These discrepancies reflected differences in complex judgments: “a coordination between the eye and the ear . . . a spatial judgment dependent upon a fixed position . . . an actual but instantaneous position of a moving object, and a remembered position no longer actual.”²⁵

In the 1820s, Bessel, an astronomer at Königsberg, studied the problem and found that such differences were not only common, but in astronomical measurements they reflected predictable individual tendencies.²⁶ By the 1830s astronomers had developed a method for calculating “personal equations” that enabled them to measure these particular kinds of

21. See Arthur S. Reber, *The Penguin Dictionary of Psychology* 669 (2d ed. 1995) (defining selective attention as “[t]he process involved in situations in which one is confronted with multiple stimulus inputs and must select but one aspect of them and attend to it”); see also John A. Bargh, *Automaticity in Social Psychology*, in *Social Psychology: Handbook of Basic Principles* 169, 174 (E. Tory Higgins & Arie W. Kruglanski eds., 1996) [hereinafter, *Social Psychology*] (describing selective attention research); James M. Olson et al., *Expectancies*, in *Social Psychology*, *supra*, at 211, 217 (describing the impact of expectancy on selective attention).

22. See Reber, *supra* note 21, at 151 (defining confirmation bias as “[t]he tendency to seek and interpret information that confirms existing beliefs”). As to the generally understood primacy of skeptically proceeding by attempting to falsify, see, for example, Donald B. Calne, *Within Reason: Rationality and Human Behavior* 220 (1999) (“If the working hypothesis withstands all attempts to refute it, new knowledge can be claimed.”).

23. Bacon, *supra* note 19, at 111, point 49. A “dry light” is a condition “in which one sees things without prejudice, uninfluenced by personal predilection.” *New Shorter Oxford English Dictionary* 758 (1993).

24. Edwin G. Boring, *A History of Experimental Psychology* 134-35 (1929).

25. *Id.* at 134.

26. *Id.* at 134-38.

observer error, adjust for them, and remove the distorting effects from their findings.²⁷

Scientists since that time have learned that observer factors can distort findings and produce misleading conclusions in myriad ways not so easily corrected for. The following are illustrations from a variety of fields.²⁸

Sir Isaac Newton failed to report absorption lines in the prismatic solar spectrum, though they would have been clearly visible with the apparatus he was using.²⁹ The most likely explanation for his failure to see them is that he held theoretically based expectations that such phenomena should not exist.³⁰ Because he believed they did not exist, he failed to see them, or at least to note their presence.

While Newton failed to see something that did exist, scientists of the early twentieth century saw something that did not exist. First reported by Rene Blondlot in 1903, "N-rays" appeared to make reflected light more intense.³¹ So long as they were believed to exist, the effects of N-rays were "observed" by many scientists.³² Of course, once it was determined that N-rays did not exist, their effects ceased to be observed.

Observer effects also have been found in the reading of scales. That is, people do not always read dials and other readouts correctly, and their errors are nonrandom. Certain numbers or patterns are more likely to be "read" than others, resulting in systematic errors in the data read from the measuring instruments.³³

For many years, laboratory technicians who counted blood cells visually were taught that correct counting would keep blood cell counts within a certain range of variation. In 1940, using a more accurate photographic method to count blood cells, researchers discovered that for years technicians had been reporting blood cell counts that were within an impossibly narrow band of variability.³⁴ The technicians made observations consistent with the expectations they held, but inconsistent with reality.

Mendel's counts of characteristics in pea plants came much closer to the theoretical predictions than is likely to have been possible.³⁵ Mendel or his assistant either deliberately misreported, or were the victims of observer effects induced by expectation.

27. *Id.*

28. For a useful discussion with many examples of error resulting from both observer effects and outright fraud in science, see Alexander Kohn, *False Prophets* (rev. ed. 1989).

29. Edwin G. Boring, *Newton and the Spectral Lines*, 136 *Science* 600, 600-01 (1962).

30. *Id.*

31. Kohn, *supra* note 28, at 18-20.

32. *Id.*

33. G. Udny Yule, *On Reading a Scale*, 90 *J. Royal Statistical Soc'y* 570 (1927).

34. Joseph Berkson et al., *The Error of Estimate of the Blood Cell Count as Made with the Hemocytometer*, 128 *Am. J. Physiology* 309, 322 (1940).

35. R.A. Fisher, *Has Mendel's Work Been Rediscovered?*, 1 *Ann. Sci.* 115, 132-34 (1936); Kohn, *supra* note 28, at 39-45.

One medical researcher found observer errors in the use of the stethoscope in cardiac diagnostics, leading him to suggest that physicians as well as their stethoscopes needed to be calibrated.³⁶ Another medical researcher, after finding medical students observing quite inaccurately when presented with two x-rays of hands to study, concluded that “[o]ur assumptions define and limit what we see, i.e., we tend to see things in such a way that they will fit in with our assumptions even if this involves distortion or omission.”³⁷

A writer on marine biology, reflecting on problems of animal observation, commented that scientists may “equate what they think they see, and sometimes what they want to see, with what actually happens.”³⁸

These realizations and attention to them have evolved into a “science of science,” a careful study of the causes of the random and systematic errors induced by observer effects and the methods for their prevention.³⁹ The results of such work can be found in the classrooms, textbooks, and laboratories of virtually all scientific fields, where methods and procedures have been developed to minimize the impact of such distorting influences. Today, awareness of such problems and their solutions is so widespread that concepts such as double-blind⁴⁰ and placebo⁴¹ have become household words popularly understood well beyond the laboratory, and analogous error-prevention techniques are employed in settings beyond the domain of science. For example, in many schools, including of course nearly every law school, teachers are required to grade examinations without knowing the identity of the student. Other common examples of such anonymous evaluations include auditions for symphony orchestras where the candidates may play behind a screen, and academic journals, many of which conduct blind peer review of submissions.

Forensic science is one of a very few fields that has not yet profited from this “science of science.” The most obvious danger in forensic science is that an examiner’s observations and conclusions will be influenced by extraneous, potentially biasing information. However, there are other

36. Alvan R. Feinstein, *The Stethoscope: A Source of Diagnostic Aid and Conceptual Errors in Rheumatic Heart Disease*, 11 *J. Chronic Diseases* 91, 100 (1960).

37. M.L. Johnson, *Seeing’s Believing*, 15 *New Biology* 60, 79 (1953).

38. Frank W. Lane, *Kingdom of the Octopus* 85 (1960).

39. See generally Robert Rosenthal, *Experimenter Effects in Behavioral Research* (1966) [hereinafter Rosenthal, *Experimenter Effects*]. As Seymour Kety observed, a “source of error which must be recognized is one which is common to all of science and which it is the very purpose of the scientific method, tradition, and training to minimize—the subjective bias.” Seymour S. Kety, *Biochemical Theories of Schizophrenia (Part I)*, 129 *Science* 1528, 1529 (1959).

40. “[A]n experimental procedure in which neither the subjects nor the experimenters know the make-up of the tests and control groups.” Webster’s Third New Int’l Dictionary 74a (1993).

41. “[A]n inert medicament or preparation given for its psychological effect esp. . . . as control in an experimental series.” *Id.* at 1727.

potentially error-producing sources of expectation beyond those induced by intentional or unintentional suggestion. In the Parts below, we will review some of the most important research on observer effects, focusing on those that result from expectancy and those that result from the context in which problems are presented for solution. We will further discuss the likely role of such effects in forensic science work as it is currently performed.

B. What This Article Is Not About (Honesty and Observer Effects)

Before turning to the principal foci of this Article, let us be clear about the problems that are not the direct concern of this Article. When we talk about distortions due to extraneous influences, we are not talking about deliberate falsification—when forensic scientists report inculpatory results when the findings were actually exculpatory or inconclusive, or when they have conducted no examinations at all. Documented examples of such misconduct, such as the false fingerprint reports of a David Harding⁴² or the blood group testimony of a Fred Zain⁴³ or a Thomas Curran,⁴⁴ are well-known, though how common such actions are is not.

According to Professor Andre Moenssens, the temptation to deliberately falsify results, whether fudging them or creating them out of whole cloth, is ever-present among forensic scientists.⁴⁵ Consider, for example, the story recounted by Evan Hodge, former chief of the FBI Firearms and Toolmark Unit, concerning a police inspector who brought a Colt Arms forty-five caliber pistol to a firearms examiner so the barrel's rifling could be compared to the marks on the murder bullet. The inspector in effect told the examiner: "We know this guy shot the victim and this is the gun he used. All we want you to do is confirm what we already know so we can get a warrant to get the scumbag off the street. We will wait. How quick can you do it?"⁴⁶ The examiner required little time to provide the requested

42. Harding was a New York State Trooper who, along with others in his unit, falsified fingerprints and other evidence to insure convictions in numerous cases. *See, e.g.*, *People v. Longtin*, 707 N.E.2d 418 (N.Y. 1998). *See generally* Nelson E. Roth, *The New York State Police Evidence Tampering Investigation: Report to the Honorable George Pataki, Governor of the State of New York*, Pursuant to Section Six of the New York State Executive Law (Jan. 20, 1997).

43. This case is summarized in Paul C. Giannelli, *The Abuse of Scientific Evidence in Criminal Cases: The Need for Independent Crime Laboratories*, 4 Va. J. Soc. Pol'y & L. 439, 442-47 (1997).

44. *See* John F. Kelly & Phillip K. Wearne, *Tainting Evidence* 13-14 (1989).

45. Professor Moenssens has written that forensic science experts are often tempted "to fabricate or to exaggerate" results. Andre A. Moenssens, *Novel Scientific Evidence in Civil and Criminal Cases: Some Words of Caution*, 84 J. Crim. L. & Criminology 1, 17 (1993). Indeed, according to Professor Moenssens, "[a]ll experts are tempted, many times in their careers, to report positive results when their inquiries come up inconclusive, or indeed to report a negative result as positive . . ." *Id.*

46. Evan Hodge, *Guarding Against Error*, 20 Ass'n Firearms & Toolmark Examiners' J. 290, 292 (1988). It should be noted that it is at least possible that even this fairly dramatic example could be the product of honest expectancy error and not of deliberate falsehood. Such

identification, which was then used as part of the interrogation that resulted in the defendant's confession.⁴⁷ The defendant then led the police to a second Colt pistol, which subsequent tests showed was the actual murder weapon.⁴⁸

That such misbehavior should be deterred, and punished when it is uncovered, cannot be doubted. But that is not a topic for the present Article. We might note, however, that to the extent examiners are prevented from knowing extraneous facts or theories of a case, those who might be tempted to falsify are handicapped in any efforts to deliberately produce false echoes of those facts and theories. The more sophisticated the error-avoidance procedures, the more difficult deliberate falsification will be rendered, and the more the temptation to do so will be reduced.

In addition to conscious falsification, there are other sources of error beyond the scope of this Article. We will not discuss, for example, the invention and use of novel but unvalidated techniques by maverick forensic scientists, such as Michael West's "blue light" for the discovery of bite marks or other impressions on skin, or Louise Robbins's footprint matching techniques.⁴⁹ Nor will we focus on the use of unvalidated techniques inherited from a time less concerned with validity, which have nevertheless come into wide use and generate conclusions that are of unknown accuracy, and which are currently being tested and corrected.⁵⁰ Finally, we will not consider ordinary incompetence; that is, forensic scientists who simply do not know how to do their technical jobs properly and as a result unintentionally reach erroneous conclusions.

The focus of this Article is on the far more pervasive but generally unnoticed error stemming from observer effects, a problem in some respects more troublesome and troubling than the intentional misconduct mentioned above. If permitted to run uncontrolled through forensic practice, observer effects can lead competent and honest forensic scientists, using well-validated techniques, to offer sincere conclusions that are, nevertheless, distorted and inaccurate. Such results may occur in large numbers, completely without examiner awareness, much less with any

"knave or fool" problems are not uncommon. Absent more particular evidence, readers must make their own decisions in this regard.

47. *Id.*

48. *Id.*

49. Michael West is a rather notorious forensic odontologist. One of his claims was that he had developed a special "blue light" that allowed him to see bite marks and other impressions on human skin where none were apparent. Unfortunately, only he could see them. Louise Robbins made similar claims for her footprint matching techniques. These now discredited practices and others are described in Giannelli, *supra* note 43, at 453-62, and Kelly & Wearne, *supra* note 44, at 13-14.

50. See Nat'l Institute of Justice, *Forensic Sciences: Review of Status and Needs 27-59* (1999) (U.S. Doc. J 28.23:F 76). Perhaps the best known example is handwriting identification expertise. See generally D. Michael Risinger, *Handwriting Identification*, Ch. 28, in David Faigman et al., *Modern Scientific Evidence* (2d ed. West 2002).

wrongful intent. Indeed, such distortions will be more ubiquitous and more insidious precisely because they are not intended and their presence goes unnoticed. In short, this Article focuses on the distorting effects that motivational bias and examination-irrelevant information can have on the conclusions of even those forensic scientists with the most sincere and honest intentions.

C. *The Psychology of Observer Effects*

1. *In General*

As we have already noted, an elementary principle of psychology is that context and expectations influence an individual's perceptions and interpretations of what he observes. Depending upon details of the process, its setting, or the theoretical model offered to explain the phenomenon, there are several terms which refer to this basic phenomenon, or particular aspects of it, including observer effects, context effects, expectancy effects, cueing, top-down processing, perceptual set, and others. In this Article, we will use the term "observer effects" to denote the general phenomenon, with other terms used to elucidate particular aspects of the general phenomenon.

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. A simple illustration of this phenomenon is provided in Figure 1:

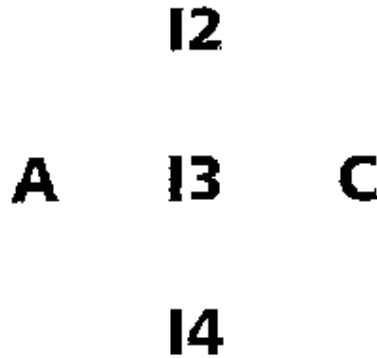
Figure 1⁵¹



What people see in the middle drawings depends upon the order in which they examine the series of drawings. People who begin at the right see the middle drawings as a woman's figure; people who begin at the left see the middle drawings as a man's face. Figure 2 presents an even simpler illustration.

51. See Gerald H. Fisher, *Ambiguity of Form: Old and New*, 4 *Perception and Psychophysics* 189 (1968).

Figure 2



Whether the character in the center is seen as the letter “B” or the number “13” depends upon the context in which it is viewed, specifically, whether one begins viewing vertically or horizontally. The context enables an observer to resolve the ambiguous symbol into one option or the other. Whether that resolution is “correct” or not is a separate matter.

Very often observer effects result from expectations about the results of an observation, and such expectations often come either from explicit messages or from subtle cues about the thing to be observed. For example, a pathologist who is told she is being presented with a slide of abnormal cells is more likely to conclude that she is seeing abnormal cells than one who is told she is being presented with a slide of normal cells.⁵²

None of this is to say that, inevitably and always, people simply see what they want to see or what they have been asked to see. The cognitive psychology underlying observer effects is best understood as a cyclical interplay between pre-existing schemata and the uptake of new information. Schemata are mental categories constructed from experience and belief that provide the framework for perception and reasoning.⁵³ Without

52. In standard parlance, the term “observer error” refers to errors that are randomly distributed, and therefore self-canceling over the long run, while the term “observer bias” refers to errors that are not random but systematic. Note, however, that random observer error may be a serious problem in any process which is not cumulative, but which relies on validity in regard to each individual result. It is of no comfort to individual patients if a pathologist “in the long run” makes as many errors calling normal cells cancerous as she does calling cancer cells normal. In principle at least, biased error of a known and stable amount is actually easier to deal with, since it can be corrected, while random error cannot. Indeed, the “personal equations” of nineteenth century astronomers were mechanisms to correct for such stable biased error.

