
# The Potential of Blind Collaborative Justice

## Testing the Impact of Expert Blinding and Consensus Building on the Validity of Forensic Testimony

Carolyn Wong, Eyal Aharoni, Gursel Rafig oglu Aliyev, Jacqueline Du Bois

For more information on this publication, visit www.rand.org/t/rr804-1

This revised edition incorporates minor editorial changes.

This project was supported by Award No. 2013-IJ-CX0002, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice.

Published by the RAND Corporation, Santa Monica, Calif.

# Preface

One of the underlying causes of wrongful convictions is scientifically invalid testimony on forensic evidence, and the National Academy of Sciences has called for novel techniques to address this problem. The National Institute of Justice (NIJ) provided funds for RAND to examine the effects of two potential techniques to quantify and reduce testimonial bias: blinding experts to their party representation (so the experts do not know whether the prosecution or defense is the hiring party), and providing consensus feedback from a panel of experts. Results of the study may help define strategies, policies, and procedures by which expert testimony can be more effectively presented and evaluated at trial, thus minimizing the probability that wrongful conviction will occur due to biased expert testimony. This report documents the blind collaborative justice study expert panel exercise and experiment.

This study marks a first attempt to systematically examine the cognitive effects of expert blinding and consensus feedback in a single representative sample of scientific experts. Evidence that expert blinding and/or consensus feedback can improve the validity of expert testimony would be of great value to the programmatic efforts of the Department of Justice and others who are developing policies and procedures that are effective in reducing rates of wrongful conviction. This report should be of interest to NIJ, criminal justice personnel, lawmakers, potential expert witnesses, and other stakeholders in the criminal justice community who influence or develop policy with respect to the elicitation and delivery of expert testimony in criminal cases.

## The RAND Safety and Justice Program

The research reported here was conducted in the RAND Safety and Justice Program, which addresses all aspects of public safety and the criminal justice system, including violence, policing, corrections, courts and criminal law, substance abuse, occupational safety, and public integrity. Program research is supported by government agencies, foundations, and the private sector.

This program is part of RAND Justice, Infrastructure, and Environment, a division of the RAND Corporation dedicated to improving policy and decisionmaking in a wide range of policy domains, including civil and criminal justice, infrastructure protection and homeland security, transportation and energy policy, and environmental and natural resource policy.

Questions or comments about this report should be sent to the principal investigators, Carolyn Wong (Carolyn_Wong@rand.org) or Eyal Aharoni (Eyal_Aharoni@rand.org). For more information about the Safety and Justice Program, see http://www.rand.org/safety-justice or contact the director at sj@rand.org.

# Table of Contents

# Figures



# Tables

# Summary

A 2009 report by the National Research Council (NRC) expressly recommended developing methods for improving the validity of forensic testimony. The purpose of this study is to evaluate the ability of expert blinding and consensus feedback to improve the validity of expert testimony, specifically in the context of forensic science. *Expert blinding* is a procedure in which experts are not told whether they are testifying for the defense or the plaintiff/prosecution, thus reducing partisanship. While this approach has garnered initial empirical support in civil contexts, it has yet to be examined in criminal settings. Expert consensus feedback is a method for potentially reducing expert drift by culling expertise from multiple sources, feeding the majority view back to the individuals, and repeating this cycle to move toward a group consensus. This technique has been successfully used in a variety of disciplines, but has not been investigated in a courtroom scenario. Empirical support for this effect carries the potential to reduce wrongful convictions and their associated personal, financial, and societal costs.

The study approach consisted of conducting a vignette-based experiment with scientific experts. We developed a hypothetical criminal case with forensic evidence, generated a consensus interpretation of the evidence with a panel of experts in relevant fields, and then conducted an experiment with scientists in the same relevant fields. The experiment randomly assigned party representation (prosecution, defense, or blinded) and consensus feedback conditions (with or without advanced consensus interpretation) among the participants. Our hypothesis was that blinded expert response would be more accurate than one or more of the known adversarial conditions (prosecution or defense) and these differences would be greater without exposure to the consensus feedback.

The hypothetical case developed for the study involves criminal trespass and grand theft larceny of property belonging to a private business. The forensic question in the case is the probability that the defendant used the stairwell. The case description included facts that would logically lead to the correct answer to the forensic question. The solution could be derived using Bayesian probability techniques familiar to those with experience or academic training in a variety of physical and social science fields such as math, chemistry, psychology, criminology, and law.

In the consensus exercise, a panel of 12 scientists with doctorates in relevant fields evaluated the case to arrive at a consensus about the best answer to the probability question. Each panelist independently provided an answer along with an explanation of the logic used to arrive at that answer. Exchanges of the reasoning arguments that explained the logic supporting the responses were conducted electronically and anonymously. No panel member was told who else was on the panel. The entire panel arrived at the mathematically correct response after two iterations of

exchanging answers and supporting logic. That correct response was used as consensus feedback in the study survey.

The experiment had 580 usable responses from participants drawn from a pool of scientifically based professional society members. This analysis produced two key findings:

1. **Expert consensus feedback resulted in a performance improvement.** Expert consensus feedback regarding the correct response to the reasoning problem demonstrated the predicted significant effect on response errors; i.e., the delivery of feedback resulted in improved performance. In the experiment, 88 percent of the pre-feedback responses (without advanced consensus interpretation) were incorrect, and higher levels of consulting experience and education were not associated with greater response accuracy. Our finding that delivery of expert consensus feedback resulted in a performance improvement also indicates that presentation of expert consensus has the potential of decreasing the rate of erroneous responses.
2. **No advantage due to blinding was found.** Modest evidence of an allegiance bias was found: Participants assigned to testify on behalf of the prosecution (the party favored by the evidence in this case) produced greater errors than those assigned to testify for the defense. However, evidence of a relative advantage due to blinding was not found.

Given the nascent evidence of blinding on accuracy of testimony in criminal contexts, it would be premature to recommend broad changes to criminal justice procedure at this time. Much more research is needed to demonstrate real-world utility and feasibility. The effect of consensus exposure suggests that consensus expert feedback could be a promising way to reduce individual error in expert testimony, but research is needed to confirm the effect of consensus exposure and address practical and fairness issues associated with implementing consensus exposure mechanisms in real-world settings.

# Acknowledgments

We thank Dr. Katharine Browning, the project officer and our study point of contact at the National Institute of Justice, for her guidance and support of this research. In addition, we would like to extend our appreciation to the professional societies for allowing us to invite their members to participate in the study experiment. We also thank the members of the expert panel and all of the participants who completed the Justice Experiment for their time and candid responses. All input received was invaluable in our research. We thank James Anderson at RAND for his insights in helping us shape the legal case. We gratefully acknowledge the support of Brian A. Jackson, director of the RAND Safety and Justice program, and Tom LaTourrette, quality assurance manager for RAND Justice, Infrastructure, and Environment. Finally, we thank the three NIJ external technical reviewers, the NIJ internal reviewer, and Kerry Reynolds at RAND for their insightful comments.

# Abbreviations

| | |
|---|---|
| AAAS | American Association for the Advancement of Science |
| ANOVA | analysis of variance |
| E-DEL+I | Electronic Decision Enhancement Leverager plus Integrator |
| DOJ | Department of Justice |
| FRE | Federal Rules of Evidence |
| ID | identification |
| NAS | National Academy of Sciences |
| NIJ | National Institute of Justice |
| NRC | National Research Council |

# 1. Introduction

## Background

The federal government has recognized that there are concerns regarding forensic testimony in the courtroom.[1] For example, Congress directed the Attorney General to form a National Forensic Science Commission as part of the DNA Sexual Assault Justice Act of 2004 (Public Law 108-405). The Department of Justice (DOJ) fulfilled this mandate by awarding a contract to the National Academy of Sciences (NAS) to conduct a study that addressed congressional concerns on the use of forensic analysis. The resulting report noted that improvements in forensic science practices should reduce the occurrence of wrongful convictions but that the interpretation of forensic evidence is not always based on science to determine its validity, and that human interpretation of forensic evidence in criminal trials could be tainted by a number of factors, including bias in the interpretation (National Research Council [NRC], 2009).

In February 2013, partially in response to the NRC report, the DOJ and National Institute of Standards and Technology jointly announced the formation of the National Commission on Forensic Science. In apparent recognition of the potential importance of forensic testimony in criminal cases, the government specified that the commission is to include forensic scientists, defense attorneys, prosecutors, and judges. Moreover, one of the commission's responsibilities is to develop guidance concerning the intersections between forensic science and the courtroom (National Institute of Standards, 2013).

According to the Federal Rules of Evidence (FRE), decisions to admit and evaluate forensic evidence are the responsibility of the judge and fact-finders, neither of whom are likely to be experts in forensic scientific methods. Rules of admissibility typically require a demonstration that the evidence meets basic standards of scientific validity and acceptance by the scientific community (FRE, 2014). But judges are often reluctant to interfere with this process, preferring to let fact-finders weigh the evidence using the adversarial process (NAS, 2002). The wide discretion available to judges and jurors has prompted scientific organizations, such as the NRC, to make formal recommendations about the limited validity of certain uses of forensic evidence (NRC, 2009; see also NRC, 2003). The American Association for the Advancement of Science (AAAS) published a handbook for courts and forensic experts that defines objective criteria for identifying qualified expert witnesses and advances guidelines for testimony (AAAS, 2002).

[1] The authors recognize that the impact of scientifically invalid testimony on the outcome of trials can vary. Examining the extent of impact of scientifically invalid testimony on the outcome of criminal trials is beyond the scope of this study.

