tapraid5/lhb-lhb/lhb-lhb/lhb00215/lhb0211d15z | xppws | S=1 | 4/16/15 | 16:32 | Art: 2014-2030

Law and Human Behavior

0147-7307/15/$12.00 http://dx.doi.org/10.1037/lhb0000134

Lay Understanding of Forensic Statistics: Evaluation of Random Match Probabilities, Likelihood Ratios, and Verbal Equivalents

William C. Thompson and Eryn J. Newman
University of California, Irvine

Forensic scientists have come under increasing pressure to quantify the strength of their evidence, but it is not clear which of several possible formats for presenting quantitative conclusions will be easiest for lay people, such as jurors, to understand. This experiment examined the way that people recruited from Amazon's Mechanical Turk (n = 541) responded to 2 types of forensic evidence—a DNA comparison and a shoeprint comparison—when an expert explained the strength of this evidence 3 different ways: using random match probabilities (RMPs), likelihood ratios (LRs), or verbal equivalents of likelihood ratios (VEs). We found that verdicts were sensitive to the strength of DNA evidence regardless of how the expert explained it, but verdicts were sensitive to the strength of shoeprint evidence only when the expert used RMPs. The weight given to DNA evidence was consistent with the predictions of a Bayesian network model that incorporated the perceived risk of a false match from 3 causes (coincidence, a laboratory error, and a frame-up), but shoeprint evidence was undervalued relative to the same Bayesian model. Fallacious interpretations of the expert's testimony (consistent with the source probability error and the defense attorney's fallacy) were common and were associated with the weight given to the evidence and verdicts. The findings indicate that perceptions of forensic science evidence are shaped by prior beliefs and expectations as well as expert testimony and consequently that the best way to characterize and explain forensic evidence may vary across forensic disciplines.

Keywords: Bayesian models, evidence, forensic science, jury decision making, probability

Supplemental materials: http://dx.doi.org/10.1037/lhb0000134.supp

Forensic scientists often compare items such as fingerprints, toolmarks, and shoeprints to determine whether they have a common source. They have traditionally reported their conclusions categorically. For example, fingerprint examiners traditionally reported one of three possible conclusions: either two prints were made by the same finger, or they were not, or the comparison was inconclusive (Cole, 2014). Forensic scientists who examine shoeprints, handwriting, tool marks, and bite marks have additional reporting categories (e.g., that two marks could have been, or probably were, or probably were not, made by the same item), but still use a limited number of categories to describe the nature and strength of their conclusions (Thompson & Cole, 2007).

Recently, forensic scientists have come under pressure to replace their traditional categorical conclusions with quantitative statements that incorporate probabilities or statistics (Redmayne et al., 2011; Koehler & Saks, 2010; NRC, 2009). Calls for quantification are driven partly by a desire to improve the scientific foundation of the forensic sciences. The National Research Council (NRC, 2009) has called for development of "measures of the accuracy of inferences made by forensic scientists" (p. 184) and declared that "[f]orensic science reports, and any courtroom testimony stemming from them, must include clear characterizations of the limitations of the analyses, including associated probabilities where possible" (p. 186).

Commentators have also questioned the logic underlying forensic scientists' categorical conclusions. Forensic scientists are trained to determine whether two items share a set of characteristics. They might also be able to estimate the rarity of shared characteristics. It is a major leap, however, to go from the observation that two items share a rare set of characteristics to the conclusion that the items probably or definitely have a common source. Various commentators have questioned whether forensic scientists should be making this leap (Evett, 1998; Berger, 2010; Morrison, 2011; Robertson, Vignaux, & Berger, 2011). They argue that forensic scientists should avoid opining on the ultimate question of whether two items have (or probably have) a common source and confine themselves to commenting on the frequency of matching characteristics or on the conditional probability of the observed results if the items have (or do not have) a common source.

A third concern is that people who rely on forensic science to make important decisions—particularly lay jurors—may be confused or misled by forensic scientists' traditional characterizations of their findings. According to the NRC report, experts have used a variety of terminology to describe their categorical conclusions—including "match," "consistent with," "identical," "similar

William C. Thompson and Eryn J. Newman, Department of Criminology, Law, & Society, University of California, Irvine.

This research was supported by a grant from the University of California Laboratory Fees Research Program.

Correspondence concerning this article should be addressed to William C. Thompson, Department of Criminology, Law, & Society, University of California, Irvine, CA 92697. E-mail: william.thompson@uci.edu

AQ: au
AQ: 2

tapraid5/lhb-lhb/lhb-lhb/lhb00215/lhb0211d15z | xppws | S=1 | 4/16/15 | 16:32 | Art: 2014-2030

in all respects tested," and "cannot be excluded"—without agreeing on the precise meaning of these terms (NRC, 2009, p. 185). Moreover, researchers have questioned whether lay people interpret such terminology in the intended manner (McQuiston-Surrett & Saks, 2008, 2009).