53. Here we are using the term “schema” (plural “schemata”) in its most general sense. It is not surprising that, in theoretical attempts to map cognitive organization, taxonomies are generated which

schemata to organize and order perception and inference, the world of perception would remain William James's "blooming, buzzing confusion."⁵⁴ However, schemata not only facilitate meaningful perception, they also limit it. The eminent cognitive scientist Ulric Neisser explains this mildly paradoxical aspect of meaningful perception as follows:

Perception does not merely serve to confirm preexisting assumptions, but to provide organisms with new information. Although this is true, it is also true that without some preexisting structure, no information could be acquired at all. There is a dialectical contradiction between these two requirements: we cannot perceive *unless* we anticipate, but we must not see *only* what we anticipate. If we were restricted to isolated and separate glances at the world, this contradiction would prove fatal. Under such conditions, we could not consistently disentangle what we see from what we expect to see, nor distinguish objects from hallucinations. This dilemma . . . can be resolved in the perceptual cycle. Although a perceiver always has at least some (more or less specific) anticipations before he begins to pick up information about a given object, they can be corrected as well as sharpened in the course of looking.

The upshot of the argument is that perception is directed by expectations but not controlled by them; it involves the pickup of real information. Schemata exert their effects by selecting some kinds of information rather than others, not by manufacturing false percepts or illusions If the environment is rich enough to support more than one alternative view (and it usually is), expectations can have cumulative effects on what is perceived that are virtually irreversible The interplay between schema and situation means that neither determines the course of perception alone.⁵⁵

Schemata may be stubbornly fixed in many dimensions in adults, and voluntarily revisable in those dimensions only with effort and training. However, schemata are adjustable in certain ways right down to the point of perception. The context of perception, including such things as emotional involvement and exterior suggestion, can set and tune by expectation the way in which schemata are brought to bear, not only on perception, but on the recall of the events of perception. Thus, not only do the rigid aspects of schemata contribute to potential observer effects, so do their flexible

use more complex terminology, with schemata representing the most concrete categories and other terms, such as "metaphor" and "theory" representing higher-order categories. See William H. Calvin, *The Cerebral Code* 161-63 (1996).

54. William James, *The Principles of Psychology*, 1890, ch. 13, *reprinted in* 53 *Great Books of the Western World* 318 (Robert Maynard Hutchins ed., 1952).

55. Ulric Neisser, *Cognition and Reality: Principles and Implications of Cognitive Psychology* 43-44 (1976) (topic heading omitted).

dimensions. These processes can occur before, during, and after observation. As we approach an occasion for observation we become “set” for what we are about to perceive. Experiments, for example, have been conducted in which investigative interviewers are given preinterview reasons to believe or to doubt the person being interviewed.⁵⁶ This assignment of expectation had effects on the interview structure, the questions posed, and other aspects of interviewing behavior of the investigators.⁵⁷

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. Indeed, “[t]he selection process is programmable, within the fixed sensory limits.”⁵⁸ For example, we can stand in a crowd of noisy people and shift our focus from listening to one person to listening to another. Figures 1 and 2, above, provide two additional examples.

Expectation, whatever its source, plays into the previously noted phenomenon of confirmation bias and lays the groundwork for selective attention to evidence. Often there is too much information for a human to process or to give equal consideration to all of it. If one has expectations about an event, or hypotheses about its cause, one tends to draw selectively from the available evidence and focus on those items that confirm the working hypothesis. As Seymour Kety suggested, “it is difficult to avoid the subconscious tendency to reject for good reason data which weaken a hypothesis while uncritically accepting those data which strengthen it.”⁵⁹

Thus, expectations, among other factors, lead us to conclude more readily that we have perceived one thing rather than another, and having done so it becomes more difficult to perceive details that run contrary to the original perception. These effects can be reinforced as we establish the initial interpretation of what we have perceived (“constructive effects”), and further still when we later try to remember what we perceived (“reconstructive effects”). Indeed, there is evidence that the most powerful effects occur during the integration and retrieval phases, as the new percepts become part of the original schema and the schema is used to recall the perception.⁶⁰ In light of this, consider the forensic scientist who takes

56. Frans W. Winkel & Leendert Koppelaar, *Perceived Credibility of the Communicator: Studies of Perceptual Bias in Police Officers Conducting Rape Interviews*, in *Psychology and Law: International Perspectives* 223 (Friedrich Lösel et al. eds., 1992).

57. *Id.* at 227.

58. Robert E. Ornstein, *The Psychology of Consciousness* 50 (2d ed. 1977).

59. Kety, *supra* note 39, at 1529.

60. See, e.g., Charles P. Bloom, *The Role of Schemata in Memory for Text*, 11 *Discourse Processes* 305 (1988); Carol Anne M. Kardash, Doris Blender, & Thomas Bliesener, *Effects of*

poor notes during an examination and prepares a skimpy report, but then goes back to “spruce them up” shortly before trial.⁶¹ Even assuming the most honest of intentions, that examiner is inviting errors to infiltrate his conclusions and his testimony. The error potential of the original skimpy report, which leaves much to be supplied from memory, facilitates the creation of testimony more consistent with assumptions and later acquired expectations than would be the case with a more detailed and complete contemporaneous account. Reconstructive errors are given room to manifest themselves during the “spruce-up” stage.

2. *Observer Effects and Decision Thresholds*

One important area of research deals with how humans perceive and process information carrying “signal” stimuli in the presence of nonsignal stimuli, generally referred to as “noise.” One well-established effect of expectation, however induced or derived, in the perception tuning process is that decision thresholds shift as a function of expectations.⁶² Thus, in response to identical stimuli, a positive decision becomes more likely, and therefore more likely to be a false positive, or less likely, and therefore more likely to be a false negative, purely as a consequence of decision thresholds that change as expectations change.⁶³ 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.

Schemata on Both Encoding and Retrieval of Information from Prose, 80 *J. Educ. Psychol.* 324 (1988).

61. This practice is noted as common by FBI Laboratory examiner Terry Rudolph in the Department of Justice Inspector General’s Report on the FBI Laboratory’s practices. *See Office of the Inspector General, U.S. Dep’t of Justice, The FBI Laboratory: An Investigation into Laboratory Practices and Alleged Misconduct in Explosives-Related and Other Cases 44* (1997) (U.S. Doc. J 1.14/2:L 11/2), available at <http://www.usdoj.gov/oig/fbilab1/fbil1toc.htm> [hereinafter *Inspector General’s Report*]. The practice was condemned by the Report. *Id.* at 50. Even more dangerous is the failure to document results at all, but to rely entirely upon memory at trial, a practice noted in regard to some test results. *Id.* at 26. Rudolph justified the practice by saying that “I don’t write my notes for the defense. I write my notes for myself.” *Id.*

62. A “decision threshold” is the point at which salient or relevant signal information, in the presence of masking noise, is taken to be sufficiently clear to decide on the presence and meaning of the signal. A mixed signal/noise stimulus presents a form of ambiguity, and thus the results of signal studies reinforce the general proposition that observer effects manifest themselves most strongly under conditions of ambiguity and high subjectivity.

63. Victoria L. Phillips, James M. Royer & Barbara A. Greene, *The Application of Signal Detection Theory to Decision-Making in Forensic Science*, 46 *J. Forensic Sci.* 294, 296 (2001).

3. *Anchoring Effects*

Another line of relevant research on perception and recall involves what are known as “anchoring effects.”⁶⁴ Anchoring research shows that estimates people make of points along a continuum are influenced by positions that have been made salient by task-irrelevant outside influences. For example, in one test the subjects were given a percentage number that came from a spin of a wheel-of-fortune.⁶⁵ They were then asked whether the percentage of African nations in the U.N. was higher or lower than the number they had been given.⁶⁶ After answering this question, they were asked to give their best estimate of the actual percentage of African nations in the U.N.⁶⁷ Those given a higher random percentage on average gave substantially higher estimates than those given the lower percentage.⁶⁸

Additional research reveals that the anchors need not come from the same dimension along which the estimate of interest falls; arbitrary anchor values can produce large differences in people’s quantitative estimates. Karen Jacowitz and Daniel Kahneman ran a series of tests in an attempt to examine the breadth of this phenomenon. They used fifteen different tasks in which different anchors were introduced by asking respondents, in essence, “is X (see list of X’s below) larger than (or longer than or more than) Y,” which was either a high or low anchor value?⁶⁹ Respondents were then asked to give their best estimate of the true value of X.⁷⁰ The anchoring effect worked on fourteen of the fifteen tasks, including:

- Length of the Mississippi River
- Height of Mount Everest
- Amount of meat eaten per year by the average American
- Number of U.N. members
- Population of Chicago
- Maximum speed of a house cat
- Number of bars in Berkeley, California⁷¹

Research also demonstrates that expertise does not insulate one from the influence of anchoring effects. For instance, in one study, experienced

64. A review of the literature on anchoring effects is found in Thomas Mussweiler and Fritz Strack, *Comparing Is Believing: A Selective Accessibility Model of Judgmental Anchoring*, in 10 *Eur. Rev. Soc. Psychol.* 215 (Wolfgang Stroebe & Miles Hewstone eds., 1999).

65. Amos Tversky & Daniel Kahneman, *Judgment Under Uncertainty: Heuristics and Biases*, 185 *Science* 1124, 1128 (1974).

66. *Id.*

67. *Id.*

68. *Id.*

69. Karen E Jacowitz & Daniel Kahneman, *Measures of Anchoring in Estimation Tasks*, 21 *Pers. & Soc. Psychol. Bull.* 1161, 1162-63 (1995).

70. *Id.*

71. *Id.* The only task which displayed no anchoring effect involved giving the number of Lincoln’s presidency. *Id.*

real estate agents were asked to appraise a house.⁷² They inspected the house and were given all the information usually used in making such appraisals, such as the characteristics of the property, recent sale prices of other houses in the community, and so on.⁷³ However, they were also told a current "listing price" which in some cases was high and in some cases low.⁷⁴ The evaluations by the agents were strongly affected by the listing price anchor.⁷⁵

One of the most dramatic demonstrations of expert vulnerability to anchoring effects is a recent study by Birte English and Thomas Mussweiler.⁷⁶ In that study, a student from a totally unrelated field gave an estimate about how large the solution to a problem should be to an expert faced with deciding the problem.⁷⁷ Some experts received high estimates and some low.⁷⁸ This information, received from a low-credibility source, was still sufficient to create an anchor impacting the estimates made by the experts.⁷⁹ Although most has research involved specifically numeric judgments, fields where the task involves making subjective probability estimates of the magnitude or rarity of certain features, such as many of the forensic sciences, would seem quite likely to be vulnerable to such anchoring effects.

4. *Role Effects*

Quite a different line of research involves the cognitive effects of "role." Role-taking studies call upon a person to adopt a particular function or perspective. The perspective adopted has a variety of effects on the information a person seeks, as well as how the person perceives that information. In one important study, some participants assumed the role of a homebuyer and others the role of a burglar.⁸⁰ They then read a story that contained a description of a house and grounds.⁸¹ Later recollections of the details of the house were quite different, depending upon the role adopted, suggesting that the role influenced participants' attention to details.⁸² This confirms the cocktail party commonplace that a barber is more likely to

72. Gregory B. Northcraft & Margaret A. Neale, *Experts, Amateurs and Real Estate: An Anchoring-and-Adjustment Perspective on Property Pricing Decisions*, 39 *Organizational Behav. and Hum. Decision Processes* 84 (1987).

73. *Id.* at 87.

74. *Id.*

75. *Id.* at 92.

76. Birte English & Thomas Mussweiler, *Sentencing Under Uncertainty: Anchoring Effects in the Courtroom*, 31 *J. Applied Soc. Psychol.* 1535 (July 2001).

77. *Id.* at 1542.

78. *Id.*

79. *Id.* at 1543.

80. James W. Pichert & Richard C. Anderson, *Taking Different Perspectives on a Story*, 69 *J. Educ. Psychol.* 309, 310 (1977).

81. *Id.*

82. *Id.* at 313.

note details of a person's haircut than a dentist, and a dentist more likely to note details about a person's teeth. In addition, similar results were obtained when the role was not assigned until after the description of the house had been provided.⁸³ Such studies demonstrate that role perception also has an impact during the retrieval phase. Given the cognitive effects of role, it is likely that role may also affect decision thresholds. If this is the case, investigators whose role is to solve a problem may become convinced of the truth of a proposed solution more easily than investigators whose role is to describe a situation, or to describe the likelihood of various options. In this regard, the following observation about forensic laboratories by James Starrs, made many years ago, appears to remain true today:

It is quite common to find . . . laboratory facilities and personnel who are, for all intents and purposes, an arm of the prosecution. They analyze material submitted, on all but rare occasions, solely by the prosecution. They testify almost exclusively on behalf of the prosecution. They inevitably become part of the effort to bring an offender to justice. And as a result, their impartiality is replaced by a viewpoint colored brightly with prosecutorial bias.⁸⁴

5. *Conformity Effects*

Research also has revealed "conformity effects," our tendency to conform to the perceptions, beliefs, and behavior of others. Research on conformity shows that people rely on the views of others in order to develop their own conclusions, sometimes to gain additional information, other times merely to be in step with their peers. For example, a classic demonstration of conformity effects used the "autokinetic effect": a stationary point of light in a completely darkened room will appear to be moving.⁸⁵ In the study, people of differing social status or authority were shown such a point of light in one another's presence, and asked to estimate over how far a range it was moving.⁸⁶ Although each person's perceptions of motion range were influenced by the announced perceptions of the others, those of perceived lower rank were more influenced by those of perceived higher rank.⁸⁷

83. Richard C. Anderson, James W. Pichert & Larry L. Shirey, *Effects on a Reader's Schema at Different Points in Time*, 75 *J. Educ. Psychol.* 271, 276 (1983); Richard C. Anderson & James W. Pichert, *Recall of Previously Unrecallable Information Following a Shift in Perspective*, 17 *J. Verbal Learning & Verbal Behav.* 1, 2-3, 10-11 (1978).

84. James E. Starrs, *The Ethical Obligations of the Forensic Scientist in the Criminal Justice System*, 54 *J. Ass'n Official Analytical Chemists* 906, 910 (1971).

85. Muzafir Sherif & Carolyn Sherif, *Social Psychology* 212-13 (1969).

86. *Id.*

87. *Id.*

6. *Experimenter Effects*

The discussion thus far has emphasized the problem of observing inanimate objects. The objects do not change, but the states of the observers do, and as a consequence, the observer's apprehensions, recordings, computations, and interpretations change. Scientists whose objects of study are living organisms face additional problems. Human and animal subjects, unlike inanimate objects, can perceive the experimenter's behavior, which results in the alteration of their own behavior.⁸⁸ The observer communicates something, usually unintentionally, to which the subject responds; what appears to be learned about the subject is actually a reflection of various aspects of the process of observing.