### *Sources of Scientifically Invalid Testimony on Forensic Evidence*

In adversarial justice systems like those in the United States, many factors can threaten the validity of forensic expert testimony. These include lawyer-driven factors, such as cherry-picking favorable experts and coaching or "woodshedding" witnesses toward a desired conclusion, as well as expert-driven factors, such as outright adversarial allegiance, random drift from the consensus view, and other unconscious cognitive biases—such as the tendency to favor outcomes that confirm one's preexisting intuitions (Orne, 1962; Rosenthal, 1966; Sheppard and Vidmar, 1980).

The standard legal response to such threats to the validity of forensic expert testimony, held by the U.S. Supreme Court itself, is that these sources of error are effectively neutralized by the structure of the adversarial system: opposing viewpoints are pitted against each other, cross-examination of these viewpoints is permitted, and fact-finders are empowered to weigh their relative merits (see Robertson, 2010). However, research has suggested that the adversarial system may not be sufficient to neutralize biased forensic testimony (Cain, Loewenstein, and Moore, 2005; Gitlin et al., 2004; Levett and Kovera, 2008; Otto, 1989; Robertson and Yokum, 2012; Sheppard and Vidmar, 1980; Simon, Stenstrom, and Read, 2009). That system strongly assumes that errors on both sides will be equivalent in magnitude, that the statistical average of these errors will represent a truer state of affairs, and that fact-finders will tend to resolve such errors using decision rules that approximate this so-called averaging rule. However, these assumptions have not been adequately tested. To the contrary, substantial literature in jury bias suggests that jurors are easily swayed by irrelevant scientific testimony (Aharoni and Fridlund, 2013; Gurley and Marcus, 2008; McCabe and Castel, 2008; Weisberg et al., 2008). And, in the face of an evidentiary disparity, they may effectively ignore the strengths and weaknesses of both parties' testimonies and favor the more confident, likable party or simply go with their pre-existing intuition (Ivkovic and Hans, 2003; Levett and Kovera 2008).

In line with the recommendations of the National Academy of Sciences, we review evidence of the two ways to improve the validity of forensic expert testimony that we will explore in this study: (1) *expert blinding*, a procedure in which experts are not told whether they are testifying for the defense or the plaintiff/prosecution, and (2) expert consensus feedback.

#### Expert Blinding Evidence

Experimental evidence has consistently shown that known adversarial party representation elicits preferential changes in witness testimony. For example, in an experimental vignette study, clinical psychology graduate students were asked to provide an expert opinion about a defendant's insanity either on behalf of the prosecution or the defense (Otto, 1989). Participants assigned to the prosecution were significantly more likely to recommend a guilty verdict, whereas those assigned to the defense were more likely to recommend not guilty by reason of insanity. In a more recent study by Murrie et al. (2013), the investigators hired forensic psychologists and psychiatrists to score risk assessments for the same offender case files but

manipulated their ostensible party representation. Those who believed they were consulting for the prosecution assigned higher risk scores than the ones submitted by those who believed they were consulting for the defense.

In another study, participants played the role of investigator in a case of academic misconduct (Simon, Stenstrom, and Read, 2009). In one condition, participants were assigned to either of the two adversarial parties. In a second condition, they took on the role of a neutral third party. The study found that participants assigned to an adversarial party recounted facts of the case that were polarized toward their representing party compared with nonadversarial participants. Moreover, verdict judgments about the occurrence of misconduct confirmed this partiality.

A recent study of mock jurors presented participants with a staged video of a medical malpractice trial containing testimony from two medical experts (Robertson and Yokum, 2012). One of these experts was blinded to party representation. The other was known to represent either the plaintiff or defense. When the plaintiff's expert was blinded, the odds of a verdict that was favorable to the plaintiff significantly increased relative to the unblinded plaintiff. Likewise, when the defense's expert was blinded, a verdict favorable to the defense was more likely. In addition, participants perceived blinded experts as more credible than unblinded experts.

The studies described above suggest that blinded and otherwise nonadversarial conditions produce testimonies that are different from unblinded, adversarial ones—but they do not, by themselves, speak to whether this testimony is any more scientifically accurate. In a role-playing study by Sheppard and Vidmar (1980), undergraduate participants bore witness to a physical altercation, then were interviewed by either an "adversarial" or "nonadversarial" lawyer. Using objective measures of testimony bias, they found that witnesses who were retained by the defense distorted the facts of their testimony to the defendant's favor, while experts retained by the prosecutor or by a nonadversarial lawyer showed less bias. They suggest this asymmetry could be due to the factual nature of the evidence being tilted toward culpability, which would have favored the prosecution at baseline.

Other studies have examined the impact of expert blinding under real-world conditions using true scientific experts. In a study of radiologists, investigators examined rates of positive diagnoses for asbestos exposure made by physicians hired by the plaintiff and compared them with the findings of blinded radiologists (Gitlin et al., 2004). Unblinded experts exhibited a dramatic increase in positive diagnosis relative to blinded experts.

Finally, researchers in the United Kingdom investigated the likelihood that fingerprint examiners would produce false-positive matches depending on whether they were blinded to extraneous information about the case (Dror, Charleton, and Peron, 2005; Dror and Charleton, 2006). The researchers identified a set of fingerprints that had previously been examined and assessed by fingerprint experts under standard blinded conditions. At a later date, the researchers presented these same experts with the same prints again (unbeknownst to the experts), but before soliciting the experts' second assessment, a confederate suggested that the matches were known

to be false positives. Consequently, a majority of the examiners' assessments changed from a "match" judgment to a "no-match" judgment. Although this study did not set out to measure effects of legal party representation, it nicely illustrates the susceptibility of real-world experts to confirm outcomes suggested by contextual, social primes.

Taken together, existing research suggests a fairly consistent confirmatory allegiance effect that produces evidence that most favors the hiring party, and a likely effect of expert blinding on the validity of expert testimony. If blinding experts and litigants to the hiring party affiliation does improve testimony accuracy, exactly how this procedure could be implemented is uncertain. One intriguing proposal (Robertson, 2010) relegates oversight duties to an independent, intermediary agency such as the AAAS. In this model, litigants could commission blind review by an expert randomly assigned from an organizational roster and whose identity and communications are mediated by that organization. If the expert's findings are unfavorable, the litigant is not required to disclose them; if they are favorable, the litigant may introduce them in discovery like any other evidence. Prior to the deposition, the blind is lifted (Robertson, 2010, pp. 206–214). While this approach has garnered initial empirical support in civil contexts (e.g., Robertson and Yokum, 2012), it has yet to be examined in criminal settings, for which the burden of proof is always on the prosecution and the standard of proof tends to be higher, among other differences. Additional research with true scientific experts will be needed to build confidence in the conclusion that expert blinding in fact produces more accurate testimony than explicit adversarial representation, and that this effect occurs in the context of criminal forensic evidence.

### Expert Feedback

Another potential way to reduce evidentiary validity problems, which was prefaced by the NAS report on expert testimony, involves solicitation of expert consensus feedback (NAS, 2002). Modern consensus feedback methods such as those based on the classic Delphi process are generally based on the principle that judgments arrived at by a panel of experts using a structured process are more likely to reflect the truth than those by unstructured groups or individuals.[2] The Delphi expert feedback consensus method was developed by the RAND Corporation in 1949 as way to systematically derive a consensus position from experts to forecast the impact of technology on warfare. Variations of the technique were used throughout the next six decades by RAND Corporation researchers to coalesce expert opinions on complex issues such as evaluating educational innovations (Helmer-Hirschberg, 1966) and assessing preventative measures for drug abuse (Thompson, 1973). Later, Taiwan used the Delphi method to prioritize their information technology industry (Madu, Kei, and Madu, 1991) and the National Cancer Institute used a Delphi variation to gain insights into funding alternatives (Hall et al., 1992). More recently, electronic versions of the technique were used to evaluate

[2] See Wong (1998) for description of the classic Delphi process and its evolvement over the last 60 years.

collaborative opportunities for the Army (Wong, 1998), estimate the number of lines of code for a software development (Smith, 2003), and prioritize research needs for the National Institute of Justice (Wong, 2011). A common thread among these applications is the recognition that consensus opinion among experts produces a better foundation upon which to base next-step actions.

As exemplified by these examples, which span more than half a century, scientific and humanities researchers seek to gain the perspectives of all stakeholders to make informed decisions based on all available knowledge. Expert consensus building techniques are particularly suited for situations where the issues are complex, actions may have far reaching consequences, or meaningful communication among stakeholders is limited or based on different assumptions. Each stakeholder must balance his focused interest with the need to interact with others, and actions require awareness of others' needs and views (Wong, 2012). Researchers note that Delphi-based consensus methods are appropriate when decisions affect strong factions with opposing preferences (Cline, 2000). A criminal courtroom scenario where expert testimony can play a critical role in a jury's decision appears to have these characteristics. A series of exercises conducted by the RAND Corporation show that experts tend to give significant consideration to viewpoints of their peers when rendering their own judgment (Wong, 2003). This observation, combined with the opposing factions characteristic of criminal trials and the NAS expert witness report—which notes that consensus in the scientific community about a particular question is unlikely to appear in the courtroom (NAS, 2002)—suggests that scientific-based evidence such as expert testimony on forensic evidence might benefit from an expert consensus building–based approach.