Although pressure is growing for forensic scientists to abandon their traditional categorical testimony, there is no consensus as yet on what the new format should be. One option is to provide a numerical estimate of the probability that a "match" or "nonexclusion" would occur by coincidence—a random match probability (RMP). This approach is frequently used with DNA evidence (Kaye & Sensabaugh, 2011). On determining that a suspect's DNA profile "matches" or "is consistent with" the DNA profile of an evidentiary sample, DNA analysts typically use population data to estimate the probability of finding a "matching" or "consistent" profile in someone sampled randomly from various reference groups (e.g., U.S. Caucasians; African Americans; U.S. Hispanics). Although the numbers are often framed as probabilities, they can also be presented as natural frequencies—for example, the analyst might say that among U.S. Caucasians only one person in 10 million would "match" or would be "included as a possible contributor" to a DNA sample. In disciplines other than forensic DNA testing there has been far less research on match probabilities, but additional research of this type is likely to emerge in the future. Where research is unavailable, experts might be able to provide a rough estimate of the RMP based on their experience.

Another option is the use of likelihood ratios (LRs) to describe the strength of forensic evidence. When assessing whether two items have the same source, a forensic expert must consider two mutually exclusive hypotheses—H: that the items have the same source; and A: that the items do not have the same source. The expert must then consider the likelihood of the observed results (D) under the two hypotheses, relying either on empirical data or subjective judgment based on experience and training (Thompson, 2012). The expert then reports the ratio of those two likelihoods—that is, $p(\mathrm{D} \mid \mathrm{H})/p(\mathrm{D} \mid \mathrm{A})$—by identifying the two hypotheses and saying something like: "The probability of obtaining this evidence is x times higher under hypothesis H than under hypothesis A." (Robertson & Vignaux, 1995; Evett, 1998; Cook, Evett, Jackson, Jones, & Lambert, 1998; Buckleton, 2005; Morrison, 2011). Those who favor the use of likelihood ratios claim that they avoid two problems associated with categorical conclusions: (a) the need to describe the evidence according to a categorical system in which the boundaries between categories (e.g., "match," "inclusion," "identification," "inconclusive") are vague and arbitrary; and (b) the leap of logic needed to go from the expert's knowledge of the rarity of shared characteristics to conclusions about the probability that two items have a common source (Berger, 2010; Robertson, Vignaux, & Berger, 2011).

Yet another option is to convert likelihood ratios to nonquantitative expressions known as verbal equivalents (VEs) (Martire, Kemp, Watkins, Sayle, & Newell, 2013). Under this approach, the analyst computes a likelihood ratio but then uses words rather than numbers to describe the strength of the evidence in accordance with a graduated scale, with higher likelihood ratios leading to stronger statements. A major proponent of this approach is the U.K.-based Association of Forensic Service Providers (AFSP, 2009). It recommends that analysts say the evidence provides "weak or limited" support for the favored hypothesis when the likelihood ratio is 1–10; "moderate support" when the likelihood ratio is 10–100; "moderately strong" support when the likelihood ratio is 100–1,000; "strong support" when it is 1,000–10,000; "very strong" support when it is 10,000 to 1 million; and "extremely strong" support when it is over 1 million. But recent research has raised questions about whether these labels convey the intended meaning to jurors (Martire et al., 2013; Martire, Kemp, Sayle, & Newell, 2014).

In the experiment reported here, we examine the way lay people respond to forensic science evidence when it is presented in one of the three formats just described (RMP, LR, or VE) in a hypothetical criminal trial. Our goal is to learn more about lay interpretations of statistical evidence of this type, and thereby to cast light on the strengths and weaknesses of various formats for communicating forensic science findings to lay audiences, such as jurors. Specifically, we hope to gain insight into how an expert's characterizations affect lay people's interpretation of forensic evidence and their ability to respond to it appropriately. We judge the appropriateness of people's responses according to three criteria that we will elaborate in the following sections: whether people's responses are sensitive to the strength of the forensic evidence, the logical coherence of their judgments about the evidence, and their susceptibility to drawing fallacious conclusions from the evidence.

Sensitivity to the Strength of Evidence

First, we examine how each presentation format affects people's sensitivity to the strength of the forensic science evidence. Because it is desirable that people give more weight to strong evidence than to weak evidence, presentation formats that render people insensitive to the strength of the evidence are obviously problematic. Following standard practice, we assess the weight that people give to forensic science evidence by measuring how much they change their estimates of the chances of a defendant's guilt after receiving the evidence.