If the problem can be serious with animal subjects, it is all the more serious with that most malleable of animals, the human being. Substantial research has been directed toward understanding the processes by which a researcher's expectancies change her behavior toward different research participants, and how the participants in turn pick up the cues and respond to them with their own changed behavior.⁸⁹ For example, the expectations, and perhaps related enthusiasm, of industrial-organizational psychologists studying the effects of workplace innovations caused workers to perform better than they had without the expectations.⁹⁰ In other words, improvements thought to be due to workplace innovations put in place by the psychologists were instead the result of communicated expectations from the researchers, the increased attention paid to the workers, or to a placebo effect. This same phenomenon can occur in the educational setting.⁹¹ When

88. An interesting example of this is found in the story of "Clever Hans," the famous counting horse of the early twentieth century. Hans's ability to count was eventually shown to be a response to subtle and unintentional cues of those who watched him. *The Clever Hans Phenomenon: Communication with Horses, Whales, Apes, and People* (Thomas A. Sebeok & Robert Rosenthal eds., 1981); Oskar Pfungst, *Clever Hans—The Horse of Mr. Von Osten* (Robert Rosenthal ed., 1965) (Carl L. Rahn trans., 1911). In fact, this phenomenon is often referred to as the "Clever Hans Effect."

89. See, e.g., *Interpersonal Expectations: Theory, Research, and Applications* (Peter David Blanck ed., 1993) [hereinafter *Interpersonal Expectations*]; Monica J. Harris & Robert Rosenthal, *The Mediation of Interpersonal Expectancy Effects: 31 Meta-Analyses*, 97 *Psychol. Bull.* 363 (1985); Robert Rosenthal, *Interpersonal Expectancy Effects: A 30-Year Perspective*, 3 *Current Directions Psychol. Sci.* 176 (1994); Ralph L. Rosnow & Robert Rosenthal, *People Studying People: Artifacts and Ethics in Behavioral Research* (1997).

90. When this phenomenon occurs or is suspected in industrial settings, it is referred to as a "Hawthorne Effect," named after the particular manufacturing plant where it was discovered. F.J. Roethlisberger, *The Elusive Phenomena* 44-48 (1977). The original study which led to the discovery of this effect is F.J. Roethlisberger & William J. Dickson, *Management and the Worker: An Account of a Research Program Conducted by the Western Electric Company, Hawthorne Works, Chicago* (1939).

91. In the educational setting this is termed the "Pygmalion Effect." See generally Robert Rosenthal & Lenore Jacobson, *Pygmalion in the Classroom* (expanded ed.

teachers are told that certain randomly selected students will blossom before the school year ends, it affects the teacher's interactions with those students.⁹² As a consequence, these students show greater improvement relative to the control group.⁹³ One final example of experimenter effects occurs when the belief of medical researchers in the potential effectiveness of an experimental treatment produces improvements in patients even though the treatment itself is worthless.⁹⁴ This placebo effect is likely the most widely known example of experimenter effects. It is due in part to experimenter expectations, which can be controlled by double-blind testing designs, and in part to the patients' own independent expectations, which cannot be controlled by keeping experimenters blind, but require matched placebo designs.⁹⁵

At first blush, it might seem that studies establishing the existence of such "experimenter effects" would have little relevance to forensic scientists, whose objects of study are by and large inanimate. For example, a forensic scientist's expectations cannot actually change the color of a paint chip or the specific gravity of a fragment of glass. The expectations of a forensic scientist, it would seem, can change only her own perceptions. This is undoubtedly true if we think only in terms of the individual examiner working a case involving inanimate paint chips or glass fragments. However, the larger organizational setting of a crime laboratory is analogous to an "experiment," where the police investigators, prosecutors, lab directors, and colleagues in the lab are the "experimenters" (occupying the same role as the managers, teachers, and doctors in the above examples), and the individual forensic examiners are the "subjects" of the experiment. From this perspective, the beliefs and expectancies of superiors, coworkers, and external personnel are manifest in their behavior toward the forensic scientist "subject," in turn affecting the behavior of those "subjects"—their observations, recordings, computations, and interpretations—not to mention the additional impact role and conformity effects may have. Thus, the more complex "experimenter effect" findings indeed appear quite relevant to what happens in the forensic science laboratory, especially in light of the findings of the effects of role and authority previously noted.

1992). The effect is named after the character in the Greek myth, and the play of the same name by George Bernard Shaw.

92. *Id.* at 249-51.

93. *Id.*

94. Rosenthal, *Experimenter Effects*, *supra* note 39, at 367.

95. Robert Rosenthal, *Designing, Analyzing, Interpreting, and Summarizing Placebo Studies*, in *Placebo: Theory, Research, and Mechanisms* (Leonard White, Bernard Tursky & Gary E. Schwartz eds., 1985).

D. *The Pervasiveness of Observer Effects*

Because human perception and judgment are inherently susceptible to influence, it is not surprising that some behavioral scientists have concentrated their research in examining all manner of observer effects. These researchers have gone beyond merely noticing the possibility of observer effects; they have conducted systematic experiments designed to better understand the conditions under which these effects occur and how to tame them.

As discussed above, where errors occur, there is the possibility that they are merely random. Alternatively, the errors might tend to reflect outcomes expected or favored by the observer, however diligently the observer is trying to report and record accurately. Examining this dichotomy, John L. Kennedy and Howard F. Uphoff studied recording errors in an experiment on extrasensory perception ("ESP").⁹⁶ In the study, twenty-eight observers recorded 11,125 attempts to detect what another person was trying to "transmit."⁹⁷ Of these, the observers mis-recorded 126, or 1.13%.⁹⁸ Because 98.87% were accurately recorded, we can infer that the observers were being honest and conscientious. But the errors that did occur were not random; observers who were believers in telepathy made nearly 75% more errors increasing the telepathy scores than observers who did not believe in ESP,⁹⁹ and ESP nonbelievers made 100% more errors decreasing the telepathy scores than did ESP believers.¹⁰⁰ In other words, believers typically erred by recording the presented stimulus and the detection attempt as being the same when in fact they were different, while nonbelievers tended to record them as different when in fact they were the same.

Other studies have been conducted to determine whether errors can be induced by creating expectations in the minds of observers. In one such study, Lucien Cordaro and James R. Ison had observers record the head turns and body contractions of *Planaria* (flatworms).¹⁰¹ A group of identical *Planaria* was randomly divided in half, and for one of those halves, observers were led to expect a high incidence of turning and contracting.¹⁰² For the other half, the observers were led to expect a low incidence.¹⁰³ The result was that observers led to expect a high rate of turns and contractions

96. John L. Kennedy & Howard F. Uphoff, *Experiments on the Nature of Extra-Sensory Perception: The Recording Error Criticism of Extra-Chance Scores*, 3 J. Parapsychol. 226 (1939).

97. *Id.* at 240.

98. *Id.* at 241 tbl.X.

99. *Id.*

100. *Id.*

101. Lucien Cordaro & James R. Ison, *The Psychology of the Scientist: X. Observer Bias in Classical Conditioning of the Planarian*, 13 Psychol. Rep. 787 (1963).

102. *Id.* at 787-88.

103. *Id.*

recorded almost five times as many head turns and twenty times as many body contractions.¹⁰⁴

In another study, one of us (Rosenthal) conducted a meta-analysis of twenty-one studies which checked the accuracy of the observers' recordings of data in those studies.¹⁰⁵ The studies involved a range of subject matter, including reaction times, person perception, human and animal learning, task ability, psychophysical judgments, questionnaire responses, and telepathy.¹⁰⁶ Together, the twenty-one studies involved over 300 observers making and recording about 140,000 observations.¹⁰⁷ Rosenthal's analysis revealed that about 1% of these observations were recorded incorrectly, and about two-thirds of all recording errors favored the hypothesis of the observer.¹⁰⁸

Similarly, L.S. Cahen carried out an experiment in which 256 prospective school teachers were asked to score exam booklets of children supposedly being tested for academic readiness.¹⁰⁹ Each of thirty test items was scored on a four-point scale using a scoring manual which gave examples of answers of varying quality.¹¹⁰ Each test booklet included some "background" information on the child, including her IQ score, to create an expectation in the grader that the child was either above average, average, or below average in intellectual ability.¹¹¹ The examination scorers gave different grades to identical performances, differences that correlated with the exam graders' expectations created by the child's IQ score.¹¹² As previously noted, in apparent recognition of such expectancy bias, many everyday academic test settings have in place procedures to avoid such biasing effects, namely, blind grading.

Consider also the phenomenon of "contrast effect" and "adaptation level," best illustrated by the mundane experience that cool water feels warm to a cold hand, and then will feel cool once the same hand has warmed up. As Donald Campbell has explained:

Whenever human judges are used as a measuring device, their calibration is subject to systematic unconscious alterations, so that the central tendency of the stimulus context to which they are adapted comes to appear as neutral or intermediate, whereas the

104. *Id.* at 788.

105. Robert Rosenthal, *How Often Are Our Numbers Wrong?*, 33 *Am. Psychologist* 1005, 1005-08 (1978).

106. *Id.*

107. *Id.*

108. *Id.* at 1006 tbl.1.

109. L.S. Cahen, *An Experimental Manipulation of the "Halo Effect": A Study of Teacher Bias* (1965) (unpublished manuscript, Stanford University) (on file with authors). The Cahen study is described in Rosenthal, *Experimenter Effects*, *supra* note 39, at 22.

110. Rosenthal, *Experimenter Effects*, *supra* note 39, at 22.

111. *Id.*

112. *Id.*

stimuli that deviate most from this adaptation level appear most striking [the “contrast effect”]. If in the course of judgments the central tendency of the presented stimuli shifts, this produces a shift in judgment standards of which the judge is unaware [the “adaptation level”]. Such effects have been found for every type of stimulus attribute for which they have been examined [citing numerous examples]. Of these studies, the last group are clearly appropriate to the psychology of science, inasmuch as they deal with an arena in which human observers have not yet been replaced by more stable instruments. In every research setting in social or clinical psychology in which raters are employed to record behavior or to code protocols, such effects will be present, and the research must be designed so as to prevent their being confounded with the crucial . . . comparisons.¹¹³

The last sentence could have been written with much of current forensic science practice in mind.

E. The Enhancement of Observer Effects by Desire and Motivation

To this point, we have discussed the problem of observer effects mostly in terms of the impact of mere expectations of what an observation is likely to reveal. There also is an extensive literature on “need-determined perception,” that is, how an emotionally heightened or “hot” motivational state, as distinct from a “cool” cognitive expectation, affects what the observer perceives.¹¹⁴ If even the mildest of expectations can affect perception, then it is not surprising to find that where 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.”¹¹⁵ And to be sure, scientists and their assistants may have strong hopes about what it is that they will “merely observe,” such as in the examples above of Newton, Mendel, and Blondlot.¹¹⁶

113. Donald T. Campbell, *Systematic Errors to be Expected of the Social Scientist on the Basis of a General Psychology of Cognitive Bias*, in *Interpersonal Expectations*, *supra* note 89, at 34-35 (essay originally written in 1959).

114. See, e.g., William N. Dember, *The Psychology of Perception* (1960); Jerome S. Bruner, *Personality Dynamics and the Process of Perceiving*, in *Perception: An Approach to Personality* (Robert R. Blake & Glenn V. Ramsey eds., 1951); Leo Postman, *Toward a General Theory of Cognition*, in John A. Rohrer & Muzafer Sherif, *Social Psychology at the Crossroads* (1951); Leo Postman, *The Experimental Analysis of Motivational Factors in Perception*, in Judson S. Brown et al., *Current Theory and Research in Motivation* (1953); Leo Postman & Jerome S. Bruner, *Perception Under Stress*, 55 *Psychol. Rev.* 314 (1948).

115. The existence of such effects was first clearly established for circumstances of high felt motivation. See the pioneering work of Leo Postman in the 1940s and 1950s, *supra* note 114. The more recent research, including most of the studies recounted in this Article, has been directed toward defining the limits of such effects under cooler and more attenuated influences.

116. See *supra* notes 29, 31, 35 and accompanying text.

Consider the case of Samuel George Morton, a leader of the objective measurement school of nineteenth-century American physical anthropology.¹¹⁷ Morton amassed a huge collection of skulls from all over the world, which he measured to determine if there were racial differences in cranial capacity, and by extension, in intelligence.¹¹⁸ He found significant differences among the races, with Caucasians enjoying the largest cranial capacity.¹¹⁹ In 1977, Stephen Jay Gould recalculated Morton's statistics using Morton's own data, and showed that racial differences Morton claimed to have found did not exist in the data when it was properly analyzed, which Morton had failed to do in a number of ways.¹²⁰ Still Gould concluded, "Yet through all this juggling, I detect no sign of fraud. . . . All I can discern is an a priori conviction about racial ranking so powerful that it directed his tabulations along preestablished lines."¹²¹

Many individuals have attitudes toward what they are observing and harbor a preference for one outcome over another. Other observers, perhaps less committed to the data and more committed to the uses to which an observation will be put, might be even more susceptible to observer effects. Here, research on the effects of the perceived role of the observer becomes relevant once again. Research on need-determined perception shows that in general the world appears different to people who have a desire to see it in different ways, and how different the world appears is related to the intensity of that desire.

In this regard, consider the following quotation from James Corby of the FBI Materials Analysis Unit, who performed the paint match that was one of the central pieces of evidence resulting in the 1987 conviction of Frank Jarvis Atwood for the abduction and brutal rape/murder of eight-year-old Vicki Lynn Hoskinson:

Usually you have no association with the victim or the family, and you work so many of these cases that you try not to get involved, but it's very difficult when a crime involves a baby or a small child, somebody that's defenseless, and you find yourself, I think, working harder to try to establish something in a case. But if it's

117. The whole story of Morton's craniometry is recounted in detail in Stephen Jay Gould, *The Mismeasure of Man*, 50-69 (1978).

118. *Id.* at 53.

119. *Id.*

120. *Id.* at 54. A full exposition of the technical details may be found in Stephen Jay Gould, *Morton's Rankings of Races by Cranial Capacity*, 200 *Sci.* 503 (1978).

121. Gould, *supra* note 117, at 69. Gould also observes that:

Conscious fraud . . . tells us little about the nature of scientific activity. Liars, if discovered, are excommunicated; scientists declare that their profession has properly policed itself, and they return to work, mythology unimpaired, and objectivity vindicated. The prevalence of *unconscious* finagling, on the other hand, suggests a general conclusion about the social context of science. For if scientists can be honestly self-deluded to Morton's extent, then prior prejudice may be found anywhere, even in the basics of measuring bones and toting sums.

Id. at 54-56.

not there, it's not there, but you certainly, I think, take a more critical look at that case, and I think it's human nature.¹²²

The examples above reflect that observer effects may occur at any of several stages of observation, from the initial observation to the conclusions drawn about what was observed. The errors at each of these stages may be described as follows:

- Errors of Apprehending (errors that occur at the stage of initial perception);
- Errors of Recording (errors that creep in at the stage where what is observed is recorded, assuming a record beyond memory is even made);
- Errors of Memory (errors that are induced by both desires and the need for schematic consistency, and that escalate over time when memory is relied on);
- Errors of Computation (errors that occur when correct observations accurately recorded or remembered are transformed into incorrect results when calculations are performed on them); and
- Errors of Interpretation (errors that occur when examiners draw incorrect conclusions from the data).