Expert panels in consensus building exercises need not be large. Researchers have noted that ten experts could be enough in homogeneous cases (Delbecq, Van de Ven, and Gustafson, 1975), and the majority of Delphi applications have used between 15 and 20 participants (Ludwig, 1997). Electronic versions of the technique have simplified its administration and greatly reduced costs and logistical elements (Wong, 2003). Although recent advances in electronic communications have made it easier to conduct expert consensus based exercises with larger panels, expert panels in recent exercises have remained fairly constant (Wong, 2003; Silberglitt et al., 2004). The reduction of logistical challenges afforded by electronic versions of expert consensus building techniques has greatly reduced the time required to conduct such exercises. Recent demonstrations have shown that a four-round consensus building exercise can be accomplished in two to three hours—the time span of a typical business meeting (Wong, 2003). Hence, expert consensus building approaches appear to be applicable, feasible, and promising for improving the validity of forensic testimony, but have yet to be explored.

### *Base Rate Reasoning*

The potential effects of expert blinding and consensus feedback are likely to depend on the exact type of evidentiary problem at hand. Many evidentiary issues necessarily rely on subjective

or open-ended judgment where it is not possible to externally verify the basis of the conclusion because the facts of the matter are unknown. For example, the proposition that a defendant meets criteria for insanity does not have an objectively measureable true or false answer, so different experts may legitimately disagree. In such a case, there is no way to determine whether a cognitive bias exists. Objective measures, in contrast, such as a DNA match or a fluid dynamics calculation, permit researchers to quantify and evaluate the presence and degree of bias in an expert's testimony, even if such examples are less common in typical legal cases.

We sought to develop a fairly general reasoning test—one that had an objective structure, would apply to a variety of experts, and was difficult enough to potentially elicit a high degree of errors. This approach would enhance understanding cognition more generally and its potentially broad impact in legal proceedings. A natural choice is an evidentiary scenario that requires reasoning about Bayesian probability. Bayes' theorem is a standard formulation of conditional probability theory, and a rich literature exists on common cognitive biases in Bayesian reasoning among both laymen and experts. For example, when asked to estimate the probability of a particular event occurring, given the observance of particular prior instances (e.g., the probability of cancer, given a particular base rate of true and false cancer diagnoses in a broader population), research has shown that responses to such inquiries are not random, but are clustered and indicative of a small number of specific, often erroneous, cognitive strategies. One of the most frequent erroneous strategies is to over-weight the diagnostic information, such as the true positive rate, while under-weighting other relevant factors, such as the relative population sizes from which the true and false positive rates derive. This phenomenon, known as "base rate neglect," has been observed among both students (Bar-Hillel, 1980; Kahneman and Tversky, 1973; Nisbett and Borgida, 1975) and professionals (Ayanian and Berwick, 1991; Bakwin, 1945; Casscells, Schoenberger, and Graboys, 1978; Eddy, 1984). In the methods section, we discuss how this phenomenon was exploited in the design of our study stimuli.

### *Study Approach*

The purpose of this study is to evaluate the ability of expert blinding and consensus feedback to improve the validity of expert testimony, specifically in the context of forensic science. Expert blinding is a procedure in which experts are not told whether they are testifying for the defense or the plaintiff/prosecution, thus reducing partisanship. The impact of blinding on testimony accuracy has not yet been examined in criminal contexts.

Expert consensus feedback is a method for potentially reducing expert drift by culling expertise from multiple sources, feeding the majority view back to the individuals, and repeating this cycle to move toward a group consensus. This technique has been successfully used in a variety of disciplines, but has not been investigated in a courtroom scenario. Empirical support for this effect carries the potential to reduce wrongful convictions and their associated personal, financial, and societal costs.

The study consisted of developing a hypothetical criminal case with forensic evidence, carrying out an exercise with a panel of relevant experts to generate a consensus interpretation of the case evidence, and then conducting an experiment in a new sample of scientists using the consensus interpretation. Participants in the experiment were randomly assigned to one of three types of party representation: prosecution, defense, or blind (where blind means the participant was not told which party was retaining his/her expertise). Approximately half the participants in each representation received the expert panel consensus feedback in advance of providing a response, and the other half did not. In addition, the group without advanced feedback received an opportunity to change their baseline response following receipt of consensus feedback (a pre/post "within-subjects" measurement). Our primary hypothesis was that blinded expert response would be more accurate than one or more of the known adversarial conditions (prosecution or defense) and these differences would be greater without exposure to the consensus feedback.

# 2. Methods and Data Collection

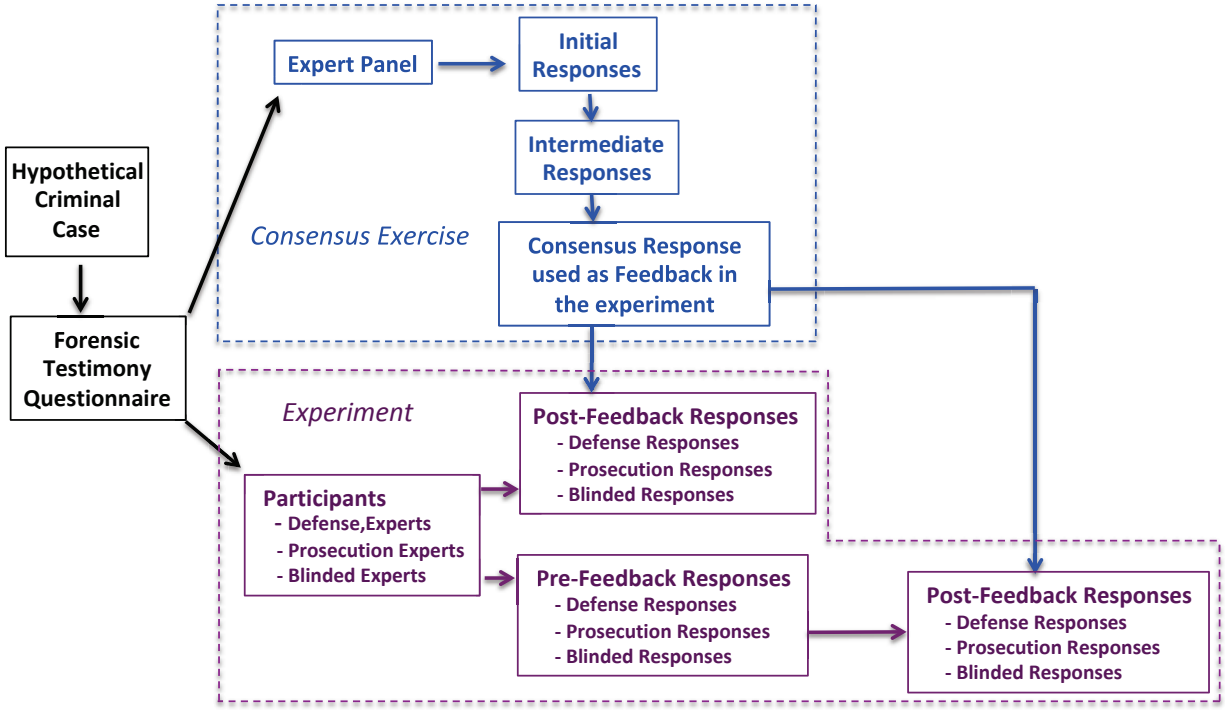
## Methods

The purpose of this study was to examine two potential contributors to biased testimony within adversarial litigation involving forensic evidence: (1) experts' knowledge of their party representation (i.e., prosecution vs. defense counsel), and (2) lack of consensus feedback from the relevant scientific community. The methods employed to address these questions implemented the following basic steps:

1. Develop a hypothetical criminal case with forensic evidence
2. Generate consensus interpretation of evidence with panel of relevant experts
3. Conduct between-subjects experiment with at least 300 scientists in relevant fields

Our first step was to generate a hypothetical criminal case with a mathematically correct answer to the forensic question. The case description with the forensic question was included as the sole question in the questionnaire for consensus exercise. The case description with the forensic question was also included in the questionnaire for the experiment. Our next step was to conduct a consensus exercise with a group of experts to generate a group consensus answer to the forensic question. After that, we conducted a between-subjects experiment with a within-subjects component, where about 50 percent of the participants were asked to answer the forensic question before knowing what the group consensus answer was (pre-feedback) and the other 50 percent answered with advance knowledge of the group consensus answer (post-feedback). The pre-feedback group was given an opportunity change their answer to the forensic question after their pre-feedback response was recorded. The relationships among the study components are depicted in Figure 2.1. We then discuss the components in detail.

**Figure 2.1. Study Procedure**




## *Hypothetical Criminal Case*

We aimed to construct a case with a solution based on theories that would be familiar to a target audience of professionals with advanced degrees in scientific fields.

The hypothetical case developed for the study involves criminal trespass and grand theft larceny of property belonging to a private business. The forensic question in the case is the probability that the defendant used the stairwell. The case stimuli included facts that would logically support a single solution to the forensic question. The solution could be derived using Bayesian techniques that are familiar to those with experience or academic training in a variety of physical and social science fields such as mathematics, statistics, chemistry, psychology, economics, criminology, and law.

### Case Stimuli

The text of the vignette is as follows.

> During preparation for the trial, you are informed that forensic engineers have developed and validated new facial recognition software for identifying specific individuals at the scene of a crime involving theft of company secrets. In their tests, video cameras were affixed inside the stairwell of a large, secured, private company building. The software attempted to match features from the recordings to an electronic record of authorized personnel who entered the building or the stairwell using electronic identification (ID) badges. The recognition software

was designed to provide an added level of security for situations in which an ID badge might have been misused. You are given three additional pieces of information:

- Of all of the people who occupied the building, 10 percent of those people were estimated to have used the stairwell during that time period.
- Of those personnel in the stairwell, the software correctly identifies that they were in the stairwell 99 percent of the time.
- Of those personnel NOT in the stairwell, the software falsely identifies them as being in the stairwell 11 percent of the time.

You are informed that the software classified the defendant as having used the stairwell during that time.

Forensic Question: What is the percent probability that the defendant used the stairwell during the time in question?