Several past studies have examined lay people's sensitivity to variations in RMPs when evaluating forensic science evidence (e.g., a blood group match) in hypothetical criminal cases. These studies have consistently found that judgments of guilt varied appropriately (Faigman & Baglioni, 1988; Goodman, 1992; Smith et al., 1996). In every case people gave more weight to the forensic evidence (greater shifts in estimated chances of guilt) when the RMP was low than when it was higher. Whether people are also sensitive to variations in LRs and VEs is less clear. Martire and her colleagues have reported that people "were only weakly sensitive to large differences in evidential strength . . ." when evaluating LRs and VEs (Martire et al., 2013; see also Martire et al., 2014). But various methodological differences (particularly the scaling issues discussed later) might also account for the difference between the findings reported Martire et al. and earlier research. We aim to resolve this ambiguity in the literature by comparing the three presentation formats in the same experiment, holding other factors constant. Based on previous research, we hypothesize (Hypothesis 1) that people's verdicts and judgments of the chances of guilt will be more sensitive to variations in the strength of forensic evidence when it is presented in the RMP format than the LR or VE format.

tapraid5/lhb-lhb/lhb-lhb/lhb00215/lhb0211d15z | xppws | S=1 | 4/16/15 | 16:32 | Art: 2014-2030

Logical Coherence

Second, we examine the logical coherence of people's judgments about the forensic evidence—specifically, whether the weight they give to this evidence (as shown by shifts in their estimates of the chances the defendant is guilty) is consistent with their judgments about the probability of three key events that could falsely incriminate an innocent defendant: (a) a coincidental match; (b) a false report of a match attributable to laboratory error; and (c) a frame-up involving planting of incriminating evidence. Logic dictates that the weight people give the forensic evidence should be inversely proportional to their estimates of the probability that an innocent person could be falsely incriminated. The exact relationship between the two variables is specified by Bayes' rule (Lempert, 1977; Robertson & Vignaux, 1995). If there are logical inconsistencies between people's estimates of the probability that an innocent person could be falsely incriminated and the weight that they give to the forensic evidence, it may signal a misunderstanding of the evidence or the use of suboptimal (non-Bayesian) strategies for drawing inferences from the evidence. Consequently, if a particular presentation format leads to such inconsistencies it is cause for concern.

Research has generally found that simulated jurors are less responsive to forensic evidence than Bayesian models indicate they should be (Thompson & Schumann, 1987; Faigman & Baglioni, 1988; Goodman, 1992; Smith et al., 1996; Schklar & Diamond, 1999; Nance & Morris, 2002, 2005; Martire et al., 2013, 2014; for reviews of the early studies see Koehler, 2001; Kaye & Koehler, 1991; Thompson, 1989). But Thompson, Kaasa, and Peterson (2013) recently questioned whether jurors always underutilize forensic evidence relative to Bayesian norms. They examined the way that people recruited from a county jury pool reacted to DNA evidence with a very low RMP (1 in 1 trillion) when evaluating the guilt of a hypothetical defendant. They reported that jurors' judgments were "generally consistent with Bayesian expectations" and that jurors' judgments actually exceeded Bayesian norms (indicating that they were overvaluing DNA evidence) when the probability of a false match attributable to laboratory error was high. They suggested that earlier researchers may have underestimated the weight that people gave to forensic science evidence (relative to Bayesian norms) as a result of two methodological problems: (a) using incomplete Bayesian models that failed to take into account all possible sources of uncertainty; and (b) eliciting probability judgments using measures that restricted the range of responses and hence the degree to which people could change their responses after receiving forensic science evidence.

There is a striking discrepancy between the findings of Thompson et al. (2013), who reported that people respond to DNA evidence in a manner consistent with Bayesian norms, and the findings of Martire et al. (2013), who reported that people grossly undervalue shoeprint evidence relative to Bayesian norms. Our experimental design allows us examine three possible explanations for this discrepancy: (a) that people's reaction to forensic statistics depends on presentation format (RMP in Thompson et al., vs. LR and VE in Martire et al.); (b) that people's reactions depend on the type of forensic science evidence involved (DNA in Thompson et al., vs. shoeprint analysis in Martire et al.); and (c) that people's reactions depend on the way the researchers elicited probability judgments (a "log scale"—see Figure 1—in Thompson et al., vs. a statement of odds in Martire et al.). By simultaneously varying the presentation format, the type of forensic evidence, and the method for eliciting probability judgments, our experiment allows us to disentangle the effects of these variables.

___Certain to be guilty
___About 9,999,999 chances in 10 million that he is guilty
___About 999,999 chances in 1 million that he is guilty
___About 99,999 chances in 100,000 that he is guilty
___About 9,999 chances in 10,000 that he is guilty
___About 999 chances in 1,000 that he is guilty
___About 99 chances in 100 that he is guilty
___About 9 chances in 10 that he is guilty
___One chance in 2 (fifty-fifty chance) that he is guilty
___About 1 chance in 10 that he is guilty
___About 1 chance in 100 that he is guilty
___About 1 chance in 1,000 that he is guilty
___About 1 chance in 10,000 that he is guilty
___About 1 chance in 100,000 that he is guilty
___About 1 chance in 1 million that he is guilty
___About 1 chance in 10 million that he is guilty
___Impossible that he is guilty

Figure 1. Log scale for estimating chances defendant is guilty. AQ: 10

Because people often find it easier to reason with frequencies than with probabilities (Gigerenzer & Hoffrage, 1995; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000), we hypothesize (Hypothesis 2) that people's estimates of the chances of guilt will be more coherent (i.e., more consistent with Bayesian norms) when the evidence is characterized with RMPs than when it is characterized with LRs or VEs. Because DNA evidence has tremendous credibility (Lieberman, Carrell, Miethe, & Krauss, 2008), whereas shoeprint comparison is less well known, we also expect (Hypothesis 3) that DNA evidence will produce larger shifts in estimates of the chances of guilt than shoeprint evidence even when the statistics that the expert uses to characterize the strength of the evidence are the same. And because the log scale may make it easier to express high and low values and thereby facilitate shifts in judgment in response to the forensic evidence, we expect (Hypothesis 4) that people's judgments will be more responsive to forensic science evidence (greater shifts in estimated chances of guilt), when these judgments are elicited on the log scale than when elicited as statements of odds.