In the case of errors of interpretation, the criteria for the “true” values of the underlying observations are often so vague, ephemeral, and submerged in the interpretation, that one often cannot discover the inaccuracy in the interpretative conclusion. Interestingly, this most error-prone circumstance corresponds to the realm of the expert testifying in a legal proceeding: the expert’s “opinion.” It is exactly where stimuli are most on the border of accurate perception and classification that conditions most favor errors of interpretation. The more ambiguous and ill-defined the stimulus and the more frustrated or motivated the observer, the more likely one or more observer effects will occur, resulting in an inaccurate result.¹²³

F. The Lack of Linkage Among Confidence, Accuracy, and Amount of Information

Generally, the more information people have, including experts, the more confident they are in their decisions. Accuracy, however, does not always increase as a function of confidence. For example, Paul Slovic studied horse-race handicappers.¹²⁴ He asked them to predict the winner and state their confidence in the prediction. As they obtained more and more information about the horse and rider, their confidence in their prediction

122. *FBI Files: The Predator* (Discovery Channel television broadcast, Nov. 29, 2000) (tape on file with authors).

123. Campbell, *supra* note 113, at 38.

124. Paul Slovic, *Toward Understanding and Improving Decisions*, in 2 *Human Performance and Productivity, Information Processing and Decision Making* 168 (W.C. Howell & E.A. Fleishman eds., 1982).

kept increasing, yet their accuracy remained unchanged.¹²⁵ Apparently, this result occurred because the new and accurate information affected outcome probabilities less than the experienced and motivated experts operationally believed it that it did.

The lack of relationship between substantial additions of information and accuracy of result under some conditions, the direct relationship between such information and confidence in one's conclusions, and the resultant lack of relationship between confidence in one's conclusion and actual accuracy, is especially troublesome in any field where subjective probability estimates are the primary conclusion. As previously noted, many traditional forensic science fields, most particularly "identification disciplines" such as toolmark, bitemark, or handwriting analysis, rely on such subjective probability estimates. Information can expand and subjective probability will go up, but the accuracy—the objective probability—may not. Indeed, if new information is sufficiently overvalued, confidence could go up while accuracy goes down.

II

Observer Effects in Forensic Science

The findings and concepts described above are no less relevant to forensic science practice than they are to physics, biomedicine, and the behavioral and social sciences. In their daily work, forensic scientists are observers of a wide variety of objects, shapes, colors, instrumentation, and test results. The observations that must be made present varying degrees of ambiguity. Subjective judgment and interpretation by the human observer remain the principal methods of reaching conclusions in most forensic disciplines, and the working environment of the forensic scientist is not lacking in sources of expectations or outcome preferences. Such circumstances facilitate the operation of observer effects, particularly when observers have armed themselves so lightly against the infiltration of distorting influences. In what follows, we explore more concretely the environment of the forensic scientist and the observer effects that are likely to impinge on a forensic examination.

A. Proper and Improper Information in the Forensic Science Practice

The proper function of a forensic scientist is to give an answer to a question appropriate to her discipline by the application of the methods of her discipline. It is not to give an answer, even an honest and accurate answer, to that same question by any other means. This may be an especially difficult distinction for forensic scientists who were drawn to their

125. *Id.*

work through an interest in law enforcement, or who began their careers as regular law enforcement officers, but it is fundamental.¹²⁶

A detective's role is to gather and consider all information in an effort to determine the material facts of a case. It is no criticism of a detective if she considers any information, even weak or undependable information, in conducting her investigation. Indeed, the very notion of an "investigative lead" involves information that is weak and often leads nowhere. When all else fails, even the employment of nonrational sources, such as psychics,¹²⁷ cannot be said to fundamentally violate the detective's role and function, although most might view them as a waste of time, money, and effort.¹²⁸ Such exercises may precipitate a change in focus leading to the discovery of more dependable information that was previously overlooked, even if the exercise is itself without rational content. All this is true because, in the end, the detective's conclusions about the material issues of the case must be backed up by legally admissible evidence, and that evidence must convince prosecutors to prosecute, judges to send the case to a jury, and a jury to convict. Most importantly, however, the detective herself is not allowed to testify concerning her conclusions. No doubt a detective's "solution" to a case is often subject to all sorts of observer effects, but the system has been built in such a way that the ultimate factfinders are insulated to a great degree from the results of those effects on the detective.

A forensic scientist is in a very different situation. A forensic scientist is not a detective. We repeat, popular television shows to the contrary notwithstanding, a forensic scientist is not a detective.¹²⁹ The conclusions of the forensic scientists are put before the jury. The reason the products of the forensic scientist's efforts are admissible is not because forensic scientists are better at drawing conclusions about the meaning of normal relevant evidentiary information than detectives or jurors; it is because the law has accepted that, as to a defined area of specialized knowledge or skill, the products of their practice are better than the jury could do alone.¹³⁰ When the forensic scientist is exposed to, relies on, or is influenced by any infor-

126. It appears that the bulk of forensic science examiners began their careers as law enforcement officers. This has certainly been generally true at the FBI laboratory until quite recently. See Inspector General's Report, *supra* note 61, at 9.

127. See, e.g., April Goodwin, *Team Set to Identify 1912 Villisca Killer*, Des Moines Register, June 6, 2000, at 1, available at 2000 WL 4961236; *I Chat with Princess Diana All the Time. She Insists the Crash Was Just an Accident*, The Express (London), Sept. 16, 2000, available at 2000 WL 24216528.

128. For a discussion including views critical of such use of psychics, see Richard N. Kocsi et al., *Expertise in Psychological Profiling: A Comparative Assessment*, 15 J. Interpersonal Violence, Mar. 1, 2000, available at 2000 WL 11328497.

129. Apropos of roles and self-images, it is perhaps worth noting here that the leading newsletter for forensic scientists is called *Scientific Sleuthing Review*.

130. D. Michael Risinger, Mark P. Denbeaux & Michael J. Saks, *Exorcism of Ignorance as a Proxy for a Rational Knowledge: The Lessons of Handwriting Identification "Expertise,"* 137 U. Pa. L. Rev. 731, 734-35 (1989).

mation outside of her own domain, she is abusing her warrant, even though she may honestly believe that such information makes her conclusion more reliable, and even, or especially, if she is right about this. Her role is not to give a conclusion based even partly on information outside her domain, which the jury can presumptively evaluate at least as well as she, but only to give the jury the reliable product of her discipline that is beyond what they could deduce on their own.

The dangers of the practice of relying on extradomain information are easily illustrated. Assume a forensic odontologist examines a bitemark on human skin. Assume further that, due to the incomplete detail of the bitemark, the odontologist would conclude that the bitemark was probably from a human adult, but that there was insufficient detail to identify any particular adult as the source of the bite. However, suppose the odontologist is told, and takes into account, that the complaining witness said she was raped by a man who also bit her, and that the DNA analysis of the sperm recovered from her identifies the defendant to a very strong random match probability. In this situation, the odontologist is fully justified in concluding that the source of the bitemark can be assigned to the defendant to a very high degree of probability, but not as a result of his expertise in forensic odontology. On one level, there is a certain apparent backwardness to his conclusion, since he is using information about the identity of the attacker to draw a conclusion about the source of the bitemark, instead of providing a conclusion about the source of the bitemark to be used as a basis for inferring the identity of the attacker.¹³¹

It is not directionality that is the problem, however, but confusion of role. If one is authorized by one's role to consider all probability-affecting factors on the issue under examination, the order of consideration of those factors is irrelevant. If the odontologist's role were to offer a conclusion about the identity of the source of the bitemark using all available information, then his conclusion would be beyond criticism. However, that is not his role. His warrant is only to provide the information derivable from his discipline alone for the jury's use. Under that warrant, if he testifies to the conclusion as described, he would be appearing to provide the jury with new and meaningful information, while in fact all he would be doing was repeating, in a disguised fashion, other information they already had. This approach results in a double impact being given to the domain-extraneous information.

The result can be an investigative echo chamber, where a few items of evidence reverberate and seem more numerous and stronger than they really are. A simple mathematical illustration demonstrates the power of

131. Indeed, this process was characterized as working "backwards" and condemned on that basis in the Department of Justice Inspector General's report on the FBI laboratory's practices. Inspector General's Report, *supra* note 61, at 104.

the effects of such contamination. Suppose I have seven items of evidence associating a suspect with a crime scene. Further, assume that each of these items of evidence is independent of the others, and each carries with it a random match probability of only .50 (that is, a person selected at random from the population would have a 50% chance of possessing the same attribute). Because I have seven items of evidence which connect the suspect to the crime, they combine to be far more compelling identification than any one of them alone: .50 raised to the seventh power equals .0078—only 78 chances in 10,000 that a person selected at random from the population would be incriminated to the same extent as this suspect. Human intuition corresponds at least roughly with the mathematics, and we have the feeling that as each item of evidence accumulates, the likelihood of erroneously convicting an innocent suspect gets smaller and smaller.

Now suppose there is only one independent item of evidence, and that all the others are the products of cross-contamination, thus: As a consequence of an interview with a suspect, a detective comes to have a hunch that the suspect is the perpetrator. Affected by that officer's hunch, a handwriting examiner concludes that the ambiguous handwriting evidence identifies the suspect as the perpetrator (an example of conformity effect, possible role effect, and confirmation bias affecting the decision threshold). An officer conducting a lineup knows of his colleague's hunch and the document examiner's match, and inadvertently steers an eyewitness to select the suspect from a lineup when, had the witness been out of the presence of the officer, a positive identification would not have been made (an example of conformity to authority and the Clever Hans effect). The bite-mark expert learns of those three items of evidence and is influenced by them when making his positive identification. Before the toolmark expert completes his examination, he knows of four other items of incriminating evidence, and so on. However, if each of these examples is largely an echo of the initial item, then the true random-match probability may be closer to .50 than to .0078 because each new incriminating item may be little more than a reflection of that initial piece of evidence.

The result is little more than an illusory consensus, but it is nevertheless a potent one. Although each expert reached the contaminated conclusion in the shadow of expectations of what the "correct" outcome was, the knowledge of the other "corroborating" conclusions reinforced the subjective confidence each had in the accuracy of his own result, even though it added nothing to the accuracy of the finding. This in turn yields powerful, confident testimony from each expert as a witness. Borrowing from a more purely scientific context:

Insofar as systematic biases have been observed, they are overwhelmingly . . . a tendency to contaminate one's reports in the direction of agreement with what others are reporting and thus fail

to report what is uniquely available from one's own perspective. In addition, the agreement achieved represents pseudo-confirmation. The tremendous literature on conformity and suggestion shows how strong and persuasive this effect is. It could scarcely fail to operate among teams of scientists.¹³²

The resulting harm can be viewed from two perspectives. If the evidence is sound and would have stood up if independently evaluated, then the informational cross-contamination undermines the true value of the evidence. Conversely, if the evidence is unsound, then the informational contamination can create guilt out of next to nothing.

B. Improper Information Contamination in Forensic Science Practice

The general principle that forensic examiners should be insulated from all information about an inquiry except necessary, domain-specific information is not novel. For instance, regarding handwriting identification, William Hagan wrote in 1894: "The examiner must depend wholly upon 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"¹³³ However, to the extent information on current practices is available, in the forensic science disciplines this fundamental principle is usually ignored. The neglect of this principle is clear from the following evidence.

First, the principle is reflected in few if any modern textbooks for forensic examiners.¹³⁴ Second, the accreditation standards of the American Society of Crime Laboratory Directors ("ASCLD") do not reflect the principle or require any specific or general review to insure that it is followed.¹³⁵ The standards do contain a section on "controlling" that sets out the subjects to be covered by a required "quality manual,"¹³⁶ but none addresses the problem of controlling domain-irrelevant information. In addition, while the standards provide for a "quality manager,"¹³⁷ none of the

132. Campbell, *supra* note 113, at 37.

133. William E. Hagan, *Disputed Handwriting* 82 (1894).

134. The most widely used introductory textbook on forensic science is Richard Saferstein, *Criminalistics: An Introduction to Forensic Science* (7th ed. 2001). This text reviews the major areas of forensic science but it includes neither mention of observer effects or examiner bias, nor discussion of blind testing or the problem of basing conclusions on collateral information not specific to the scope of the expertise being applied. Other texts with similar lacunae include Ordway Hilton, *Scientific Examination of Questioned Documents* (rev. ed. 1982); Brian J. Heard, *Handbook of Firearms and Ballistics: Examining and Interpreting Forensic Evidence* (1997); Claude W. Cook, *A Practical Guide to the Basics of Physical Evidence* (1984).

135. See American Society of Crime Laboratory Directors, *Laboratory Accreditation Board Manual* (2000) [hereinafter ASCLD Manual].

136. *Id.* § 1.4.2.1.

137. *Id.* § 1.4.2.2.

manager's described functions, including administration of a "quality audit,"¹³⁸ touches on unnecessary biasing information. This is especially significant since the ASCLD Manual was substantially revised in 2000 after the publication of the Department of Justice Inspector General's Report on FBI laboratory practices, yet it still did not address the problems of biasing information revealed in that report.¹³⁹ The standards pay significant attention to preventing contamination of evidence and none to preventing contamination of examiners.¹⁴⁰

Third, there have been no formal studies on the actual practices in forensic science laboratories that would document the statistical incidence of the use of domain-extraneous information. Responsibility for the absence of such studies can only be placed on the forensic science community itself, since no one else is in a position to conduct such studies.

Finally, the anecdotal evidence is extensive and uniform in indicating that extraneous information is rife in most, if not, all areas of forensic practice. Consider, for instance, the following examples:

To start with an example close to home for one of the authors (Risinger), the forms for the submission of evidence for examination at the New Jersey State Police Laboratory have a section marked "Brief History of Case." The submitting local agency can and generally does use this section to include whatever case information it wishes. This form is passed along to the examiner with the evidence.¹⁴¹

There is no reason to believe this practice, or similar ones, is confined to New Jersey, and substantial reason to believe it is normal general practice. For instance, Smithsonian forensic anthropologist Douglas Ubelaker recounts receiving numerous letters of transmittal on submissions forwarded to him in his role as a consultant to the FBI which routinely included extensive case information. He adopted the practice of reading nothing except the bare minimum necessary to log in the specimen,¹⁴² at least before doing his initial examination, and lists "being influenced by someone else's expectations" as one of the three biggest dangers in forensic practice.¹⁴³

In an effort to make up for the lack of formal studies on the incidence of examination-irrelevant biasing information in letters of transmittal, one of the authors (Saks) contacted the director of one ASCLD-certified labora-

138. *Id.* § 1.4.2.3.

139. *See* Inspector General's Report, *supra* note 61, at 11, 24, 104.

140. *See* ASCLD Manual, *supra* note 135, § 1.4.1.4. It should be noted that the principle of blind testing to prevent artifacts makes its first appearance in the ASCLD standards with the insertion of a new provision providing for blind proficiency testing. *See id.* § 1.4.3.5.