0% - - - - - - - - - 100%

### *Eligibility Criteria for Consensus Exercise and Experiment Participants*

Since the solution to the forensic question is readily derivable using Bayesian techniques familiar to those with experience or academic training in a variety of physical and social science fields, the eligibility criteria for participation in the consensus exercise or the experiment were an advanced degree in a relevant scientific field where training would normally include exposure to Bayes' theorem or equivalent experience.

The RAND employees invited to participate on the expert panel were identified using the search function of the RAND internal People Finder capability. We searched for membership in one or more of the 14 professional scientific societies we describe next, then culled our resulting list to include only those with doctorates. Eighteen employees were invited to participate, but none were informed about the goals or purpose of the study before the invitations were issued. Six declined or failed to respond. Examining the range of expertise and experience of the individuals who accepted showed that these 12 constituted a balanced expert panel.

Fourteen professional scientific societies were identified as having membership requirements likely to meet our criteria. Eligible societies were those with a focus on social/behavioral sciences or statistics and with a primarily U.S. membership. Nine societies agreed to distribute the invitation to participate in the experiment to their membership lists via e-mail or newsletter or to post our invitation on their websites. This approach enabled the investigators to collect responses anonymously. Six hundred and eighty five individuals accessed the experiment questionnaire and provided consent; 580 of these submitted a completed form. The reasons for attrition after consent are not known, but we consider their potential impact on sample representativeness in the Discussion and Limitations section of Chapter 4.

### *Questionnaire*

The case stimuli, along with the forensic question, formed the basis of the forensic testimony questionnaire for both the consensus exercise and the experiment. In the case of the consensus exercise, the questionnaire consisted of the case stimuli and the forensic question only. For the experiment, supplemental questions were added to ascertain personal elements, such as how confident the participant was in his/her response to the forensic question, and demographic characteristics, such as academic degree and experience serving as an expert consultant in criminal cases. The questionnaires are publicly available on the National Archive of Criminal Justice Data (2010).

### *Consensus Exercise*

The Electronic Decision Enhancement Leverager plus Integrator (E-DEL+I)[3] technique was used to conduct the consensus exercise. E-DEL+I is an iterative technique for integrating the science of expert opinion in research and analysis to enhance informed decisionmaking through structured consensus building. E-DEL+I has four primary components:

1. An expert panel whose collective knowledge base spans the issues being addressed comprises the participants in an E-DEL+I exercise.
2. A metric is developed to assess dimensions critical to the issue.
3. A questionnaire is formulated and used to solicit the independent assessments from the expert panel members.
4. A standard for consensus that is higher than a simple majority is specified for achievement of a consensus position.

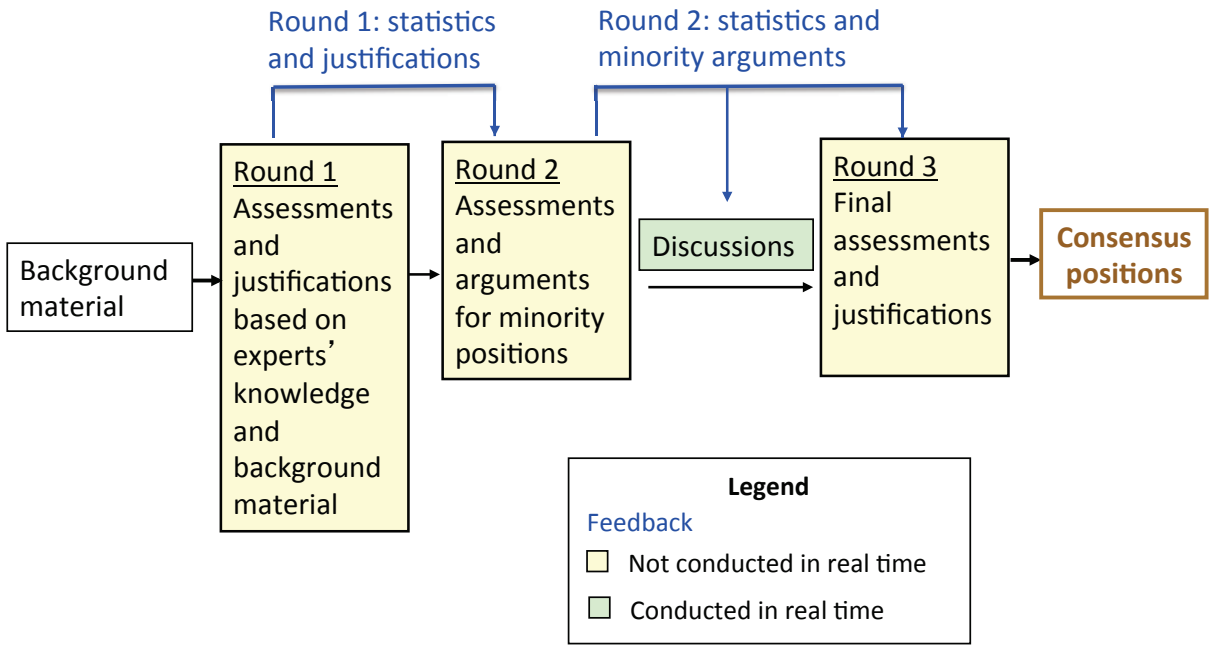
In an E-DEL+I exercise, a balanced panel of experts is selected and provided with background material and the questionnaire in electronic format by email or on a dedicated protected website. In the first round of the exercise, each panel member submits assessments (responses to the questions on the questionnaire) and justifications for the responses. An exercise coordinator computes the relevant statistical feedback, synthesizes the justifications, indicates the dominant position of the panel, and identifies any issues where the panel has reached consensus. Relevant statistical feedback can include the percentage of experts specifying each response or centralization statistics such as the mean, mode, or median response. This de-identified material is sent with the same questionnaire to the expert panel to start the second round. In Round 2, each expert gives the feedback material any consideration he or she deems appropriate and completes the same questionnaire justifying any position that differs from the dominant position. The coordinator computes the relevant statistics, synthesizes the minority position arguments, announces the Round-2 dominant position, again indicates where consensus has been achieved,

[3] RAND was awarded two provisional patents for the E-DEL+I processes with Dr. Carolyn Wong listed as the inventor on both provisional patents in Patent File 6847136P2, U. S. Patent Office, Washington D.C., 2001.

and sends this feedback back to the expert panel along with another copy of the questionnaire. The coordinator then conducts a real-time moderated discussion of the issues that have not yet achieved consensus. At the end of the discussion period, panel members are instructed to weigh all of the feedback and discussions they as deem appropriate to make a decision and submit their final responses. The consensus or dominant position of the Round-3 responses is the expert panel's assessment of the issue.

Figure 2.2 shows the components of the E-DEL+I technique.

**Figure 2.2. E-DEL+I Consensus Building Technique**

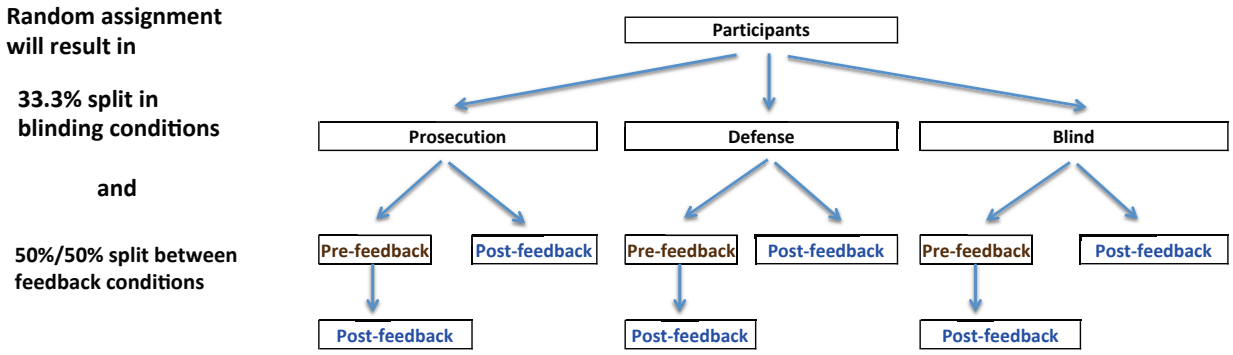



### *Experiment*

We conducted a between-subjects experiment with a within-subjects component. The design of the experiment randomly assigned participants among six independent groups defined by pairing the three forms of party representation (prosecution, defense, and blind) with the two feedback conditions (pre-feedback where the participants were not informed of the consensus interpretation before being presented with the forensic question and post-feedback where the participants were told what the consensus interpretation was before they were presented with the forensic question). For the within-subjects component, the group that provided a response before receiving feedback (pre-feedback) was given the opportunity to provide a different answer after their initial responses had been recorded. The random assignment produced a relatively even split among the various subsamples. In terms of party representation, there were three possibilities (prosecution, defense, and blind), so random assignment resulted in approximately

one third of the participants being assigned to each of the three party representations. Random assignment also resulted in approximately half of each party representation being assigned the pre-feedback and post-feedback conditions. Figure 2.3 shows the design of the experiment.

**Figure 2.3. Design of Experiment**




## Data Collection

### *E-DEL+I Consensus Building Exercise*

The blind collaborative justice consensus exercise was conducted over a ten-day period in February 2014. The questionnaire used for the exercise consisted of the case description and the forensic question only.

#### Expert Panel

The expert panel consisted of 12 RAND employees, each of whom held a doctorate in a relevant field where expertise in Bayes' theorem application would be the norm. In this case, the expert panel collectively held doctorates in behavior decision theory, biostatistics, clinical psychology, criminal justice, economics, experimental forensic psychology, industrial and organizational psychology, social psychology, psychology, and statistics. The panel's areas of research specialization included criminal and civil justice, economics, environment, defense, health, and social communications. Four panel members had one to five years of professional research experience, one had ten to 15 years of experience, and seven members had more than 15 years of research experience.