To address the questions raised by Thompson et al. (2013) about the adequacy of Bayesian modeling in earlier studies, we developed a Bayesian network model (discussed further in the Results section) that is more sophisticated and complete than the models used previously. It takes into account participants' own estimates of the probability that an innocent person could be incriminated as a result of coincidence, lab error, or a frame-up and it tells us how much each participant should change his or her estimate of the chances of the defendant's guilt (after receiving the forensic evidence) in light of those estimates. Comparing the actual shifts in participants' guilt judgments to the shifts specified by the model thus allows us to assess whether the weight participants give to the forensic evidence is logically consistent with

tapraid5/lhb-lhb/lhb-lhb/lhb00215/lhb0211d15z | xppws | S=1 | 4/16/15 | 16:32 | Art: 2014-2030

their estimates of the chances of a coincidental match, lab error, and frame-up and to make this assessment for each presentation format (RMP, LR, or VE), type of forensic evidence (DNA or shoeprint) and response measure.

Fallacious Reasoning

From a theoretical perspective, it will not be surprising to find discrepancies between the weight people give to forensic evidence and the weight specified by a Bayesian model. As Kahneman and Tversky famously declared: "In his evaluation of evidence, man is apparently not a conservative Bayesian: he is not Bayesian at all." (Kahneman & Tversky, 1972, p. 450). People employ a variety of heuristic strategies for evaluating evidence. These strategies generally work well although they can cause people to overvalue or undervalue evidence in specific situations (see generally, Tversky & Kahneman, 1974, 1982; Kahneman, Slovic, & Tversky, 1982; Gigerenzer & Hoffrage, 1995). For example, Koehler and his colleagues have shown that people give significantly less weight to evidence that the defendant "matches" the DNA profile of the perpetrator when the match probabilities are described in a manner that makes it easier to imagine that someone else could also match (Koehler, 2001; Koehler & Macchi, 2004).

But people sometimes evaluate forensic science using illogical strategies that arise from a fundamental misunderstanding of probabilistic evidence. For example, people sometimes mistakenly assume that they can infer the probability that matching items have (or do not have) a common source from the random match probability (RMP). If an expert reports that a defendant matches a DNA sample and that the probability a random person would match is 1 in 1 million, for example, people sometimes assume that this necessarily means there is one chance in a million that the DNA sample came from someone other than the defendant—a mistake of logic that has been called the "source probability error" (Koehler, 1993: Koehler, Chia, & Lindsey, 1995) and the "fallacy of the transposed conditional" (Evett, 1995). The conclusion is mistaken because, although the DNA evidence places the defendant among a relatively small group of potential contributors (1 person in 1 million), it cannot distinguish the defendant from other individuals in that group. (Consider that in a country the size of the United States, there may well be more than 300 people who would "match."). Consequently, the DNA evidence cannot, by itself, indicate the probability that the defendant, rather than another group member, is the source. In fact, the probability that the defendant is "not the source" may be higher or lower than the RMP depending on the strength of other evidence in the case. The danger of this fallacy is that it leads people to think they can determine the probability the defendant is (or is not) the source from the forensic evidence alone, without considering the other evidence.

The same erroneous logic (arising from transposition of conditional probabilities) might also lead to fallacious interpretation of likelihood ratios. If an expert says the DNA evidence is one million times more likely if the defendant, rather than a random person, is the source of a sample, for example, then people might mistakenly assume that this means it is one million times more likely that the defendant, rather than a random person, is the source of the sample.

In cases where the defendant's identity as the perpetrator is the sole issue, the prosecution can often prove the defendant's guilt by proving that he was the source of a sample left by the perpetrator. In such cases, people sometimes mistakenly equate the RMP with the probability the defendant is innocent—an error known as "the prosecutor's fallacy" (Thompson & Schumann, 1987; Balding & Donnelly, 1994; Nance & Morris, 2002, 2005; Kaye, Hans, Dann, Farley, & Albertson, 2007; Murphy & Thompson, 2010; de Keijser & Elffers, 2012; Thompson, Kaasa, & Peterson, 2013). The "prosecutor's fallacy" arises from the same transposition of conditional probabilities that underlies the "source probability error" (Thompson, 1989; Thompson, Taroni, & Aitken, 2003; Thompson, Kaasa, & Peterson, 2013). People equate the RMP with the probability the defendant is innocent because they equate the RMP with the probability the defendant is not the source of an incriminating sample while also assuming that if he is the source he must be guilty and if he is not the source he must be innocent.