141. Form on file with authors.

142. Douglas Ubelaker & Henry Scammell, *Bones: A Forensic Detective's Casebook* 86-87, 228 (1992).

143. *Id.* at 279.

tory and a supervisor in another ASCLD-certified laboratory. He inquired about the practice of allowing submitting agencies to include whatever information they wished in their submission documents and the practice of passing those documents to the testing examiner along with the evidence to be tested. Both confirmed that, to the best of their knowledge, the practice was virtually universal.¹⁴⁴

Even the design of proficiency tests can reveal a lack of sensitivity to domain irrelevant information and its effects. For example, in an apparent effort to be realistic, the 1988 Forensic Sciences Foundation proficiency test for document examiners contained much domain-irrelevant information that was explicitly relied on by at least one of its respondents, and potentially by all of them.¹⁴⁵

Perhaps the single most egregious miscarriage described in the Inspector General's investigation of the FBI laboratory was the testimony in the World Trade Center bombing case by David Williams of the FBI explosives unit. Mr. Williams identified the main charge as a urea nitrate bomb, not based upon residues found at the scene, but "on speculation based on evidence linking the defendants to that explosive."¹⁴⁶ As the report goes on to say:

Williams portrayed himself as a scientist and rendered opinions as an explosives expert. As such, he should have limited himself to conclusions that logically followed from the underlying data and the scientific analyses performed He should not have based his opinions, in whole or in part, on evidence that was collateral to his scientific examinations, even if that evidence was somehow connected to the defendants By basing his urea nitrate opinion on the collateral evidence, Williams implicitly accepted as a premise the prosecution's theory of guilt. This was improper.¹⁴⁷

In a similar vein, in the Psinakis case, which involved a claim that the defendant produced a large amount of explosive by stripping it out of detonating cord,¹⁴⁸ examiner Terry Rudolf

Acknowledged that his identification of PETN on the tools was based in part on the fact that stripped detonating cord was found in the defendant's garbage. In his interview with the OIG, Rudolf observed that given this information, he presumed the material on

144. Identities and documentation of responses on file with authors. It should be noted that, in some laboratories, in regard to certain high-volume, routine, and highly instrumented examinations such as toxicological applications of gas chromatography, the examiner may not ordinarily be privy to biasing information in the transmission documents, not as a matter of policy, but as a practical result of the internal arrangements of the particular laboratory.

145. See D. Michael Risinger & Michael J. Saks, *Science and Nonscience in the Courts: Daubert Meets Handwriting Identification Expertise*, 82 Iowa L. Rev. 21, 53-55 (1996).

146. Inspector General's Report, *supra* note 61, at 11.

147. *Id.* at 128-29.

148. *Id.* at 24.

the knife was PETN Rudolph failed to distinguish between the separate and distinct roles of an investigator and a forensic scientist.¹⁴⁹

Examples could be multiplied, and indeed many examples given both previously and below in particular contexts could as easily have been set out here. As to the FBI in particular, it is clear from the Inspector General's Report as a whole that domain-irrelevant information was routinely available to examiners in the FBI laboratory at the time of that report in 1997. The Inspector General claimed to be hopeful that ASCLD accreditation and ASCLD quality control standards would solve whatever problems the report had identified.¹⁵⁰ But, as previously noted, this seems unlikely, since the revised ASCLD standards do not address these problems.

Indeed, some forensic scientists actively promote reaching into improper domains for assistance in making the determinations they are called upon to make in their own proper domain. For instance, the recently published Document Examiner Textbook advises: "Before an attempt by the examiner to identify a handwriting, the investigator should consult and [obtain] as much circumstantial evidence as possible about the case."¹⁵¹

None of this is to say that it is always an easy or trivial exercise to formulate standards of domain-relevant information. To give an example from forensic anthropology, the appearance of damage done to bones by plant roots growing through fissures may be indistinguishable from that

149. *Id.* at 30.

150. *Id.* at 481-85.

151. J.E. Dines, Document Examiner Textbook (1998), *quoted in* Duayne J. Dillon, Book Review, 22 *Sci. Sleuthing Rev.* 4, 5 (1998). The reviewer comments that this advice is "diametrically opposed to proper professional practice In a genuinely scientific approach, to insure maximum objectivity, an examiner should focus on the documents in the case and avoid extraneous proffered information concerning details of the matter in dispute." *Id.* at 5. Mr. Dines is by no means alone, however. For example, in *Truth and Deception*, John E. Reid and Fred E. Inbau cite with approval and reproduce as an appendix an article recounting a study of the effect of polygraph examiners reading the whole case file before interpreting results. John E. Reid & Fred E. Inbau, *Truth and Deception: The Polygraph ("Lie Detector") Technique* 304, app. A-4 (2d ed. 1977). The authors recommend the practice because when polygraph examiners "consider collateral factors such as we have discussed, they can not only increase their diagnosis accuracy but also decrease the number of indefinite reports." *Id.* at 406.

Finally, forensic pathologists apparently consider absolutely any information to be within their domain. Consider the following passage from Dominick J. Di Maio & Vincent J. M. Di Maio, *Forensic Pathology* (1989):

How does a medical examiner (forensic pathologist) approach a case? He or she approaches it just like any other physician approaches a patient. In medical school, one is taught that to make a correct diagnosis, one has to take a history, perform a physical examination, and order relevant laboratory tests. Based on this, a diagnosis is made. The forensic pathologist performs all these functions but with some variance. Thus, the history is not obtained from the patient, but is an account of the events leading up to and surrounding the death obtained from witnesses, relatives of the deceased, police agencies, treating physicians, and/or records (medical, nonmedical, police, governmental, etc.).

Id. at 16.

caused by other sources of trauma.¹⁵² Information concerning the existence or absence of roots among the bones at their discovery site would thus be domain-relevant to a forensic anthropologist but not necessarily to a toxicologist. We will return to this problem below, but for now it is enough to say that such difficulties hardly justify the complete absence of published or analyzed standards that now exists.

In the past, there has been little motivation to develop standards of domain-relevant information. Now, however, there is some indication that, just as *Daubert v. Merrell Dow Pharmaceuticals*¹⁵³ has driven some efforts to develop validity data, at least under test conditions, for some forensic science disciplines,¹⁵⁴ *Kumho Tire Co. v. Carmichael*¹⁵⁵ may be spurring courts to become more sensitive to the distortions of suggestion and other sources of observer effects. Note, for instance, the court's critical observations concerning the suggestive way exemplars were presented to the forensic examiner in *U.S. v. Rutherford*.¹⁵⁶ If this is so, then the development by practitioners in each forensic specialty of appropriate and defensible standards for distinguishing domain-specific from domain-irrelevant information, coupled with mechanisms for screening the latter, may become a necessary precondition to admissibility.¹⁵⁷

C. *Specific Sources of Induced Observer Error in Forensic Science Practice*

1. *Direct Communication Between Investigators and Examiners*

As previously noted, transmittal letters which accompany a submission to a crime laboratory often communicate more about the case than is required to perform the necessary examinations. This information sometimes tells examiners about other inculpatory evidence that has been found in the case, and may include what the investigator making the submission expects or hopes the requested tests will conclude.

As an illustration, consider the cases of Christopher Boots¹⁵⁸ and Eric Proctor.¹⁵⁹ The two were indicted in 1986 for the 1983 murder of a convenience store clerk who was shot to death in the store cooler.¹⁶⁰ Boots called the police to report discovering the murder, and was present with Proctor

152. Ubelaker & Scammell, *supra* note 142, at 107.

153. 509 U.S. 579 (1993).

154. See, e.g., Moshe Kam, Gabriel Fielding & Robert Conn, *Writer Identification by Professional Document Examiners*, 42 J. Forensic Sci. 778 (1997).

155. 526 U.S. 137 (1999).

156. 104 F. Supp. 2d 1190 (D. Neb. 2000). *Rutherford* is discussed more fully at *infra* note 236.

157. See *infra* § 4.

158. *State v. Boots*, 767 P.2d 450 (Or. App. 1989), *rev'd*, 780 P.2d 725 (Or. 1989).

159. *State v. Proctor*, 767 P.2d 453 (Or. App. 1989).

160. *Boots*, 767 P.2d at 451.

when the police arrived.¹⁶¹ The following letter accompanied evidence submitted to the FBI laboratory by the California authorities:

As per our phone conversation of March 6, 1986 I am submitting the partially burned flakes of double base powder out of our Oliver homicide.

This is a murder case that took place in June 1983. The killer or killers entered a local 7-11 store in the late evening hours and forced the young male clerk into the back room (cooler) and broke a full 10 ounce bottle of Orange Crush over his head and then shot him in the head three times with a .22 caliber weapon (probably a Hi-Standard revolver.) Due to some interagency problems the case to date has not been prosecuted, but will be soon.

Going through the trace evidence, some of which had been analyzed by SEM-EDAX, I found a partially burned double base powder flake on one of the planchets. The flake was originally found on the trousers of one of our suspects. We want, if possible for you or Ed to compare this flake (B) to some partially burned flakes (A) found on the body of our victim. The only difference between the treatment of the flakes is that flake B has been carbon coated to prepare it for SEM work.

Exhibits:

Both A and B are sandwiched between the glass slides and clearly circled and labeled. (I have tried to get them to move by tapping the slide but they appear to be stationary.)

(Sample A) Several partially burned flakes of double base powder from the victim.

(Sample B) One piece of partially burned flake of double base powder from the trousers of a suspect.

Request:

If possible, please compare A to B.

Time is of the essence now because of a lawsuit one of the suspects is bringing against the police department for false arrest.

I would appreciate any help you can give. Thank you very much.
/S/¹⁶²

The resulting laboratory report was incriminatory. Boots and Proctor were both convicted, and were imprisoned for eight years until the identity of the real killer was established by independent evidence and they were finally released.¹⁶³

161. *Id.*

162. Letter from Charles H. Vaughan, Lieutenant, Crime Laboratory Director, Oregon Department of State Police to Terry Rudolph (Mar. 7, 1986) (on file with author).

163. Kelly & Wearne, *supra* note 44, at 94. Boots and Proctor later settled a civil action against the city of Springfield, Oregon and two of the investigating officers for two million dollars. *Men Wrongly Imprisoned Settle for \$2 Million*, *Seattle Times*, May 8, 1998, at B2.

Earlier, we pointed out that even what is often referred to as the “gold standard” of forensic science, DNA testing, can present substantial problems of ambiguity in reading and interpreting results under some conditions, especially with specimens that might contain DNA from more than one person.¹⁶⁴ Consider then what might be the effect in such a case of the following information found in the DNA laboratory notes in a California case, documenting a phone call from a Detective Miller: “Suspect—known crip gang member—keeps ‘skating’ on charges—never serves time—this robbery—he gets hit in head w/ barstool—left blood trail. Miller wants to connect this guy with the scene—DNA—if blood on swabs.”¹⁶⁵

On occasion, examiners may become more intensely involved with investigators, with calls or visits back and forth as the evidence in a case develops. The following account of an interaction suggests how intense such communications between investigator and forensic scientist can become:

Over the next several days, Dabbs [the forensic scientist] found herself talking to Horgas [the investigator] numerous times during the day, sometimes three to five calls per hour.

“I had lots of questions about different pieces of evidence,” she said later. “Anytime I called he was always available, and if he wasn’t available he would call me right back. He was trying to help me and of course I was trying to help him.”

.....

Like most scientists, Dabbs prided herself on objectivity. Her role was simply to analyze specimens and write up results. Whether her efforts resulted in an arrest was entirely incidental to her task. With each passing day, however, she found herself becoming more deeply involved with the progress of the Tucker investigation. Her daily conversations with Horgas routinely went beyond a particular lab-related inquiry, and she found herself asking him how his leads were developing, whether he had received any responses to his teletypes, and so on.¹⁶⁶

Such cases are by no means uncommon. Take, for instance, the recent New Jersey case of *State v. Fortin*.¹⁶⁷ Steven Fortin was charged with the August 11, 1994, murder of Melissa Padilla. One piece of evidence against Fortin was testimony by forensic odontologist Dr. Lowell Levine that cer-

164. See *supra* note 17.

165. Note of Mar. 4, 1996, phone call entered in “examination results” section of Orange County (California) Sheriff Coroner’s Department, Request for Evidence Examination Form, Case 96-01-0445 (originally filed Jan. 9, 1996, in exam later conducted July 15, 1996, or thereafter) (on file with authors).

166. Paul Mones, *Stalking Justice* 137 (1995).

167. The case was tried between November 2000 and April 2001. The case generated two reported opinions prior to trial: 724 A.2d 818 (N.J. Super. Ct. App. Div. 1999) and 745 A.2d 509 (N.J. 2000). It appears likely to generate more.

tain marks on Padilla's left breast and chin were bitemarks, and that Fortin's teeth had clearly produced the marks on the breast. The defense expert, the equally credentialed Dr. Norman Sperber,¹⁶⁸ asserted that it was unclear whether any of the marks were bitemarks, and that if they were, they clearly did not match Fortin's dentition. In his testimony Sperber called Levine's conclusions "totally inaccurate."¹⁶⁹ In closing the prosecutor called Sperber a liar.¹⁷⁰

What is important for this Article is not who was correct, but the process which led to Dr. Levine's conclusions. Along with specifically domain-relevant information, such as the photographs of the wounds and the casts of Fortin's teeth, Dr. Levine was provided with a report containing information irrelevant to his particular claimed expertise that suggested Fortin's likely guilt. Perhaps even worse, investigators traveled to Dr. Levine's office on April 19, 1995 and apparently sat with him discussing the case while he did his preliminary comparisons and gave his initial conclusions.¹⁷¹ Such circumstances can undermine the reliability of conclusions even when they are rendered with the purest of conscious intent.

Where this can lead should be apparent. Evan Hodge wrote the following in regard to the "wrong Colt" episode discussed earlier:

[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.¹⁷²

2. *Revision of Findings in Light of New Test-Irrelevant Information*

Examiners can also be influenced by learning of findings regarding other evidence in the case that are inconsistent with their own conclusions. Occasionally, upon learning such information, examiners will change their initial conclusions. We are not concerned here with the examiner who, in light of the other findings, deliberately alters her own opinion to achieve a false consistency. That is the perpetration of an intentional fraud on the justice system, and there are appropriate ways with which such falsification

168. Both Levine and Sperber had decades of experience, both had law enforcement positions, Levine had been president of the American Board of Forensic Odontology ("ABFO"), and Sperber had been chair of the ABFO Committee on Standards. All such details are drawn from the testimony of Levine, Trial Transcript, State v. Fortin, No. 1197-09-05 (N.J. Super. Ct. Law Div. Nov. 15, 2000) [hereinafter Trial Transcript] (on file with authors), and Sperber, Trial Transcript (Nov. 30, 2000).

169. Trial Transcript, *supra* note 168, at 45 (Nov. 30, 2000).

170. *Id.* at 103 (Dec. 5, 2000).

171. Aff. & Request for Search Warrant, Detective Gerard Madden (on file with authors).

172. Hodge, *supra* note 46, at 292. Hodge refers to the "wrong Colt" case as "just one of the many we have seen over the years." *Id.*

should be dealt. Of greater interest for the present Article is the examiner who, upon learning of the contrary findings of other case evidence, begins to rethink and re-perceive and reinterpret his own findings, coming to sincerely believe his revised conclusion.