#### Round 1

In the first round of the consensus exercise, ten of the 12 responses argued for the correct answer (50-percent probability that the defendant was in the stairwell). Two Round-1 responses

included arguments for incorrect answers. Both of these incorrect responses included erroneous applications of statistical theory.

#### Round 2

All reasoning arguments (correct and incorrect) were sent back in de-identified form to the panel for the second round of the exercise. The Round-1 dominant response (50-percent probability that the defendant was in the stairwell) and the mean, median, and mode of the Round-1 responses were also included as Round-1 feedback. In this case, the mean, median, and mode were all equal to 50-percent probability that the defendant was in the stairwell. In Round 2, the ten panelists who submitted correct responses in the first round made no changes to their responses. The comments received for the second round included the observation that the text arguments provided clear logic and were just as effective as the mathematical proofs submitted in support of Round-1 responses. The Round-1 feedback led to unanimous consensus. Of the two panelists who submitted incorrect answers in the first round, one was convinced by Round-1 responses to submit the correct response for Round 2, and the other provided the correct response and valid argument without reference to Round 1 feedback.

#### Consensus Interpretation of Evidence

Since the exercise resulted in a unanimous correct answer backed by the same reasoning argument after two rounds, the discussion and third round were not held. The consensus exercise showed that the case was suitable for scientists with backgrounds in the physical as well as social sciences, and that scientists at all levels of experience may be qualified to answer the forensic question. The unanimous solution comprised the feedback for the experiment by appending the following paragraph to the end of the case stimuli for the post-feedback condition:

> In preparation for this survey, the study investigators asked a panel of 12 scientists with doctorates in relevant fields to discuss and arrive at a consensus about the best answer to the probability question above. Each panelist independently provided an answer along with an explanation of the logic used to arrive at his/her answer. Discussions of the reasoning arguments that explained the logic supporting the responses were conducted electronically and anonymously. No panel member was told who else was on the panel. The panel's final unanimous consensus response was 50 percent.

For participants initially assigned the pre-feedback condition, the following sentence was appended to the end of the above paragraph: "You now will have the opportunity to change your response or to leave it unchanged." The extended paragraph was presented to those assigned the pre-feedback condition after the participant's initial response was recorded.

### *Justice Experiment*

The experiment was designed and staged on SurveyMonkey, a web-based commercial tool for survey design and administration. The experiment was conducted April 11–30, 2014. It took

an average of 18 minutes for participants to complete the questionnaire. Following completion, respondents were invited to an independently hosted website form (SelectSurvey.net) to request compensation. Since compensation required collecting contact information, this approach enabled us to partition identifying information from responses on the questionnaire.

### Participants in the Experiment

Invitations to participate in the justice experiment were delivered by the nine professional scientific societies who agreed to distribute the invitation to their membership. Six hundred and eighty five individuals accessed the questionnaire and provided consent, of whom 580 submitted a completed form. The final sample was 51.9 percent female, 41.9 percent male, and 6.2 percent unreported. Mean age was 42.7 ($SD$ = 14.2). Ethnically, 3.3 percent identified as Hispanic or Latino, 82.8 percent were not Hispanic or Latino, and 13.9 percent were unspecified. Racially, 81.0 percent of the sample identified as white; fewer than 5 percent reported for each of the other races: American Indian or Alaska Native (1.2 percent), Asian (4.5 percent), black or African American (2.2 percent), Native Hawaiian or Pacific Islander (0.2 percent), mixed race (2.1 percent), other race (1.9 percent); 10.6 percent were unspecified. Most respondents had an advanced degree in the sciences: 36.2 percent reported a doctoral degree, followed by juris doctorate (4.1 percent), master's (33.3 percent), bachelor's (17.8 percent), and other/unspecified (8.6 percent). Degree fields represented physical sciences (including forensic science) (50.7 percent) and social sciences (36.2 percent), with the remainder classified as other/unspecified. More than half (57.2 percent) of the sample reported serving as an expert witness for a legal case; 38.3 percent had not served and 4.5 percent did not specify.

### Data Cleansing

The key dependent measure was derived from the probability (0 to 100 percent) that the defendant in the experiment scenario was in the stairwell at the time in question (i.e., the positive predicted value). It was defined as the arithmetic absolute difference in each respondent's probability score from the correct score of 50 percent. We refer to this calculated variable as our "error index." It is an inverse measure of response accuracy.

Other calculated variables included a dummy code describing whether the respondent was assigned to the pre/post or prior feedback condition. The question gauging the highest degree earned was consolidated into a dichotomous variable describing whether the respondent carried a doctoral degree or not. Similarly, the question about the number of times having served as an expert witness was consolidated into a dichotomous variable representing "any experience" versus "no experience."

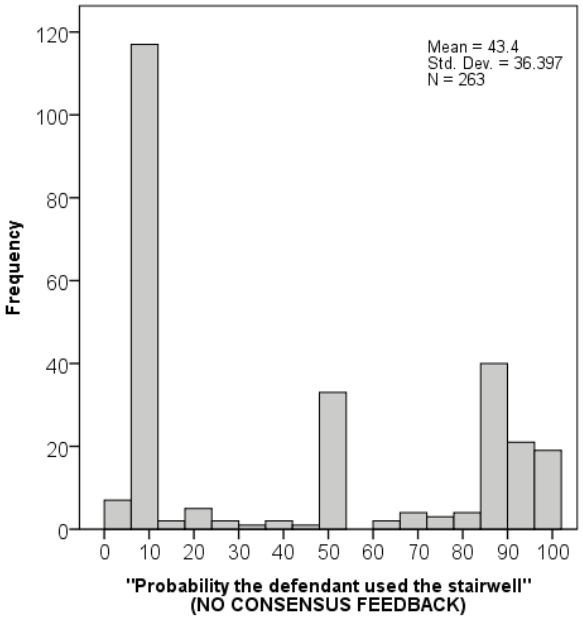
# 3. Analyses and Findings

## Distribution of Responses

First, we sought to understand the distribution of correct and incorrect responses to the probability question prior to introduction of consensus feedback (see Figure 3.1). Among respondents in this condition, 88.2 percent provided incorrect responses (greater or less than a score of 50 percent). On average, respondents estimated a 43.4 percent (*SD* = 36.4) probability that the defendant used the stairwell during the time in question (median = 25 percent, mode = 10 percent). However, such measures of central tendency are limited due to the multimodal nature of the distribution, which reveals possible evidence of discrete cognitive strategies. As expected, evidence of base rate neglect was found for approximately 30.4 percent of the sample in a modal response ranging from a score of about 85 percent to 100 percent. Scores in this range suggest a strategy of placing excessive weight on the diagnostic information (the true positive rate of 99 percent) while neglecting the low base rate of stairwell users.

The larger modal score of 10 percent was not anticipated. Individuals providing this response might have been cued by the prior base rate of people stipulated to be in the stairwell without regard for the proportion of positive tests that were true. In Appendix A, we report a test of our hypotheses with these individuals removed, and observed partial replication of the sample-wide effects.

**Figure 3.1. Distribution of Responses (No Consensus Feedback)**



## Association Between Pre-Feedback Accuracy and Demographic Characteristics

To examine whether pre-feedback accuracy could be explained or suppressed by an unpredicted association with basic demographic characteristics, we subjected each to a Pearson correlation with the error index. Neither age ($r$ [251] = 0.03, $p$ = 0.66) nor ethnicity ($r$ [251] = –0.08, $p$ = 0.24) was associated with the error index, but male respondents produced significantly smaller errors on average than females ($t$[249] = 2.92, $p$ < 0.01). Therefore, gender was entered as a control variable in our hypothesis tests to more clearly discern the independent effects of party representation and consensus feedback on error rates.

## Effect of Blinding and Consensus Feedback on Accuracy

Next, we examined whether erroneous responses differed by party representation. We first conducted a two-way analysis of variance (ANOVA) with party representation and feedback as between-subjects independent factors, controlling for gender (see Table 3.1). First, a main effect of feedback was found. As predicted, participants who received advanced consensus feedback exhibited significantly smaller error differentials than did those who received no advanced feedback, $F$ (1, 545) = 6.23, $p$ < 0.05, $MD$ = 3.51; $SE$ = 1.41. There was also a marginal main effect of the party representation manipulation, $F$ (1, 545) = 2.41, $p$ = 0.09. Pairwise comparisons using Fisher's least significant differences test (Fisher, 1935) revealed that participants who were informed that they were retained by the prosecution exhibited larger errors than those ostensibly retained by the defense ($MD$ = 3.66, $SE$ = 1.74, $p$ < 0.05). Errors in the blinded condition were not significantly different from that of the defense condition ($MD$ = 0.86, $SE$ = 1.70, $p$ = 0.63), but were marginally smaller than that of the prosecution condition ($MD$ = 2.83, $SE$ = 1.73, $p$ = 0.10). However, this test violated the assumption of homogeneity of variance, using Levene's test of equality of error variances, $F$ (5, 545) = 10.52, $p$ < 0.001, so we re-examined the direct effect of party representation using the Games-Howell procedure, which does not assume equality of variances. This test produced a similar marginally significant effect of party representation ($p$ = 0.08) with a tendency for the prosecution to exhibit larger errors than the defense ($MD$ = 3.78, $p$ = 0.07). There was no interactive effect of feedback by party representation, $F$ (2, 545) = 0.51, $p$ = 0.60.