When evaluating forensic science evidence people sometimes fall victim to another error called the "defense attorney's fallacy" (Thompson & Schumann, 1987; Thompson, Kaasa, & Peterson, 2013). Victims of this fallacy mistakenly assume that a forensic match has little or no probative value for incriminating the defendant if someone other than the defendant could also have matched. This error may cause people to underutilize forensic evidence, or ignore it entirely, when evaluating a case.

Does the presentation format (RMP, LR, or VE) affect people's susceptibility to drawing fallacious conclusions from forensic evidence? We make that assessment by presenting a series of correct and fallacious statements about the meaning of the forensic evidence and asking participants to tell us whether each statement is a correct interpretation of what was said by the expert witness who presented the evidence. This approach has been used by de Keijser and Elffers (2012) to detect fallacious understanding of statistical evidence by judges, lawyers and forensic scientists. Because the source probability error arises from confusion about the meaning of numbers, and no numbers are presented in the VE condition, we expect that the rate of agreement with statements consistent with the source probability error will be higher in the RMP and LR conditions than the VE condition (Hypothesis 5).

This experiment also allows us to check whether there is an association between fallacious interpretations and the weight that people give to the forensic evidence—an issue that has not previously been examined. Compared with other participants, we expect that those who fall victim to the source probability error (as shown by their agreement with statements consistent with that error) will be more likely to convict (Hypothesis 6a) and will give higher estimates of the probability of the defendant's guilt (Hypothesis 6b). Compared with other participants, we expect that those who fall victim to the defense attorney's fallacy (as shown by their agreement with fallacious statements) will be less likely to convict (Hypothesis 7a) and will give lower estimates of the probability of the defendant's guilt (Hypothesis 7b).

Method

Participants

We placed a solicitation on Amazon's Mechanical Turk (MTurk), an online labor pool, inviting people to participate in an

tapraid5/lhb-lhb/lhb-lhb/lhb00215/lhb0211d15z | xppws | S=1 | 4/16/15 | 16:32 | Art: 2014-2030

online jury simulation study for a fee of 70 cents. The invitation was available only to workers with IP addresses in the United States. We used a web utility to screen out respondents whose MTurk ID numbers had previously been used to participate either in this experiment or in other similar experiments administered by our research group. We also eliminated respondents who failed to affirm that they were American citizens at least 18 years of age, those who failed to successfully complete a series of screening questions designed to detect random or robotic responders, and those whose responses to screening questions indicated poor comprehension. The remaining respondents ($N = 541$) were a diverse group of jury-eligible adult Americans, and thus appeared suitable for our purpose (ages 18–74, $M = 34$; gender: 44% male, 49% female, 7% declined to state). The online supplementary materials for this article include a detailed demographic breakdown (Table S1) that compares our participants with a sample of actual jurors that Thompson et al. (2013) recruited from a county jury pool.

Procedure

Our experimental materials were mounted on a survey administration website (www.Qualtrics.com) which assigned each respondent randomly to an experimental condition, after which participants read the study materials. To assure comprehension, we posed questions about pertinent factual details at the end of each section. Participants who answered incorrectly were directed to read the relevant materials again. They had to respond correctly to every comprehension question to continue with the study. Most participants completed the study in 20 to 30 minutes.

Design

We used a between-subjects factorial design that varied the nature of the forensic evidence presented in the criminal case (DNA or shoeprint), the strength of the forensic evidence (very strong or moderate), and the presentation format that the forensic expert used to describe the strength of the forensic evidence (RMP, LR, or VE). In the very strong evidence condition, the RMP was 1 in 1 million, the LR was 1 million, and the VE was "very strong support." In the moderate strength condition, the RMP was 1 in 100, the LR was 100, and the VE was "moderate support."

To assess the impact of forensic evidence, we asked participants to render verdicts and estimate the chances of guilt at two points. The initial judgments occurred after they had read about the nonforensic evidence in the case but before they had received a full account of the forensic evidence. At this stage they were told that a forensic comparison (either DNA or shoeprint) had been attempted but had produced inconclusive results because the amount of evidence left at the crime scene (either a shoeprint or DNA sample) was too limited. After their initial verdicts and chance-of-guilt estimates were recorded, participants were asked to reevaluate the case under the assumption that the forensic testing had come out differently. They were then given a full account of the forensic tests, including the expert's statements about the strength of the findings (which varied across conditions in accordance with the experimental design).

We examined the shift caused by moving from inconclusive forensic evidence to incriminating forensic evidence, rather than the shift from no forensic evidence to incriminating forensic evidence, because we believed participants might draw negative inferences about the thoroughness of the investigation and hence the strength of the case if no forensic testing was reported. If that happened, any difference we observed between initial and final judgments could have been influenced by changes in participants' perceptions of the thoroughness of the investigation as well as by the strength of the forensic evidence. By making the forensic evidence inconclusive for the initial judgments we avoided creating that confound.