The prosecution of Bruno Richard Hauptmann for the kidnap and murder of the son of Charles A. Lindbergh may provide an example. Albert O. Osborn and his son Albert D. Osborn were among the leading handwriting identification experts of the time. According to a report of FBI special agent Thomas Sisk, quoted by Ludovic Kennedy in his book *The Airman and the Carpenter*, Albert D. Osborn initially doubted that Bruno Hauptmann's writings came from the same source as the ransom notes.¹⁷³ However, Kennedy recounts, within an hour after having been informed of the discovery of the bulk of the ransom money in Hauptmann's garage, both Osborns came to the conclusion that Hauptmann did in fact write the ransom notes.¹⁷⁴

It is clear from the Inspector General's investigation of practices at the FBI crime laboratory that the FBI laboratory contemplates that examiners in at least some units will know of the findings of other examiners, and that they will meet to arrive at resolution in case of conflicting results.¹⁷⁵ The Inspector General's Report even seems to approve of this practice.¹⁷⁶ However, the inherent dangers of such a practice should by now be readily apparent. Any process for refining inquiry after the return of apparently conflicting findings by different examiners must be much more sensitive to observer effects than appears to be the case at present.¹⁷⁷

3. *Selective Re-examination of Evidence*

Sometimes police or prosecutors respond to test results that are negative or inconclusive by suggesting to forensic scientists what they should have found and asking them to test again in hopes of obtaining a "better"

173. Ludovic Kennedy, *The Airman and the Carpenter* 178-83 (1985).

174. *Id.*

175. See Inspector General's Report, *supra* note 61, at 491.

176. *Id.*

177. A special case of mandated cross-communication involves the "peer review" process followed in most forensic laboratories and mandated by ASCLD standard 1.4.2.16. While it is true that "peer review" is thus common, it is unclear what it is supposed to accomplish. The ASCLD standard indicates that the purpose of a laboratory's peer review process is "to ensure that the conclusions of its examiners are reasonable and within the constraints of scientific knowledge." ASCLD Manual, *supra* note 135, § 1.4.2.16. Regardless of whether "peer reviews" are exposed to the contaminating information that an initial examiner was exposed to, the reviewing examiner typically knows the conclusions of the initial examiner, itself a strong form of contamination. If the peer reviewer serves merely as a process check on the procedures used, making sure the report adequately documents and explains its findings and conclusions, then the fact that the reviewer knows the outcome is arguably necessary. But no one should have any illusions about such a peer review being much of an independent confirmation of the initial conclusions' correctness in the event that normal practice has been followed.

result. The contamination here can be quite crude; the investigator or prosecutor might be signaling to the examiner that a more inculpatory result is desired and inviting the examiner to rethink the conclusions with that in mind. For example, Peter DeForest has described investigators who responded to inconclusive results by saying to forensic examiners: “Would it help if I told you we know he’s the guy who did it?”¹⁷⁸

On a less crude level, a man being charged with the murder of a police officer claimed that the officer was beating him and that he took the officer’s gun in self-defense and shot the officer.¹⁷⁹ A state medical examiner concluded that the entry wound was in the officer’s back, which was inconsistent with the defendant’s claim of self-defense, but consistent with the prosecution theory that the officer was shot while he was trying to run from the defendant.¹⁸⁰ FBI examiners, on the other hand, concluded that the entry wound was in the officer’s chest, which was consistent with the shooter’s self-defense claim.¹⁸¹ In an effort to resolve the discrepancy, the district attorney contacted the FBI, pointed out that the state examiner had reached the opposite conclusion, and asked them to double-check their findings, just to make sure they were correct.¹⁸² The mere choice of whom to call and ask to conduct a re-examination skews the results. It leaves the preferred set of conclusions in place while inviting revision of the non-preferred conclusion.¹⁸³ In the end, the federal examiner switched to the state examiner’s conclusion, asserting that he had actually reached that conclusion, but had misrecorded it.¹⁸⁴

178. Peter DeForest, Address at the 2d International Conference on Forensic Document Examination (June 14-18, 1999) (notes of Michael Saks, who was present). The quoted statement can easily be interpreted as an invitation to fraud, and if so interpreted, is not within the principal focus of this Article. We are interested in the more subtle bias created by nonfraudulent but selective re-examination.

179. National Inst. Just., National Conference on Science and the Law: Proceedings 228 (2000) (statement of E. Michael McCann, Milwaukee County District Attorney).

180. *Id.*

181. *Id.*

182. *Id.*

183. This is an aspect of a more general strategy sometimes referred to as “cherry-picking.” Cherry-picking generally refers to any process in which numerous tests or evaluations are performed, often without the knowledge of any given evaluator that there are multiple evaluations being sought. These evaluations, predictably, will yield a range of results, with only the favorable results being reported, and the others either being discarded and suppressed or, as in the example in the text, made to conform to the preferred results. For a discussion of cherry-picking in litigation-developed statistical evidence, see David W. Peterson & John M. Conley, *Of Cherries, Fudge and Onions: Science and its Courtroom Perversion*, Law & Contemp. Probs., Autumn 2001, at 213, 227-32. Some may regard various cherry-picking procedures as the hallmark of good lawyering.

184. National Inst. Just., *supra* note 179, at 229. Mr. McCann manifests awareness of the power of observer effects in the next case he recounts, which dealt with a fingerprint examiner who testified that prints were “fresh” (a very helpful piece of testimony in the particular case) even though there is no way of determining the “freshness” of a print. McCann concludes, “I firmly believe that the error was inadvertent in that the technician’s keen desire to support the prosecution and anticipate the

Interestingly, the District Attorney offered this incident as an example of his efforts to make sure that truly correct and proper results are obtained from forensic examinations.¹⁸⁵ But the mere making of a request for reconsideration conveys information and sets up expectations, so it has to be done with care if it is not to bias the outcome. Imagine what result might have emerged if the District Attorney had called the state examiner and asked him to reconsider his conclusions because they were in conflict with those of the FBI. Or, if he had called both and merely pointed out the conflict, not telling either of them anything about his theory of the case so that neither knew whether the wounded subject was a police officer, or someone shot by a police officer.

Another example is provided by the proceedings in *United States v. Mitchell*,¹⁸⁶ a case in which the accuracy of fingerprint identification was challenged under Rule 702.¹⁸⁷ In response to the defendant's challenge and in an effort to prove the claim that all fingerprint examiners reach the same conclusion on the same evidence, an FBI supervisory fingerprint specialist sent two latent prints and a known fingerprint card to each of the fifty state crime laboratories.¹⁸⁸ In the results that came back seven of the labs concluded that one of the latent prints could not be matched to the suspect, and five concluded that the other one could not be matched.¹⁸⁹ The FBI fingerprint specialist then enlarged the exhibits and annotated the latent prints, indicating the argued-for points of similarity on which a conclusion of identification would rely and sent these embellished exhibits to those experts who had reached contrary conclusions.¹⁹⁰ Those experts were then asked to reconsider their conclusions.¹⁹¹ This rather glaring attempt to persuade the "errant" examiners, and only the "errant" examiners, that they were incorrect and should change their opinions succeeded; they all acquiesced to the opinion being urged upon them. The obviously skewed nature of that process of "inquiry" illustrates the biasing effects of selective re-examination. Suppose, instead, the FBI had selected a sample of "non-errant" examiners and sent them similar exhibits pointing out the bases on which those who found no match had reached their conclusions. One then would have been able to assess the extent to which the reversed opinions were a product of reconsideration of the actual evidence or acquiescence to the cues being sent. But the structure of the request for recon-

defense caused him to subconsciously put the word "fresh" before the words "palm print." *Id.* at 241-42.

185. *Id.*

186. *United States v. Mitchell*, No. Crim. #96-407-1 (E.D. Pa. 2000). A previous trial and conviction resulted in a reversal, reported at 145 F.3d 572 (3d Cir. 1998).

187. *See generally* Simon Cole, *The Myth of Fingerprints*, *Lingua Franca*, Nov. 2000, at 54.

188. *Id.*

189. *Id.*

190. *Id.*

191. *Id.*

sideration insured that the situation could only get "better." All of the opinions that had come back in the favored direction were allowed to remain set; the opinions in the nonpreferred direction were invited, indeed encouraged, to be reversed.

The bias in this kind of situation is powerful. In the first place, some findings but not others are being re-examined, thereby leaving the preferred results in place but inviting change in the nonpreferred results. Second, the examiner to whom the re-examination request is made is told that another examiner reached different, and more pleasing, conclusions. We are not saying that this kind of situation always will produce the results being sought by the party requesting re-examination, but it will at least sometimes. And there is no possibility for the opposite to happen, because a selection bias has been created that allows only the nonpreferred result to be subject to revision.

We do not suggest that the examples above represent a deliberate perpetration of fraud on the courts that have the duty to weigh the evidence in these cases. Putting aside claims that prosecutors ought to be held to higher standards, an argument could even be made that what has been done in these examples represents diligent lawyering. But it surely represents poor science.

III

Minimizing Observer Effects in Forensic Science: Conclusions and Recommendations

As a result of the growing number of DNA exonerations,¹⁹² and the analyses of those cases to determine what went wrong, it is beginning to appear that forensic science contributes more to convicting the innocent than anyone previously suspected. The data indicate that forensic science error rivals eyewitness error as the leading cause of erroneous convictions.¹⁹³ This trend should be enough to give anyone pause at the continuation of business as usual in many areas of forensic science.

Before proceeding further, we would ask the reader to perform the following mental experiment. Assume that you have been called to a distant planet and asked to set up a new system of forensic science laboratories with the goal of producing results of maximal accuracy. Which of the following options would you choose in establishing such a system?

192. See Barry Scheck, Peter Neufeld & Jim Dwyer, *Actual Innocence: Five Days to Execution and Other Dispatches from the Wrongly Convicted* (2000) (documenting exonerations of persons convicted of murder and awaiting execution through the use of DNA evidence).

193. *Id.* at 263.

- Either laboratories would be set up in the new system as arms of criminal law enforcement, or laboratories would be freestanding entities available to both prosecution and defense.
- Either examiners in the new system would be drawn largely from the ranks of current law enforcement officers, or examiners in the new system would be recruited from the ranks of people interested in science with no pre-existing law enforcement bias.
- Either examiners in the new system would be socialized in such a way, and would interact with case detectives in such a way, as to feel themselves an integral part of a law enforcement “team,” or examiners would be insulated from such influences and trained to form no such role view, but instead to view their role solely in terms of the maximal integrity and maximal accuracy of their own results.
- Either examiners in the new system would be exposed to much domain-extraneous information in the process of conducting an examination, including the emotionally gripping details of the underlying case and the hopes and expectations of the case detectives, or specific procedures would be put in place to separate relevant information from extraneous information, and to insulate the examiners from exposure to the latter.

We suggest that the answers to the above choices are obvious. Yet our current system, largely for reasons of historical accident, has generally answered these questions the wrong way, if reliability is what we are after. Historically, criminal defendants as a group benefited from the unavailability of information. It is hardly surprising that the law enforcement arm of the state organized efforts to apply science and quasi-science methods to problems of solving and proving criminal cases. In so doing, law enforcement utilized the tools available: officers trained as “technicians” by the small number of scientists with law enforcement interests.¹⁹⁴

In seeking to change these historical remnants, we do not pretend that we are writing on a clean slate. In regard to some organizational reforms, as the vaudeville punch line says, “you can’t get there from here,” at least not within the foreseeable future. The establishment of freestanding gov-

194. While the product of whatever science an era might muster has made its way into the courtroom for centuries, see, e.g., *The Trial of Spencer Cowper, Ellis Stephens, William Rogers, and John Marson, at Hertford Assized, for the Murder of Ms. Sarah Stout* (1699), in *13 A Complete Collection of State Trials* 1105 (T.B. Howell ed., 1812), until the early twentieth century it was the ad hoc product of individual practitioners. General forensic science laboratories in the United States have generally been set up as an adjunct to law enforcement organizations. Stuart Kind & Michael Overman, *Science Against Crime* 31 (1972). The first laboratory worthy of the name was set up in the Berkeley (Cal.) Police Department by August Volmer around 1918. Jurgen Thorwald, *Crime and Science* 149 (1966). The characterization of most of the personnel in such labs as “technicians” rather than “scientists” is from Andre A. Moenssens, *Novel Scientific Evidence in Civil and Criminal Cases: Some Words of Caution*, 84 *J. Crim. L. & Criminology* 1, 5 (1993).

ernment forensic laboratories, though occasionally advocated,¹⁹⁵ would require such a revolution in thinking and organization, and diminish so many established bureaucratic empires, that it would take a generation of patient lobbying to have a chance of success.

The winds of change are beginning to blow, however, for reasons independent of any explicit calls for reform. The biggest single factor contributing to this change appears to be the increased forensic use of academic science disciplines which cannot be adequately taught to law enforcement personnel as “technicians,” such as forensic chemistry, forensic anthropology, and DNA analysis. Sometime over the past quarter century, the percentage of trained personnel in the larger forensic science laboratories with advanced degrees in science appears to have begun to grow.¹⁹⁶ This has created a culture collision of significant proportions, the most public manifestation of which was “l’affaire Whitehurst” at the FBI laboratory.

Frederick Whitehurst is a Ph.D. chemist who was hired by the FBI laboratory in 1982.¹⁹⁷ From the beginning of his employment he seems to have been shocked by the unscientific methods of some of his colleagues, and he complained about them.¹⁹⁸ Whitehurst’s complaints led to the Inspector General’s investigation,¹⁹⁹ which substantiated many of his allegations,²⁰⁰ and, prospectively at least, adopted recommendations aimed at insuring the existence of more defensible methods in the FBI laboratory.²⁰¹ The Whitehurst affair is merely a manifestation of the leavening of the traditional forensic science laboratory culture with personnel seriously trained

195. See Giannelli, *supra* note 43, at 472-74. In forensic pathology, at any rate, such independence is recognized as an important value: “No medical examiner’s office should function under a police agency. There is a direct conflict in values, goals and philosophies. The police want to make an arrest and clear a case. The medical examiner’s office wants to determine the cause and manner of death independent of who did what.” Di Maio & Di Maio, *supra* note 151, at 12. How closely reality approaches this goal may be another matter.

196. For instance, until 1993, except in the fingerprint section, only persons who underwent full agent training and had served as normal investigatory agents could qualify to become forensic examiners for the FBI. Thus, any Ph.D. chemist interested in working at the FBI lab would have to be willing to undergo both normal law enforcement training and service. This policy was changed in 1993, and the relative percentages of agent and nonagent examiners have begun to shift. See Inspector General’s Report, *supra* note 61, at 9-10.

197. *Id.* at 13.

198. *Id.* It should be noted that one of the main targets of his complaints, Terry Rudolph, was also a Ph.D. chemist, a circumstance which may have intensified Whitehurst’s contempt for Rudolph’s unprofessional sloppiness. *Id.* at 6, 13.

199. *Id.* The Report is not entirely kind to Whitehurst. *Id.* at 476-79. One senses a somewhat bureaucratic motivation in some of the Report’s less kind conclusions, though it does appear that over the years Whitehurst grew increasingly eccentric. However, most of his eccentricities can be accounted for as the reactions of an embattled man of stubborn integrity being harassed both by his immediate colleagues and by an institution for not going along with business as usual.

200. *Id.* at 479.

201. *Id.* at 480-516.

in the methods of academic science, who come to their new jobs primarily for the science and less for the law enforcement satisfactions involved.

While some desirable structural changes seem unrealistic, and other desirable changes are happening by evolution and infusion, the serious problems of observer effects can only be solved, or at least ameliorated, by intentionally embraced changes in forensic practice. These changes will be neither tremendously complex nor excessively expensive; fortunately, many of these problems already have solutions that are in routine use in most scientific fields, and that can be found in the standard research methodology textbooks of those fields.