**Table 3.1. Between-Subjects Comparison of Party Representation and Feedback Conditions**

| Feedback Condition | Party Representation Condition | Error Index (%) | Standard Error | 95% Confidence Interval |
|---|---|---|---|---|
| Pre-feedback | Defense | 32.56 | 1.79 | [29.03, 36.08] |
| | Blind | 33.63 | 1.76 | [30.16, 37.09] |
| | Prosecution | 34.82 | 1.82 | [31.26, 38.39] |
| Post-feedback | Defense | 28.28 | 1.60 | [25.15, 31.42] |
| | Blind | 28.87 | 1.62 | [25.58, 32.05] |
| | Prosecution | 33.34 | 1.73 | [29.93, 36.74] |

In addition to testing the between-subjects effect of feedback, we also sought to examine the within-subjects effect of feedback, namely the error differential before and after exposure to feedback. (The within-subjects approach is more sensitive in detecting small effects.) Using a two-way mixed ANOVA otherwise identical to the between-subjects model, a similar pattern of results emerged (see Table 3.2). We found that post-feedback error differentials ($M = 27.72$; $SE = 1.16$) were smaller than pre-feedback errors ($M = 33.68$; $SE = 0.91$), $F\,(1, 249) = 16.27$, $p < .001$. However, no significant effects were observed for party representation, $F\,(2, 249) = 0.49$, $p = 0.62$, or the interaction, $F(2, 249) = 0.50$, $p = 0.61$. This model did not violate the assumption of homogeneity of variances, using Levene's test ($p = 0.26$ for pre-feedback errors and $p = 0.56$ for post-feedback errors).

**Table 3.2. Comparison of Respondents Before and After Consensus Feedback for Each Party Representation Condition**

| Party Representation Condition | Feedback Condition | Error Index (%) | Standard Error | 95% Confidence Interval |
|---|---|---|---|---|
| Defense | Pre-feedback | 32.60 | 1.57 | [29.50, 35.70] |
| | Post-feedback | 27.57 | 2.02 | [23.60, 31.54] |
| Blind | Pre-feedback | 33.61 | 1.55 | [30.56, 36.65] |
| | Post-feedback | 26.41 | 1.98 | [22.51, 30.31] |
| Prosecution | Pre-feedback | 34.83 | 1.59 | [31.69, 37.96] |
| | Post-feedback | 29.19 | 2.04 | [25.17, 33.21] |

## Association Between Accuracy, Education Level, and Expert Witness Experience

To further explore the observed relationship between probability score accuracy and consensus feedback, we examined whether this relationship might depend on education level or expert witness experience. Using two-way mixed ANOVA comparing consensus feedback (pre

vs. post) and education level (advanced science degree vs. no advanced science degree), the expected main effect of consensus feedback was observed with accuracy increasing following the feedback, $F$ (1, 218) = 50.56, $p < .001$, $MD$ = 7.40 ($SE$ = 1.04). However, no interactive effect was found, $F$ (1, 218) = 1.42, $p$ = 0.23. Moreover, education level did not exert a main effect on accuracy, $F$ (1, 218) = 0.11, $p$ = 0.74.

An equivalent two-way ANOVA was constructed to test the moderating effect of expert witness experience (has provided witness consulting vs. has not). Again, consensus feedback exhibited the predicted main effect upon probability accuracy, $F$ (1, 247) = 39.12, $p < .001$, $MD$ = 6.06 ($SE$ = 0.97), but this effect was not dependent on expert witness experience, $F$(1, 247) = 0.02, $p$ = 0.88. Experience did exert a main effect on accuracy: Those reporting experience with expert witness consultation were less accurate than those reporting no experience, $F$(1, 247) = 9.05, $p < 0.01$, $MD$ = 5.75, $SE$ = 1.91.

Last, we examined whether those with advanced science degrees or expert testimony experience might report greater confidence in their (pre-feedback) response accuracy relative to those without such education or experience. Using separate independent-samples $t$-tests, those with an advanced degree reported significantly higher confidence in their response ($M$ = 1.00, $SD$ = 0.92) than those without an advanced degree ($M$ = 1.66, $SD$ = 0.94), $t$(249) = –2.77, $p < .01$. No such effect was observed for experience level, $t$(247) = –1.11, $p$ = 0.27.

Together, these tests suggest that those with greater education or experience did not perform significantly better in the probabilistic reasoning tasks even though the former group reported greater confidence in their response accuracy.

# 4. Conclusions and Closing Remarks

## Conclusions

This analysis produced two key findings:

1. Expert consensus feedback resulted in a performance improvement.
2. An advantage due to blinding was not observed.

### *Expert Consensus Improved Performance*

Expert consensus feedback regarding the correct response to the reasoning problem demonstrated the predicted significant effect on response errors: Delivery of feedback resulted in improved performance. This result is also consistent with previous research (Wong, 2003), and the fact that expert consensus building approaches have been used for over half a century in a variety of disciplines suggests that this approach could help reduce biased expert testimony in criminal cases.

Our finding that delivery of expert consensus feedback resulted in a performance improvement also indicates that presentation of expert consensus has the potential of decreasing the rate of erroneous responses. Eighty-eight percent of survey participants provided an erroneous response. Many of these matched the base rate, which may represent an unintended demand of the wording of the task. However, a substantial amount of these approximated the true positive rate provided in the hypothetical scenario, suggesting broad neglect of the base rate information, as predicted. This latter finding is highly consistent with classic research on Bayesian reasoning. Our own pattern of results could indicate that more than 40 years of research into the cognitive biases that distort probabilistic reasoning may not be reflected in the professional sector on a broad enough scale to improve the overall performance of scientific professionals en masse. These individuals continue to be susceptible to classic erroneous tendencies such as the base rate fallacy.[4] Further exploration of the effect of consensus feedback on accuracy did not find any evidence that Bayesian reasoning improves with expert witness experience or scientific education, yet those with an advanced science degree reported greater confidence in the accuracy of their responses. It is possible that this false confidence reflects an

[4] See Ayanian and Berwich, 1991; Bakwin, 1945; and Eddy, 1984. For instance, using a similarly structured reasoning problem, Bar-Hillel (1980) asked university students to estimate the probable color of a cab involved in a hit-and-run accident given a base rate of color options. She found that only about 10 percent of respondents reached the correct Bayesian answer while the majority of respondents based their answer on the diagnostic information (true positive rate) alone. A similar pattern of over-weighting the diagnostic information and under-weighting the base rate was found in a sample of Harvard Medical School physicians and students challenged to estimate the probability that a patient with a positive result for cyanosis actually had the disease (Casscells, Schoenberger, Graboys, 1978). Most participants said a true positive diagnosis was more likely than not, and only 18 percent provided the correct probability of 2 percent.

overly routinized or heuristic approach to problem solving with less regard for the particularities of the case. Although this study demonstrated that expert consensus feedback could be beneficial to reducing biased testimony, further research that examines approaches for delivery of expert consensus needs to be conducted to address whether social persuasion was responsible for the improvement, rather than recognition of the correct response. If improvements can be demonstrated even in the absence of social persuasion, additional questions about cost-effectiveness and implementation will need to be examined.

### *Advantage Due to Blinding Not Found*

We found modest evidence that an expert's mere knowledge of the hiring counsel (prosecution vs. defense) could influence probabilistic reasoning. Increased accuracy tended to be higher in the party least favored by the evidence (the defense, in this case), suggesting a greater motivation to evaluate the evidence critically. Although motivations for critical thinking provide a plausible explanation for this effect, additional research would be needed to determine whether such motivations consistently evoke more accurate results or simply more effort or creativity of approach. Finally, we did not find conclusive evidence that blinding experts to their party representation conferred an advantage—accuracy within this condition was only marginally greater than the prosecution condition and not greater than that of the defense.

Previous research includes several examples of studies, described in Chapter 1, showing that both laypeople and experts tend to produce testimony that favors the counsel that ostensibly hired them, suggesting that blinding such individuals to that information could potentially improve the quality of their testimony. There are several possible reasons these patterns were not observed. First, though the sample was sizable, it is possible that a true effect of blinding was too small to be detected. If true, this would suggest that blinding is unlikely to be an important focus for practical intervention. Another explanation might be the relative lack of ecological cues needed to activate an adversarial bias. In real-world circumstances, it is possible that such biases only result from highly persuasive cues, such as coaching by counsel or high levels of compensation. Indeed, studies that have used deception to collect testimony from forensic experts have shown substantial effects (e.g., Dror and Charleton, 2006; Murrie et al., 2013). Further research with larger samples and greater ecological validity could help determine whether such effects can be observed using vignette studies. Finally, it is possible that the blinding is no more effective at inspiring critical reasoning than the knowledge that the evidence is unfavorable to one's case. In this case, it would still be valuable to know whether blinding procedures could be of practical use for parties already favored by the evidence. Future research will be necessary to distinguish between these alternative possibilities.

The theoretical mechanism of the allegiance effect is that it is a reflexive, unconscious, confirmatory response to cues associated with the hiring party, in this case. As such, it may be applicable to expert witnesses as a whole. However, it is not inconsistent with conscious, deliberate favoritism toward a hiring party. It is assumed that conscious, deliberate favoritism

will draw on some of the same cues; however, this behavior likely describes only a subset of the expert witness population. Future research could examine whether the allegiance effects might be stronger among so-called hired guns compared with experts who explicitly renounce any party affiliation.

## Discussion and Limitations

The high level of attrition at the consent stage of the experiment raises questions about sample representativeness. Unfortunately, it is not possible to know the reasons for participant dropout. We speculate that the level of difficulty of the reasoning problem might have deterred some individuals from participating. If so, the true error rate of this population could be even higher than observed had these individuals been retained. Raising incentive pay could provide one potential method of increasing the response rate in future versions of this study.