The experiment also varied the method used to elicit judgments of the chances of the defendant's guilt. When making initial judgments (based on the inconclusive forensic evidence), about half of subjects used the same measure that was used by Martire et al. (2013): if they voted guilty they stated how many times more likely the defendant was to be guilty than not guilt; if they voted not guilty, they stated how many times more likely the defendant was to be not guilty than guilt. The other half used a log scale (Thompson et al., 2013; see Figure 1). When making final judgment of the chances of guilt (based on the incriminating forensic evidence), subjects first used the same measure they had used initially, and then they were asked to restate their judgment using the other measure.

Materials

Case description. Our hypothetical case was similar to the case used by Thompson, Kaasa, and Peterson (2013). (A complete copy of this case may be found in the online supplementary materials for this article.) The case concerned a woman who was sexually assaulted in her home. She did not get a good look at the attacker's face. The defendant, Brian Kelly, became a suspect because he was spotted near the crime scene, but the victim could not identify him. Moreover, Kelly did not match her initial description of the perpetrator and he presented a seemingly credible alibi. Nevertheless, the police had the crime laboratory compare evidence left by the rapist in the victim's bathroom (either DNA on a faucet or a shoeprint on the freshly waxed floor) with reference samples (DNA or a shoe) taken from the defendant.

Inconclusive forensic evidence. In the DNA condition, the laboratory detected a small quantity of human DNA on a swab from the bathroom faucet but the amount was too small to determine whether it could have come from the defendant. In the Shoeprint condition, the laboratory detected a faint shoeprint on the freshly waxed bathroom floor, but it was too faint to determine whether it could have been made by the defendant's shoe.

Initial verdicts and probability estimates. Before making their initial judgments on the case, participants read a list of "Points to Consider" which summarized the arguments in favor and against guilt that we would expect the prosecutor and defense lawyer to make in a case of this type. Participants were then asked to state a verdict. Then those in the log-scale condition (about half of the sample) were asked to give an estimate based on the available evidence of the "chances that Mr. Kelly is guilty" using the log scale (see Figure 1). Those in the odds condition who voted guilty were asked to fill in a textbox to complete the sentence: "Based on the available evidence I believe it is ____ times more likely that Mr. Kelly is guilty than not guilty." Those in the odds condition who voted not guilty were asked to complete the sen-

tapraid5/lhb-lhb/lhb-lhb/lhb00215/lhb0211d15z | xppws | S=1 | 4/16/15 | 16:32 | Art: 2014-2030

tence: "Based on the available evidence I believe it is ____ times more likely that Mr. Kelly is not guilty than guilty."

Presentation of forensic science evidence. At this point, participants were asked to suppose that the forensic testing had come out differently than previously described. They were then given a statement in which the forensic expert described his conclusions. The expert began by saying that Brian Kelly's DNA profile is consistent with a partial DNA profile found on the faucet handle (DNA condition) or that Brian Kelly's shoe is consistent with the print impression found on the bathroom floor (shoeprint condition). The expert went on to characterize the strength of this evidence in a manner that varied across conditions.

For instance, in the moderate strength RMP condition, those who received the DNA evidence read:

> Based on scientific data on the genetic characteristics of the human population, I estimate that approximately one in one hundred people has a DNA profile that is consistent with the partial DNA profile on the faucet. That means that there is one chance in one hundred of finding a consistent profile in a randomly chosen person.

Those who received the shoeprint evidence read:

> Based on scientific data on the characteristics of the tread patterns of shoes, I estimate that a shoe that would make a consistent print would be found in only one in one hundred pairs of shoes. That means that there is one chance in one hundred of finding a consistent tread size and pattern in a randomly chosen pair of shoes.

In the comparable moderate strength, LR condition, those who received the DNA evidence read:

> Based on scientific data on the genetic characteristics of the human population, I estimate that the results I obtained are one hundred times more likely if the partial profile came from Brian Kelly than if it came from a randomly chosen person.

Those who received the shoeprint evidence read:

> Based on scientific data on the characteristics of the tread patterns of shoes, I estimate that the results I obtained are one hundred times more likely if the print impression came from Brian Kelly's shoe than if it came from a shoe from a randomly chosen pair.

In the comparable moderate strength VE condition subjects who received the DNA evidence read:

> Based on scientific data on the genetic characteristics of the human population, I estimate that the evidence offers moderate support for the hypothesis that the DNA on the faucet came from Brian Kelly.

Those who received the shoeprint evidence read:

> Based on scientific data on the characteristics of the tread patterns of shoes, I estimate that the evidence offers moderate support for the hypothesis that the shoeprint on the bathroom floor came from Brian Kelly's shoe.

In the very strong evidence conditions, the wording was the same as above except that the word "million" was substituted for "hundred" in the RMP and LR conditions and the words "very strong support" were substituted for the words "moderate support" in the VE condition.

Cross-examination of the expert witness. Participants also read a description of information elicited during cross-examination of the forensic expert. During cross-examination the expert admitted that it is possible that an innocent person could falsely be incriminated as a result of a coincidental DNA/Shoeprint match, a laboratory error, or planting of the incriminating evidence. The expert conceded that his statements about the strength of the forensic evidence (the RMP, LR, or VE) were based solely on his estimate of the chances of a coincidental match and did not take into account the chances of a laboratory error or frame-up. He insisted however, that he had taken great care to avoid making an error in this case and that his work had been thoroughly checked by another analyst.