The first step is awareness, which we hope has been fostered by this Article. Such awareness of the phenomenon of observer effects is a necessary, potentially powerful, but entirely inadequate step. As in other areas of practice, awareness alone is not enough; action is required: "The discovery of suggestibility in patients undergoing experimental treatments necessitated the introduction of the placebo experiment, and the possibility of similar suggestibility on the part of experimenters led to the double-blind experiment."²⁰² Forensic scientists have no less need, and no less ability, than so many other serious scientists around the world to institute procedures to protect their findings against avoidable sources of error. "[T]he psychological fact of an omnipresent tendency toward motivational bias fully justifies those many aspects of experimental procedure, objective scoring, instrumentation, and the like that guard against self-deception."²⁰³

A. Preventing Distortions Due to Expectation and Suggestion: Blind Testing

It would be hard to disagree with the Inspector General's affirmation that examiners should not "base forensic conclusions on unstated assumptions or information that is collateral . . ."²⁰⁴ Obviously, forensic conclusions cannot be based on such extraneous information if the examiner is not exposed to it. The simplest, most powerful, and most useful procedure to protect against the distorting effects of unstated assumptions, collateral information, and improper expectations and motivations is blind testing. An examiner who has no domain-irrelevant information cannot be influenced by it. An examiner who does not know what conclusion is hoped for or expected of her cannot be affected by those considerations.²⁰⁵

202. Campbell, *supra* note 113, at 29.

203. *Id.* at 36-37.

204. Inspector General's Report, *supra* note 61, at 511.

205. Because forensic science deals mostly with inanimate objects, the blinding procedure will be simpler than in fields that work with humans and animals, such as biomedical research and psychology. Those fields must construct double-blind studies, while forensic science needs only single-blind procedures.

A wall of separation must be created between forensic science examiners and any examination-irrelevant information about a case. That means properly controlling information flowing to examiners from external investigators,²⁰⁶ from laboratory managers, and from fellow examiners. Controlling this information will not always be simple and straightforward; sometimes examiners need to know certain details of a crime to develop meaningful hypotheses and to determine what tests need to be done. The solution is to provide examiners with the information they need to perform the tests, and only that information. At times, good practice might require sharing information in stages—giving examiners certain information necessary to performing a test, then, following the results of that test, providing additional information that might lead to additional testing. Doing so protects the soundness of the early testing without losing the benefit of the later testing.

This kind of information management can easily be made to fit in with administrative structures common in forensic laboratories. For instance, the FBI laboratory historically has had a three-step procedure for processing evidence submissions.²⁰⁷ The first contact would be in the Evidence Control Center, where an employee would log in the submission, give it an evidence control number, and then route the submission to a relevant investigatory unit.²⁰⁸ The Unit Chief would then receive the submission and decide which examiner would act as the primary case agent. This examiner was responsible either for testing the submission or coordinating its testing by various units if more than one set of tests is necessary.²⁰⁹ At each of these stages the personnel had available the “submission” document and were free to contact or be contacted by the case investigators. While the current administrative structure of the FBI lab has changed somewhat, there is no evidence that this general structure has changed. Moreover, small changes to this structure could do much to eliminate observer effects.

The most important change would be to convert the personnel in the Evidence Control Unit from fundamentally clerical personnel²¹⁰ to the most highly trained and highly respected personnel in the laboratory, true

206. One significant consideration necessitated by this regime would be how to deal with the criminalist whose specialty is “crime scene” as it is referred to in ASCLD Standard 2.11. ASCLD Manual, *supra* note 135, § 2.11. Serious thought must be given to either insuring their insulation from inappropriate suggestion or insuring that the products of such suggestion do not leak through to the examiners who do actual testing. This concern is complicated by the fact that in some smaller settings the criminalist may also perform tests. Similar considerations apply to the control of crime scene visits by testing examiners.

207. David Fisher, *Hard Evidence* 22-23 (1995).

208. *Id.* at 22.

209. *Id.* at 22-23.

210. Although the duties the Evidence Control Unit performs are almost entirely clerical in nature, they do include determining the initial order of testing based on such considerations as relative destructiveness. Ubelaker & Scammell, *supra* note 142, at 64.

“Evidence Control” and “Quality Control” officers. Such officers would be required to have advanced degrees in some normal science discipline and to undergo rigorous training. This training would enable them to implement programs designed to filter out all domain-irrelevant information from submissions, to formulate questions in the least suggestive way, and to route and coordinate the submission of the evidence to the appropriate section or sections.²¹¹ The Evidence and Quality Control Officer²¹² would be responsible not only for coordinating work among examiners in different specialties, but also for being the sole contact point between the entity requesting the test and the laboratory. She would also serve as the filter between each examiner and any information about the case, whether it originated from without or from within the lab. She would decide not only generally what kinds of tests were needed, but what information about the case was needed to perform those tests, and her primary duty would be to maintain appropriate masking between the examiners and all sources of domain-irrelevant information.²¹³

Put simply, good scientific practice is to “keep the processes of data collection and analysis as blind as possible for as long as possible,”²¹⁴ and to accurately document what was done, making that documentation automatically available to anyone concerned with the reliability of the test procedures, including criminal defendants. Such a regime may well produce fewer “positive” results, but it is hard to see how any defensible positive results would be lost, and the number of “false positives” will be minimized.

*B. Preventing Distortions Due to Assumed Base Rates of
Inculcation: Evidence Lineups*

The forensic scientist’s situation is unusual in that the job often comes with an almost built-in expectation that tested evidence will inculcate, even in the absence of a domain-irrelevant suggestion. For example, in a detailed study of four different crime laboratories, Joseph Peterson, Steven Mihajlovic, and Michael Gilliland found that, on average, fewer than 10% of all reports disassociated a suspect from the crime scene or from connec-

211. In 1997, the FBI announced its intention to create four “supergrade level science positions” whose duties would include “problem solving, liaison with the relevant scientific communities, and quality assurance.” *Inspector General’s Report*, *supra* note 61, at 509. Such supergrade positions could easily be utilized in the manner suggested in this Article.

212. This could be conveniently abbreviated “EQC,” but we are hesitant to adopt this usage in the first article recommending the position’s creation.

213. A somewhat similar process has been under development in the United Kingdom’s Forensic Science Service. It is based on Bayesian principles and involves a more formalized process of “pre-assessment” of hypotheses and what would be required of the evidence to test those hypotheses, as well as careful documentation of every step of the process. R. Cook, I.W. Evett, G. Jackson, P.J. Jones & J.A. Lambert, *A Model for Case Assessment and Interpretation*, 38 *Sci. & Just.* 151 (1998).

214. Rosenthal, *supra* note 105, at 1007.

tion to the victim.²¹⁵ This high rate of inculcation comes from the fact that each piece of evidence connected with any suspect has a heightened likelihood of being inculpatory, since investigators do not select suspects or evidence at random, but only those they have some reason to think were connected to the crime. Thus, forensic scientists have a continuing expectation that the evidence before them is inculpatory, which is perhaps reinforced by the role effects noted earlier, a situation likely to strengthen confirmation bias and selective attention effects.²¹⁶

Whatever the reasons, the inclusion rate is high, and examiners come to expect it to be high. Blind testing procedures, while fundamental and fairly easily and cheaply instituted, cannot remove these base rate-induced expectations that most examinations will lead to inculcation. In addition, as indicated earlier, the more subjective and less instrumented a forensic technique is, the more subject to expectation-induced errors it is, and the more important finding a solution to such sources of expectation-induced error becomes.

Fortunately, there is a technique that can provide a solution to this problem, namely, an evidence lineup. In an evidence lineup, the examiner would be presented with multiple specimens, some of which were "foils." The examiner would, of course, be blind to which items of evidence in the evidence lineup are foils and which are the true questioned evidence. For example, a firearms examiner might be presented with a crime scene bullet and five questioned bullets labeled merely "A" through "E." Four of those bullets will have been prepared for examination by having been fired through the same make and model of firearm as the crime scene bullet and the suspect's bullet had been. The task for the examiner would then be to choose which, if any, of the questioned bullets was fired through the same weapon as the crime scene bullet had been.

Appropriately designing such lineups and submitting evidence to examiners in this form would be another responsibility of the Evidence and Quality Control Officer. The evidence lineup would perform many of the same functions that an eyewitness lineup does. Examinations in forensic science labs are currently the equivalent of show-ups in the eyewitness

215. Joseph L. Peterson, Steven Mihajlovic & Michael Gilliland, *Forensic Evidence and the Police* 117 (National Institute of Justice Research Report, 1984).

216. Indeed, the high rate of inculcation might be a consequence of police investigative work performed so well that labs rarely are troubled with evidence that turns out to be exculpatory. Alternatively, it may reflect expectations on the part of examiners that most of what is given to them is going to incriminate, or reflect policies or cultures of labs that evidence ought to incriminate. Support for the latter possibilities comes from the Peterson study's finding that laboratories varied greatly in their criteria for conclusions, so that the same evidence reported by one as "not sharing a common origin" was reported by the others to be "inconclusive." *Id.*

realm.²¹⁷ In both settings, the test is structured to be single-suspect, implying that the correct suspect is in hand and preventing the ultimate decisionmakers (the courts) from evaluating the quality of the test and the likely validity of its results. A properly constructed and controlled lineup solves these problems. The Justice Department's recent guidelines on scientifically informed eyewitness identification procedures discusses these details which, except for those guidelines pertaining to interviewing, apply equally well to evidence lineups.²¹⁸

In brief, for forensic science examiners, a properly constructed evidence lineup would accomplish at least the following. The examiner would know from the sheer structure of the test situation that most of the questioned evidence items are not associated with the suspect, and that a failure to exercise real expertise in an unbiased fashion is likely to lead to an incorrect conclusion. That would erase much of the impact of base rate-induced expectations. In contrast to the show-up situation that today is common practice, an examiner would not be able simply to conclude "inclusion" or "inconclusive" in virtually every test. On the other hand, when an examiner rejects all of the foils and concludes that the one known evidence item matches the crime-scene item, this conclusion will be far more powerful and persuasive evidence because it has not been affected by observer effects.²¹⁹ Moreover, the lineup structure, unlike the customary show-up procedure, allows a direct and exact calculation of the probability that the examiner would have reached a correct conclusion by chance.

Proper evidence lineups present some nontrivial problems of design, requiring the Evidence and Quality Control Officer both to determine what

217. A "show-up" is an identification procedure where the witness is presented with a single suspect for identification.

218. Technical Working Group for Eyewitness Evidence, United States Dep't of Justice, *Eyewitness Evidence: A Guide for Law Enforcement* (1999). The recommendations involve presentations of multiple candidates, one at a time, by an investigator who does not know which is the actual suspect. At least one jurisdiction, New Jersey, has adopted these recommendations. See *Witnesses, Victims Get New Way to ID Suspects*, Sunday Record (Bergen Co., N.J.), July 22, 2001, at A-3.

219. Evidence lineups are currently rare but not unheard of. In *State v. Stokes*, 433 So. 2d 96 (La. 1983), a murder case, the trial court, as a condition of compelling the defendant to submit to dental casting for comparison to bitemarks found on the victim's body, required that the defendant's casts be presented to the forensic odontologist identified only by a code number, and accompanied by four other casts of the teeth of males of the same general age as the defendant, two to be selected by the prosecution and two by the defense. The expert was unable to form a conclusion, saying only that he could not rule out any of the sets of teeth represented by the various casts as the source of the bitemarks. *Id.* at 103. Compare this procedure to the procedure undertaken by Dr. Lowell Levine in *State v. Fortin*, discussed *supra* note 167 and accompanying text. Incidentally, Dr. Levine gave an opinion of similar certainty to that which he gave in the Fortin case (apparently under similar circumstances) in regard to the 1998 Maine murder of Irene Kennedy. Levine identified the police's prime suspect, Edmund Burke, as the source of bitemarks found on Kennedy's body. Later DNA tests of saliva from the bitemarks and comparison of a bloody palmprint found on the victim's thigh with Burke's palmprint exonerated Burke, and the prosecution was discontinued. *The Justice Files: I Am Innocent* (Discovery Channel television broadcast, Sept. 6, 2001) (tape on file with authors).

would constitute appropriately similar foil specimens and to arrange to obtain them. This process would obviously be easier for some types of examinations than for others. Unfortunately, it may often be most difficult precisely where it is most needed, in those areas, such as handwriting identification, with the least instrumentation and greatest subjectivity. Nevertheless, if forensic science is to move to a more defensible scientific model, the effort must be made to eliminate the inclusion bias that currently appears to be endemic.²²⁰

The fundamental tasks of the eyewitness and of the forensic examiner share notable similarities, suffer from remarkably similar sources of potential systematic error, and enjoy the same potential for elimination of those problems merely by structuring the tasks in a rigorous fashion. The goal of the Justice Department in promulgating its eyewitness guidelines, namely, reducing the incidence of false positive errors without reducing the incidence of true positive identifications, can be achieved equally well for forensic science.²²¹

C. Likely Objections to the Recommendations

Although the use of blind (and double-blind) testing protocols has been readily embraced by a great multitude of scientific fields, forensic science remains a prominent exception. What might be the special concerns forensic scientists would offer to justify excusing them from adopting these more rigorous procedures?

One argument might be that certain context information is needed to make a proper interpretation of the evidence. No suggestion has been made, however, that examiners be denied information that is appropriate and necessary to doing their proper job. At the simplest level, without latent prints from a crime scene, a fingerprint examiner cannot evaluate whether the suspect's fingerprints match or not. No suggestion has been made that examiners be required to guess. At the other extreme, it is hard to imagine how knowing that a purse belonging to the crime victim was

220. In conducting such an evidence lineup, it would seem that, along with each decision made, and before receiving any post-test feedback or other extraneous information, the examiner ought to be required to record a rating of his or her subjective confidence that the selected questioned evidence item shares a common origin with the crime scene evidence. The rating would be a check on extraneous postexamination information creating an exaggerated confidence in an opinion originally reached with less confidence.

221. Blind testing and evidence lineups are two procedures that, in combination, would solve the majority of the problems resulting from observer effects that occur when any human being sets about to make decisions of the sort made in forensic examinations. But, for those labs that are interested in thinking about developing procedures that go even further, there are additional techniques that may be adopted, including the following: cancellation of biases (creating counterbalanced and mutually self-canceling expectations), production of biases (on a periodic experimental basis to monitor their effects), and increased development and use of mechanical or electronic recorders and apprehenders (thereby reducing the human role in the observation).

found at the suspect's home could ever play a valid role in resolving uncertainties in comparisons of DNA or firearms or handwriting. The difficult problem comes in between, in making a judgment about what is or is not appropriate and necessary. No doubt that would vary with the details of the case and the nature of the tests being conducted.²²² But, inevitably, the question what context information is needed to make a proper examination must be left to the informed judgment of the Evidence and Quality Control Officer, based on protocols developed for each forensic specialty. The coordinating examiner can achieve the benefits of blind testing by supplying the minimum amount of information initially and then taking those test results, as well as other test results and other case information, into account when final conclusions are to be drawn. If need be, examiners can do further tests with the benefit of additional context information. At the end of the day, the coordinating examiner would have the advantage of contamination-free scientific information, as well as the advantage of full context information and any consultation with colleagues that was necessary. Yet all of them would then know what went into a particular judgment and what did not.