This study utilized an objective measure (i.e., classification probability) to assess allegiance bias because it permitted us to quantify the extent of the potential bias and assess whether it could operate even in the face of a correct solution. The drawback of this choice is that many evidentiary questions for expert witnesses do not take this objective form, thus limiting the generalizability of the predicted effects. However, efforts to study *subjective* allegiance effects in expert testimony face a potentially more difficult problem: They cannot attribute those differences to cognitive bias, per se. To bridge this gap, the example of classification probability reasoning was selected because it can be solved using basic logical inference, thus carrying the potential to cut across multiple domains of expertise.

In guaranteeing anonymity for our participants, it was not feasible to examine differences between participant groups (the academic societies) because this would have required individual-level attribution. As a consequence, any group-level differences influencing our outcome measures (e.g., potential differences in education or experience) are unknown, and further research on this topic should consider techniques to overcome this limitation without compromising perceptions of privacy.

## Closing Remarks

Given the nascent evidence of blinding on accuracy of testimony in criminal contexts, it would be premature to recommend broad changes to criminal justice procedure at this time. Much more research is needed to demonstrate the real-world utility and feasibility of expert blinding. However, to the extent that this study's results can be replicated in more naturalistic trial settings, it suggests that greater focus on common threats to sound scientific reasoning (such as base rate neglect) and engagement of the broader scientific community could improve the

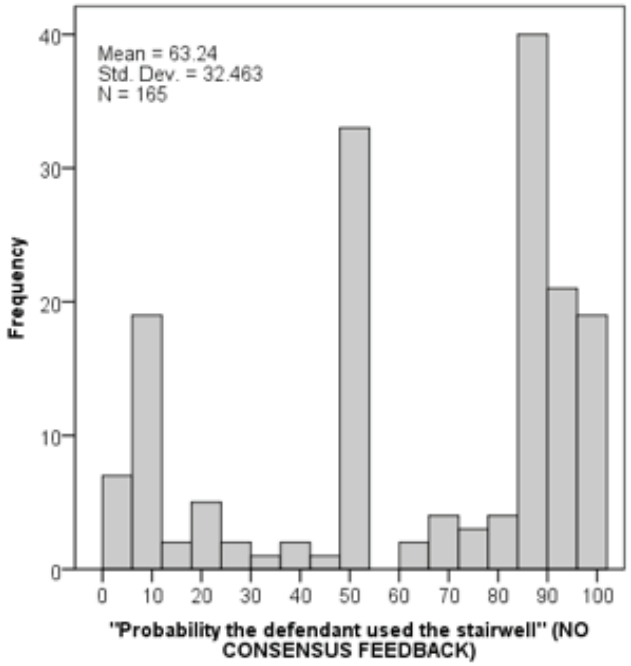
validity of expert testimony in court cases.[5] Such focus may be particularly important among those with expert witness experience and those who may be hired to testify on evidence that ostensibly favors the side that hired them. In addition, the effect of expert consensus exposure suggests that expert consensus feedback could be a promising way to reduce individual error in expert testimony. Future research is needed to confirm the effect of consensus exposure and to address practical and fairness issues associated with implementing expert consensus exposure approaches in a criminal justice environment.

[5] We experimented with an online means for providing such refresher education with a web application called TeachApp: Are You An Expert? (See the bibliography for the link.)

# Appendix: Supplemental Analysis

The high frequency of scores of "10 percent" suggested that these respondents might have believed they were being asked to report the defendant's prior probability of being in the stairwell (i.e., to simply restate premise 1 of the case stimuli, that 10 percent of the occupants were estimated to have used the stairwell) without regard to the accuracy of the identification software (premises 2 and 3). If so, this would constitute a misunderstanding of the instructions, which were to calculate the *posterior* probability of being in the stairwell given such priors. Although it is impossible to determine whether these participants misunderstood the instructions, it was possible to infer a basic understanding from their textual comments in answer to a subsequent question about the purpose of the study. Participants who submitted a score of "10 percent" but demonstrated an expressed recognition that the problem related to the keywords involving Bayesian probability, predicted values, or true or false positives were classified as demonstrating sufficient understanding of the instructions and were thus retained for analysis. Those who scored "10 percent" but made no such references were excluded from subsequent analysis. This approach resulted in a sample of 482, which we will call the reduced sample.

**Figure A.1. Distribution of Responses (Pre-Feedback), N = 482**




To examine whether pre-feedback accuracy could be explained or suppressed by an unpredicted association with basic demographic characteristics, we subjected each to a Pearson

correlation with the error index. However, age ($r$ [157] = 0.04, $p$ = 0.66), gender ($r$ [159] = –0.04, $p$ = 0.65), and ethnicity ($r$ [157] = –0.05, $p$ = 0.51) were not associated with the error index, suggesting that they need not be controlled in our hypothesis tests.

Next, we examined whether erroneous responses differed by the three party representation conditions. We first conducted a two-way ANOVA with party representation and feedback as between-subjects independent factors (see Table A.1). No effects on the error index were observed for party representation, $F(2, 476) = 1.94$, $p = 0.15$; feedback, $F(2, 476) = 0.01$, $p = 0.92$; or their interaction, $F(2, 476) = 0.24$, $p = 0.78$.

**Table A.1. Between-Subjects Comparison of Party Representation and Feedback Conditions (Reduced sample = 482)**

| Feedback Condition | Party Representation Condition | Mean Error Rate (%) | Standard Error | 95% Confidence Interval |
|---|---|---|---|---|
| Pre-feedback | Defense | 28.52 | 2.41 | [23.79, 32.25] |
| | Blind | 31.00 | 2.28 | [26.51 35.49] |
| | Prosecution | 31.84 | 2.48 | [26.98, 36.71] |
| Post-feedback | Defense | 28.16 | 1.69 | [24.85, 31.48] |
| | Blind | 29.46 | 1.66 | [26.20, 32.71] |
| | Prosecution | 33.22 | 1.83 | [29.61, 36.82] |

In addition to testing the between-subjects effect of feedback, we also sought to examine the within-subjects effect of feedback, namely the error differential before and after exposure to feedback (see Table A.2). (The within-subjects approach is more sensitive in detecting small effects.) Using a two-way mixed ANOVA otherwise identical to the between-subjects model, the predicted main effect of feedback was observed: Post-feedback error differentials were smaller than pre-feedback errors, $F(1, 159) = 24.01$, $p < .001$, ($MD = 5.73$; $SE = 1.17$). However, no significant effects were observed for party representation, $F(2, 159) = 0.09$, $p = 0.91$, or the interaction, $F(2, 159) = 1.14$, $p = 0.32$.

**Table A.2. Comparison of Respondents Before and After Consensus Feedback for Each Party Representation Condition (Reduced sample =482)**

| Party Representation Condition | Feedback Condition | Mean Error Rate (%) | Standard Error | 95% Confidence Interval |
|---|---|---|---|---|
| Defense | Pre-feedback | 28.302 | 2.385 | [23.59, 33.01] |
| | Post-feedback | 25.075 | 2.708 | [19.78, 30.42] |
| Blind | Pre-feedback | 30.695 | 2.261 | [26.23, 35.16] |
| | Post-feedback | 23.898 | 2.567 | [18.83, 28.97] |
| Prosecution | Pre-feedback | 31.720 | 2.456 | [26.87, 36.57] |
| | Post-feedback | 24.540 | 2.788 | [19.03, 30.05] |

# Bibliography


AAAS—*See* American Association for the Advancement of Science.

Aharoni, Eyal, and Fridlund, Alan J., "Moralistic Punishment as a Crude Social Insurance Plan," In Thomas A. Nadelhoffer, ed., *The Future of Punishment,* New York: Oxford University Press, 2013, pp. 213–229.

American Association for the Advancement of Science, *Court Appointed Scientific Experts (CASE), A Handbook for Judges*, Version 3.0**,** Washington, D.C., 2002.

Ayanian, John Z., and David M. Berwick, "Do Physicians Have a Bias Toward Action? A Classic Study Revisited," *Medical Decision Making,* Vol. 11, No. 3, 1991, pp. 154–158.

Bakwin, Harry, "Pseudodoxia Pediatrica," *New England Journal of Medicine,* Vol. 232, No. 24, 1945, pp. 691–697.

Bar-Hillel, Maya, "The Base Rate Fallacy in Probability Judgments," *Acta Psychologia*. Vol. 44, No. 3, 1980, pp. 211–233.

Burnam, M. Audrey, "Selecting Performance Measures by Consensus: An Appropriate Extension of the Delphi Method?" *Psychiatric Services*, Vol. 56, No. 12, December 2005, pp. 1583–1584.

Cain, Daylian M., George Loewenstein, and Don A. Moore, "The Dirt on Coming Clean: Perverse Effects of Disclosing Conflicts of Interest," *Journal of Legal Studies,* Vol. 34, No. 1, 2005, pp. 1–25.

Casscells, Ward, Arno Schoenberger, and Thomas Graboys, "Interpretation by Physicians of Clinical Laboratory Results," *New England Journal of Medicine,* No. 299, Vol. 18, 1978, pp. 999–1000.

Cline, Alan, *Prioritizing Process Using Delphi Technique*, Dublin, Ohio: Carolla Development Incorporated, 2000.

Cohen, Jacob, *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed., Hillsdale, N.J.: Erlbaum, 1988.

Delbecq, Andre L., Andrew H. Van de Ven, and David H. Gustafson, *Group Techniques for Program Planning.* Glenview, Ill.: Scott, Foresman, and Co., 1975.