Final verdicts and probability estimates. Before making their final judgments on the case, participants again read a list of "Points to Consider" which summarized the arguments in favor and against guilt that we would expect the prosecutor and defense lawyer to make in a case of this type. Participants then rendered final verdicts and gave final estimates of the chances of the defendant's guilt by responding to the same questions they had been asked for their initial judgments. Those who gave probability estimates using the log scale were then asked to also give odds estimates, and vice versa.

Recognizing correct and fallacious statements. To assess people's comprehension of the expert testimony, and their susceptibility to fallacious interpretations, we presented six statements about the evidence and asked them to indicate whether each was a "correct" or "incorrect" interpretation of what the expert had said. They could also respond "I don't know." There was some variation in these statements across conditions, as indicated (in parentheses) in the following examples, to make the statements consistent with the forensic evidence that participant in each condition had received. Some of the statements were correct and some were incorrect. For example, one of the correct statements was "[Brian Kelly/Brian Kelly's Shoe] is not the only [person/shoe] that could have [left the DNA found on the bathroom faucet/made the shoeprint found on the bathroom floor]."

We were primarily interested in how participants would respond to three incorrect statements that were consistent with fallacious interpretations. Two statements were consistent with the source probability error:

- (E1): "It is [one hundred/one million] times more likely that the [DNA on the bathroom faucet/shoeprint on the bathroom floor] came from [Brian Kelly/Brian Kelly's shoe] than from a random [person/shoe]."
- (E2): "There is only one chance in [one hundred/one million] that the [DNA on the faucet/shoeprint on the floor] came from any [person/shoe] other than [Brian Kelly/Brian Kelly's shoe]."

And one statement was consistent with the defense attorney's fallacy:

- (D1): "The [DNA/shoeprint] evidence has little value for proving Brian Kelly is guilty because a lot of other [people besides Kelly/shoes besides Kelly's] could have left the [DNA/shoeprint]."

To assess whether people have insight into their own level of understanding of the scientific evidence, we also asked them to

tapraid5/lhb-lhb/lhb-lhb/lhb00215/lhb0211d15z | xppws | S=1 | 4/16/15 | 16:32 | Art: 2014-2030

estimate how well they understood the forensic evidence on a 7-point scale, 1 = *I did not understand at all*, 7 = *I completely understood*.

Perceived chances of a coincidental match, frame-up, and laboratory error. To assess the logical coherence of participants' probability judgments, we asked them to give estimates of the chances that "an innocent man in a case like this one" would be incriminated falsely by a coincidental match, a frame-up, or a laboratory error. They gave estimates of the chances of each of these errors on 9-point scales in which the options were *1 chance in 10*, *1 chance in 100*, *1 chance in 1,000*, *1 chance in 10,000*, *1 chance in 100,000*, *1 chance in 1 million*, *1 chance in 1 billion*, *1 chance in 1 trillion*, *zero-chances-impossible*.

Results

Verdicts

When asked to render an initial verdict under the assumption that the forensic evidence was inconclusive only 3.6% of our subjects voted guilty. At this point the evidence presented to subjects was the same in every condition except for the nature of the forensic examination that was attempted (examination of DNA or a shoeprint), and subjects were told that the results of the forensic examination were inconclusive. So no statistically significant differences across conditions were expected and none were found.

When subjects rendered final verdicts after receiving the incriminating forensic evidence the conviction rate was significantly higher overall and there were differences among the conditions as shown in Figure 2. We used logistic regression to predict verdicts based on the experimental variables and found a significant effect of type of forensic evidence (higher conviction rate with DNA evidence, 24.1%, than with shoeprint evidence, 9.9%), and a significant two-way interaction between presentation format and strength of evidence. No other main effects or interactions were significant. (For the overall model, pseudo $R^2 = .10$, $\chi^2(6) = 32.62$, $p < .01$; see Table S2 in the supplementary materials for additional details). Separate logistic regression analyses within each of the presentation format conditions showed that the interaction arose because the variation in strength of evidence had a significant effect on conviction rates in the RMP condition (b = −1.443, *SE* = .403; Wald(1) = 12.839, $p < .01$; Exp(B) = .236, 95% CI [.107, .520]), but did not affect conviction rates in the LR condition (b = .230; *SE* = .42; Wald(1) = .31, $p = .58$; Exp(B) = 1.26, 95% CI [.558, 2.84]) or VE condition (b = 0.00; *SE* = .44; Wald(1) = 0.00, $p = 1.00$; Exp(B) = 1.00, 95% CI [.42, 2.38]). Taken together, these findings support both hypotheses 1 and 3—that people are more sensitive to the strength of the forensic evidence when it is presented in a RMP format and that people give more weight to DNA evidence than shoeprint evidence.