A second objection might come from forensic scientists' assertions that they are taught to disregard biasing information that comes into their possession, and can by the exertion of their will rid themselves of distorting influences.²²³ This argument, however, reflects little understanding of the nature of the problem. Every field that has considered the problem has concluded that it cannot be solved merely by trying to will it away. When everyone from Nobel Prize winners to average citizens, who informally subject themselves to homemade "blind taste tests," take steps to make sure their judgments are not distorted by extraneous context information, then it is hard to conceive of what it is that makes forensic scientists think they are immune from the same effects.²²⁴

222. Ironically, the more science-based the tests (for example, chemical analyses, DNA typing), the more easily they can be conducted in something that approaches blind testing.

223. Apparently many forensic pathologists take this position. While recognizing that the effects of police suggestion are a serious problem for the "unqualified," they appear confident that their qualifications will eliminate the problem. Consider the following passage from Di Maio & Di Maio, *supra* note 151, at 14:

[The police] prefer the charlatan who tells them what they want to hear to the expert who tells them unpalatable truths or that conclusions cannot be made. One of the characteristics of the unqualified expert in forensic pathology is an ability to interpret a case in exquisite detail. This "expert" sets the time of death, plus or minus a few minutes, accurately positions the deceased, and gives detailed analysis of the events surrounding the death and precise deductions about the assault. If the police have expressed prior opinions, it is not uncommon for the opinions of the "expert" to agree almost in complete detail with the police hypotheses. The experienced forensic pathologist tends to hedge, knows there may be more than one interpretation of a set of facts, and is more "wishy-washy" than the charlatan.

224. This is not a question that need be the subject of speculation and argument. If forensic scientists believe that something in their training, which is lacking in the training of all other scientific disciplines, makes them immune to context effects, it would be a relatively simple matter to design

Another possible argument is that blind testing is insulting. Though this argument is not often heard from forensic scientists, it has been offered by police in resisting double-blind testing in the conduct of lineups. Though conceding the benefits of blind testing, they feel their colleagues would be “insulted,” and would feel they are “not trusted,” if it were required that lineups be conducted only by officers who do not know which lineup member is the actual suspect.²²⁵ This rationale is something of a puzzle. While the feelings of police officers (and forensic scientists) are not unimportant, surely this concern pales when placed alongside of the primary goal: developing the most valid possible evidence for criminal courts to use in making sober and weighty decisions. Moreover, if scientists of all other kinds do not feel insulted to be expected to carry out their research in proper blind or double-blind fashion, why should forensic scientists feel any differently?

Related to opposition based on a perceived “insult” is opposition based on suspicions of increased bureaucratization and associated loss of job satisfaction.²²⁶ Put bluntly, the forensic examiner is used to being a kind of free agent as regards the individual case, and to having the excitement and drama of following cases as they unfold. Blind testing would put the examiner more in the position of a technician in a medical lab. While this perception is true, the cost in inevitable and undiscoverable error from allowing such job satisfaction considerations to prevail is simply too high. The forensic examiner must learn to delay curiosity and dramatic gratification until after the examinations are completed and the results are in.

Finally, it can be argued that costs will rise if testing is conducted blind, under the guidance of Evidence and Quality Control Officers, and especially if evidence lineups were to be adopted as the standard of practice. This argument is certainly true. However, virtually all other fields of science have determined that the risk of harm due to observer effects is so

appropriate experiments and test the claim empirically. The one study of which we are aware that has actually tested forensic examiners for the effects of biasing context information found (not surprisingly) that the results of hair comparison varied as a function of the manner in which the samples were presented to the examiner: traditional paired comparison of a questioned with a known exemplar versus a lineup style presentation. Larry Miller, *Procedural Bias in Forensic Science Examinations of Human Hair*, 11 L. & Hum. Behav. 159-62 (1987).

225. See, e.g., Gary Wells et al., *From the Lab to the Police Station: A Successful Application of Eyewitness Research*, 55 Am. Psychologist 581, 594 (2000). The DOJ recommendations on proper lineup procedures reject this position.

226. Though not explicitly asserted, this source of opposition was suggested by the nature of many of the comments made in an online discussion among forensic scientists during a discussion about the desirability of blind testing. The discussion took place in June 1998 on Forens-L, a forensic science online discussion listserve, and is reproduced under the title *The Need for “Blind” Procedures in Forensic Science, Scientific Testimony: An Online Journal*, at <http://www.scientific.org/open-forum/articles/blind.html> (hard copy on file with authors). It should be noted that one of the authors (Thompson) runs the site and was a participant in the discussion.

great, and the need for valid findings is so important, that the increased costs are worth paying in order to gain the benefits that proper testing procedures bring. If cost is to be a consideration, it should be noted that at least some of the proposed reforms will likely add very little to operating expenses of laboratories once the transition in structure and training is completed.

IV

Observer Effects and Admissibility Under Federal Rule of Evidence 702

Prior to the decision in *Kumho Tire*, the problem of observer effects managed to fly below the law's radar. The usual frame of reference that courts adopted to make Rule 702 reliability judgments was the global reliability of proposed expertise, and not taking into account the nonoptimum conditions under which the particular conclusion was rendered.²²⁷ By its emphasis on reliability under the conditions of the particular case, *Kumho Tire* has changed this.²²⁸ So what should now be the judicial response to a claim that particular expert testimony ought to be excluded as unreliable pursuant to Rule 702 because of the presence of a substantial risk that the expert's results were contaminated by observer effects? There are a number of potential responses, and we will try to deal with them in turn.

One possible response would be to conclude that observer effects pose insufficient dangers to the reliability of forensic science expertise to warrant attention in the Rule 702 reliability calculus.²²⁹ This response might be the initial instinct of some judges, given the longstanding admissibility of such evidence²³⁰ and the heavily precedent-oriented and inertial nature of

227. See, for example, the extensive analysis showing the global approach regarding the handwriting identification cases in Risinger, *Defining the "Task at Hand," supra* note 7, at 778-98.

228. See *supra* notes 1-15 and accompanying text.

229. Fed. R. Evid. 702. At the time of the decision in *Kumho Tire*, the Rule provided: "If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise." Fed. R. Evid. 702. It was subsequently revised effective Dec. 1, 2000, to reflect more particularly the *Daubert* decision. It now reads: "If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case." Fed. R. Evid. 702.

230. The only reported decision we have been able to discover that directly decided a challenge to proffered forensic science expertise based on the suggestive context in which the expert arrived at a conclusion is *State v. Asherman*, 478 A.2d 227 (Conn. 1984), another case involving bitemark identification. The court rejected the challenge, which was based on a claimed violation of Constitutional due process requirements, relying in part on dicta in *United States v. Wade*, 388 U.S. 218, 227-28 (1967), to the effect that the problem of suggestion in forensic science identification was not as serious as that of eyewitness identification. Perhaps the empirical record might lead to a

the legal process, so frankly captured in the common law maxim “better a fiction than a novelty.”²³¹ However, the weight of the research and the condition of normal forensic science practice render such a response so irrational that in the long run it cannot prevail over the responsibility to evaluate the reliability of such testimony pursuant to *Daubert* and *Kumho Tire*. In particular, *Kumho Tire*’s mandate to evaluate the reliability of expert conclusions whenever their “factual basis, data, principles, methods or their application are called sufficiently into question,”²³² and new Rule 702’s requirement that the testimony in the case be “the product of reliable . . . methods,”²³³ would seem to foreclose this instinctive response. For what could more centrally call into question the methodology by which a particular conclusion was reached than the uncontrolled presence of the precursors of various observer effects, which render it impossible to say with confidence whether or not the conclusion is merely an artifact of these conditions? This would seem the very definition of “unreliability.” And what more is needed “sufficiently” to call the methodology of a particular conclusion into question than the current generally uncontrolled state of normal forensic science practice? If more is needed, it can be supplied through an examination of the conditions actually prevailing during the consideration of the particular evidence under review, pursuant to the explicit requirement of revised Rule 702(3) that such reliable methods have been “applied . . . reliably to the facts of the case.”²³⁴

A second possible response is that, at least as to the products of forensic science based on experience and subjective evaluation, such evidence should be excluded until the proponent shows it to have been the product of a process uncontaminated by domain-irrelevant information or the effects of institutional influence and expectancy. Such a response might be salutary, not only because of the unreliability of results generated by such processes, but also because no general reform of practice is likely to be forthcoming unless that reform is required by the courts through decisions excluding evidence.

There is very little likelihood, however, that any judge will adopt such a general position, and perhaps with justification, since such a decision would arguably be too global to comport with the individualized “task at

reassessment of that position today. At any rate, future attacks on admissibility are likely to be premised on the proper construction of Rule 702 or its state analogues, issues not addressed in *Asherman* or *Wade*.

231. Perhaps not surprisingly, there seems to be exquisite resistance on the part of judges to being the first to exclude evidence which has been routinely admitted for generations. See the explicit invocation of this reluctance in regard to handwriting identification testimony in *United States v. Jones*, 107 F.3d 1147, 1158 (6th Cir. 1997). One of the authors (Saks) has personally heard at least one other judge make similar comments from the bench.

232. *Kumho Tire*, 526 U.S. at 149.

233. Fed. R. Evid. 702(2).

234. Fed. R. Evid. 702(3).

hand” analysis mandated by *Kumho Tire*.²³⁵ Nevertheless, it seems clear under *Kumho Tire*, that in making a Rule 702 reliability determination, a judge ought appropriately to consider whether, and how well, the institutional setting in which an expert’s conclusion was reached addresses role bias and built-in expectancy, how unmasked in fact were the procedures utilized, and how contaminated individual conclusions have been by exposure to domain-irrelevant information. These considerations are to be weighed with other information, such as data on the demonstrated ability of examiners to reach accurate results in the particular “task at hand” under test conditions, the subjectivity of the process, the intensity of such drawbacks as low “signal-to-noise” ratio in the case before the court, and so forth.²³⁶ Not only is it appropriate to weigh observer effects, but also the research indicates that these effects should constitute fairly heavy weights in the resultant determination of threshold reliability.

In addition, in making a determination of the reliability of the task performed in a case, there are things a court clearly should not do. Early in this Article, we said that it would be inappropriate for a judge to exclude expert testimony merely because other evidence unrelated to the expert’s domain had convinced the judge that the expert’s conclusions were in error.²³⁷ It is similarly clear, and perhaps even more so, that, just as an expert should not reach a conclusion based on domain-irrelevant information,²³⁸ a judge should not admit unreliable expert testimony just because the judge is convinced from other independent evidence in the case that the expert’s conclusions are correct. This would merely implicate the judge in the “echo chamber” phenomenon previously discussed.²³⁹

Further, both judges and attorneys should keep in mind that the factors which comprise the ground conditions for observer effects ought a fortiori to be proper subjects for discovery.²⁴⁰ These are factual conditions which

235. See *supra* notes 6-7 and accompanying text.

236. This seems to be the approach adopted by Judge Bataillon in *United States v. Rutherford*, 104 F. Supp. 2d 1190 (D. Neb. 2000), the only case we have discovered where the suggestive context in which the expert’s opinion was formed was raised in a reliability challenge. Partly based on this, and on other questions concerning the general reliability of handwriting identification, Judge Bataillon substantially restricted the expert’s testimony and disallowed his conclusion. *Rutherford* was subsequently acquitted. See the discussion of the case in Risinger, *Defining the “Task at Hand,” supra* note 7, at 796-97.

237. See *supra* text accompanying notes 13-14.

238. See *supra* text accompanying notes 126-156.

239. See *supra* text accompanying notes 131-132.

240. Whether or not such information would be subject to discovery under the current Rule 16 of the Federal Rules of Criminal Procedure is open to some question. The rule was drafted without reference to the not-yet-extant implications of *Kumho Tire*. Presumably, any written records bearing on the issue would be discoverable documents under Rule 16(1)(c), but undocumented procedures might be more difficult to discover. See *United States v. Shue*, 766 F.2d 1122 (7th Cir. 1985) (finding no obligation to reveal that a government photograph expert had used a magnifying glass, since no written report of it was made). After *Kumho Tire*, a strong argument can be made that evidence bearing on the existence of the preconditions of observer effects constitutes “Brady material,” at least in many cases

affect not just the weight but the threshold admissibility of such proffered expertise. Beyond this, in the event the expert testimony is admitted over challenge,²⁴¹ such conditions are appropriate topics of cross-examination and impeachment. As the Court said of such expert testimony in *Daubert*, “[v]igorous cross examination, presentation of contrary evidence, and careful instruction on the burden of proof are the traditional and appropriate means of attacking shaky but admissible evidence.”²⁴² Finally, concerning the “presentation of contrary evidence,” in the event such contaminating conditions are exposed it would seem appropriate to call an expert familiar with the somewhat counterintuitive power of such effects revealed by the research. The expert would educate the jury on the results of that research so that they may better evaluate what weight to give the product of a process contaminated by expectation and suggestion.²⁴³

And so we come to the end. We hope that this Article has brought to light a serious problem concerning the reliability of much of the expertise upon which the life and liberty of those charged with crime is often made to turn. *Daubert* and *Kumho Tire* commit at least the federal courts to take steps to deal with such problems when they are identified. And, in the end, the steps we take will testify eloquently to how much we really mean the well-worn slogans we so blithely repeat about the search for truth.

where the expert testimony is central. Brady material is information sufficiently exculpatory that it is required to be given to the defense as a matter of Due Process independent of formal discovery rules, pursuant to *Brady v. Maryland*, 373 U.S. 83 (1963), and its progeny. See generally Charles Alan Wright & Arthur R. Miller, 2 *Federal Practice and Procedure, Criminal* § 254.2 (3d ed. 2000). In any event, if a *Kumho/Daubert* hearing is held, a court can and should inquire into these matters.

241. There is reason to believe that criminal defendants’ challenges to proffered expertise have been systematically less successful than those of civil plaintiffs or, indeed, of prosecutors. See generally D. Michael Risinger, *Navigating Expert Reliability: Are Criminal Standards of Certainty Being Left on the Dock*, 64 *Alb. L. Rev.* 99 (2000). There are many factors which may account for this lack of success, including relative lack of resources, but one such factor appears to be a “systemic failure to seriously litigate these issues on the part of the criminal defense bar.” *Id.* at 135. These issues require both a sophistication of nonlegal knowledge and the kind of substantial advance planning that altogether too often get lost in the press of time and the shortness of money. It is time for some form of collective action on the part of the criminal defense bar to make such challenges practically available when the nature of the proffered evidence rationally demands it. For similar, though less explicitly critical, observations coupled with suggestions on how to proceed, see generally Richard H. Underwood, *Evaluating Scientific and Forensic Evidence*, 24 *Am. J. Trial Advoc.* 149 (2000).

242. *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579, 596 (1993).

243. One may anticipate judicial hostility to such witnesses, though the rationale for such hostility is anything but clear. On the function of such “summarizational” or “educational” experts, see D. Michael Risinger, *Preliminary Thoughts on a Functional Taxonomy of Expertise for the Post-Kumho World*, 31 *Seton Hall L. Rev.* 508, 511-18 (2000). The use of such “educational” witnesses is generally recognized as proper in the advisory committee notes to both the original and revised rule 702. What is clear is that courts have been less receptive to such witnesses proffered by criminal defendants than one might suppose appropriate. See Risinger, *Navigating Expert Reliability*, *supra* note 241, at 131-35. Differential treatment of such “educational” experts proffered by the prosecution and defense is one of the clearest indicators of an element of pro-prosecution bias in the judicial handling of expert reliability issues in criminal cases after *Daubert*. *Id.*