Dror, Itiel E., and David Charleton, "Why Experts Make Errors," *Journal of Forensic Identification,* Vol. 56, No. 4, 2006, pp. 600–615.

Dror, Itiel E., David Charleton, and Ailsa E. Peron, "Contextual Information Renders Experts Vulnerable to Making Erroneous Identifications, *Forensic Science International,* Vol. 156, No. 1, 2005, pp. 74–78.

Eddy, David M., "Variations in Physician Practice: The Role of Uncertainty," *Health Affairs,* Vol. 3, No. 2, 1984, pp. 74–89.

Federal Rules of Evidence, Article VII, Rule 702, U. S. Government Printing Office, 2014.

Fisher, Ronald A., *Design of Experiments*, London: Oliver and Boyd, 1935.

FRE—*See* Federal Rules of Evidence.

Gitlin, Joseph N., Leroy L. Cook, Otha W. Linton, and Elizabeth Garrett-Mayer, "Comparison of 'B' Readers' Interpretations of Chest Radiographs for Asbestos Related Changes," *Academic Radiology,* Vol. 11, No. 8, 2004, p. 843.

Gould, Jon B., Julia Carrano, Richard A. Leo, and Joseph Young, *Predicting Erroneous Convictions: A Social Science Approach to Miscarriages of Justice*, Washington, D.C.: National Criminal Justice Reference Service, 2012.

Gurley, Jessica R., and David K. Marcus, "The Effects of Neuroimaging and Brain Injury on Insanity Defense," *Behavioral Sciences & The Law,* Vol. 26, No. 1, 2008, p. 85.

Hall, Nicholas G., John C. Hershey, Larry G> Kessler, and R. Craig Stotts, "A Model for Making Project Funding Decisions at the National Cancer Institute," *Operations Research*, Vol. 40, No. 6, 1992, pp. 1040–1052.

Helmer-Hirschberg, Olaf, *The Use of the Delphi Technique in Problems of Educational Innovations*, Santa Monica, Calif.: RAND Corporation, 1966.

Innocence Project, *Unvalidated or Improper Forensic Science,* undated. As of April 17, 2013: http://www.innocenceproject.org/understand/Unreliable-Limited-Science.php

Ivkovic, Sanja Kutnjak, and Valerie P. Hans, "Jurors' Evaluations of Expert Testimony: Judging the Messenger and the Message," *Law & Social Inquiry,* Vol. 28, No. 2, 2003, p. 441.

Kahneman, Daniel, and Amos Tversky, "On the Psychology of Prediction," *Psychological Review,* Vol. 80, No. 4, 1973, pp. 237–251.

Landis, J. Richard, and Gary G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics,* Vol. 33, No. 1, 1977, pp. 159–174.

Levett, Lora M., and Margaret Bull Kovera, The Effectiveness of Educating Jurors About Unreliable Expert Evidence Using an Opposing Witness," *Law and Human Behavior,* Vol. 32, No. 4, 2008, pp. 363–374.

Ludwig, Barbara, "Predicting the Future: Have You Considered Using the Delphi Methodology?" *Journal of Extension,* Vol. 35, No. 5, 1997, pp. 1–4. As of November 6, 2005:
http://www.joe.org/joe/1997october/tt2.html

Madu, Christian, Chu-Hua Kei, Assumpta Madu, "Setting Priorities for the IT Industry in Taiwan—A Delphi Study," *Long Range Planning*, Vol. 24, No. 5, 1991, pp. 105–118.

McCabe, David, and Alan Castel, "Seeing is Believing: The Effect of Brain Images on Judgments of Scientific Reasoning," *Cognition*, Vol. 107, No. 1, 2008, p. 343.

Murrie, Daniel C., Marcus T. Boccaccini, Lucy A. Guarnera, and Katrina Rufino, "Are Forensic Experts Biased by the Side That Retained Them?" *Psychological Science,* Vol. 24, No. 10, 2013, pp. 1889–1897.

NAS—*See* National Academy of Sciences.

National Academy of Sciences, *The Age of Expert Testimony, Science in the Courtroom, Report of a Workshop***,** Washington, D.C., 2002.

National Archive of Criminal Justice Data, homepage, 2010. As of April 8, 2015:
http://www.icpsr.umich.edu/icpsrweb/NACJD

National Institute of Standards, "Department of Justice and National Institute of Standards Announce Launch of National Commission on Forensic Science," Washington, D.C., February 15, 2013.

National Research Council. *The Polygraph and Lie Detection*. Committee to Review the Scientific Evidence on the Polygraph, Division of Behavioral and Social Sciences and Education. Washington, D.C.: The National Academies Press, 2003.

———, *Strengthening Forensic Science in the United States: A Path Forward*, Committee on Identifying the Needs of the Forensic Sciences Community, Washington, D.C.: Document Number 228091, August 2009.

Nisbett, Richard E., and Eugene Borgida, "Attribution and the Psychology of Prediction," *Journal of Personality and Social Psychology,* Vol. 32, No. 5, 1975, pp. 932–943.

NRC—*See* National Research Council.

Orne, Martin T., "On the Social Psychology of the Psychological Experiment," *American Psychologist,* Vol. 17, 1962, pp. 776–783.

Otto, Randy K., "Bias and Expert Testimony of Mental Health Professionals in Adversarial Proceedings: A Preliminary Investigation," *Behavioral Sciences & the Law,* Vol. 7, No. 2, 1989, pp. 267–273.

Paul, Christopher, Russell W. Glenn, Beth Grill, Megan McKernan, Barbara Raymond, Matthew Stafford, and Horacio R. Trujillo, "Identifying Urban Flashpoints: A Delphi-Derived Model for Scoring Cities' Vulnerability to Large-Scale Unrest," *Studies in Conflict and Terrorism,* Vol. 31, No. 11, 2008, pp. 981–1000.

Public Law 108–405, Title III, Sec. 306, National Forensic Science Commission, Oct. 30, 2004, 118 Stat. 2274.

Robertson, Christopher T., "Blind Expertise," *New York University Law Review,* Vol. 85, 2010, pp. 174–257.

Robertson, Christopher T., and David V. Yokum, "The Effect of Blinded Experts on Juror Verdicts," *Journal of Empirical Legal Studies,* Vol. 9, No. 4, 2012, pp. 765–794.

Roman, John, Kelly Walsh, Pamela Lachman, and Jennifer Yahner, *Post-Conviction DNA Testing and Wrongful Conviction,* Washington, D.C., Urban Institute, 2012.

Rosenthal, Robert, *Experimenter Effects on Behavioral Research.* New York: Appleton-Century-Crofts, 1966.

Sackman, Harold, *Delphi Critique: Expert Opinion, Forecasting, and Group Process*, Lexington, Mass.: Lexington Books, 1975.

Schneps, Leila, and Coralie Colmez, *Math on Trial: How Numbers Get Abused in the Courtroom*, New York: Basic Books, 2013.

Sheppard, Blair H., and Neil Vidmar, "Adversary Pretrial Procedures and Testimonial Evidence: Effects of Lawyers' Roles and Machiavellianism," *Journal of Personality and Social Psychology,* Vol. 39, No. 2, 1980, pp. 320–332.

Silberglitt, Richard, Lance Sherry, Carolyn Wong, Michael S. Tseng, Emile Ettedgui, Aaron Watts, and Geoffrey Stothard, *Portfolio Analysis and Management for Naval Research and Development*, Santa Monica, Calif.: RAND Corporation, MG-271-NAVY, 2004. As of April 9, 2015:
http://www.rand.org/pubs/monographs/MG271.html

Simon, Dan, Doug Stenstrom, and Stephen J. Read, "Adversarial and Non-Adversarial Investigations: An Experiment," West Coast Experimental Political Science Conference, May 15, 2009.

Smith, John, *The Estimation of Effort Based on Use Cases*, New York: Rational Software Corporation, 2003.

Thompson, L. T., *A Pilot Application of Delphi Techniques to the Drug Field*, Santa Monica, Calif.: RAND Corporation, R-1124, 1973. As of April 9, 2015:
http://www.rand.org/pubs/reports/R1124.html

Weisberg, Deena Skolnick, Frank C. Keil, Joshua Goodstein, Elizabeth Rawson, and Jeremy R. Gray, "The Seductive Allure of Neuroscience Explanations," *Journal of Cognitive Neuroscience,* Vol. 20, 2008, pp. 470–477.

Wong, Carolyn, *An Analysis of Collaborative Research Opportunities for the Army*, Santa Monica, Calif.: RAND Corporation, MR-675-A, 1998. As of April 9, 2015: http://www.rand.org/pubs/monograph_reports/MR675.html

———, *How Will the e-Explosion Affect How We Do Research?* Santa Monica, Calif.: RAND Corporation, DB-399-RC, 2003. As of April 9, 2015: http://www.rand.org/pubs/documented_briefings/DB399.html

———, *Facilitating Informed Decision Making: Consensus Building Using E-DEL+I*, Presentation at the Center for Discrete Mathematics and Theoretical Computer Science, National Science Foundation Science and Technology Center, Workshop on the Science of Expert Opinion, 2011.

———, *Using E-DEL+I to Integrate the Science of Expert Opinion in Informed Decision Making*, Presentation at 2012 INFORMS, Phoenix, Arizona, 2012.

Wong, Carolyn, and E. Aharoni, TeachApp: Are You An Expert? web application, 2015. https://smapp2.rand.org/surv4/TakeSurvey.aspx?SurveyID=m2436n7

Wong, Carolyn, Paul Steinberg, Kenneth Horn, and Elliot Axelband, "An Approach for Efficiently Managing DoD R&D Portfolios," *Acquisition Review Quarterly*, Fall 1998.