Likelihood of Guilt

Next, we examined people's estimates of the likelihood of guilt and how they updated those estimates when forensic evidence was

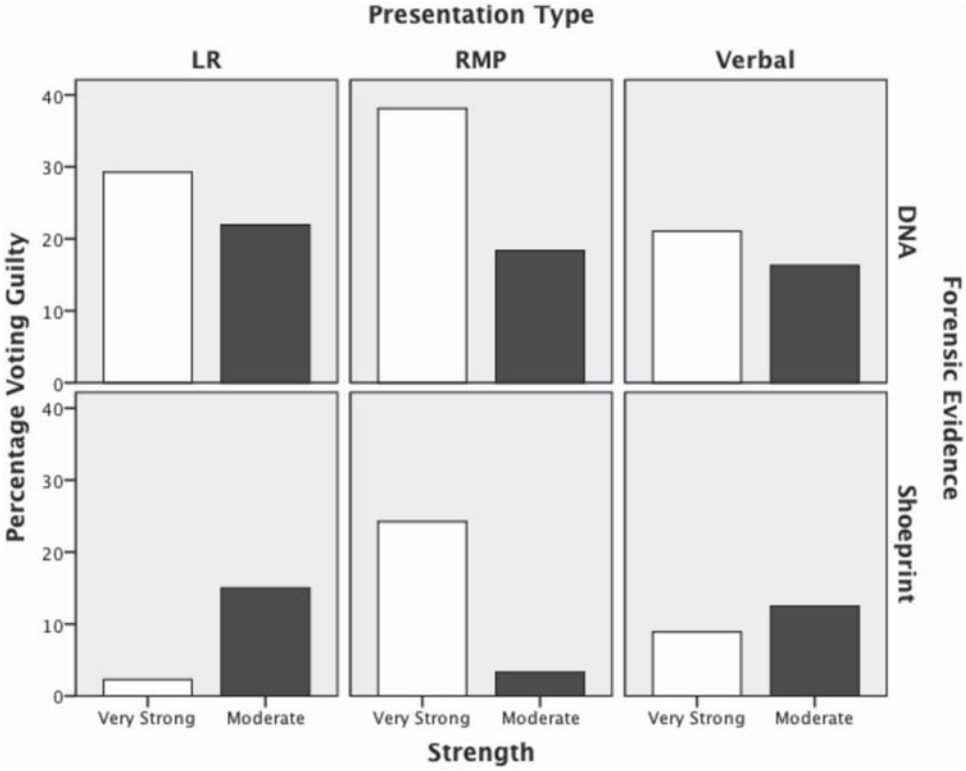


Figure 2. Conviction rates by type of forensic evidence (DNA or shoeprint), presentation format, and strength of evidence.

tapraid5/lhb-lhb/lhb-lhb/lhb00215/lhb0211d15z | xppws | S=1 | 4/16/15 | 16:32 | Art: 2014-2030

introduced. Approximately half of participants used the log scale measure (see Figure 1) to estimate the likelihood of guilt before and after receiving the forensic evidence; the other half used the odds measure. We describe the results for each measure.

Log scale. The log scale is a 17-point scale. With the exception of the highest point (*certain to be guilty*) and the lowest point (*impossible that he is guilty*), the intervals between the points are approximately equal on a scale of log odds. Each step up or down the scale represents a change by approximately a factor of 10 in the odds ratio (which corresponds to a change of one unit in log odds).

When making their initial judgments, based on the inconclusive forensic evidence, most subjects used the lower half of the scale; very few gave judgments higher than 1 chance in 2; the median response was 1 chance in 1,000 and the modal response was 1 chance in 10. When the incriminating forensic evidence was presented their responses were generally higher: 15% said the chances of guilt were 9 in 10 or higher; the median response was 1 chance in 100 and the modal response was 1 chance in 2.

We computed a log-scale change score for each subject indicating how many steps up (positive values) or down (negative values) they moved on this scale after receiving the incriminating forensic evidence. Means of these change scores in each experimental condition are displayed in Figure 3 (light bars). The change scores were larger for DNA evidence ($M = 2.14$, $SD = 3.07$) than for shoeprint evidence ($M = 0.93$, $SD = 2.12$), $F(1, 238) = 13.84$, $p < .01$, $\eta_p^2 = .06$; the mean difference was -1.21, 95% CI [$-.54$, -1.87]. Change scores were also larger for the very strong evidence ($M = 2.11$, $SD = 2.88$) than for the moderate evidence ($M = 1.00$, $SD = 2.44$), $F(1, 238) = 11.79$, $p < .01$, $\eta_p^2 = .05$; the mean difference was -1.11, 95% CI [$-.45$, -1.78]. Change scores were not significantly affected by presentation format and no interactions among the independent variables were detected, all Fs < 1.

One way to understand the change scores is to consider the likelihood ratio that would be needed to persuade a rational Bayesian to change in the same manner. On average, our participants jumped two steps on the log scale after receiving DNA evidence. This is the way a rational Bayesian would respond if she thought the value of the DNA evidence, as measured by likelihood ratio, was about 100. By comparison, participants jumped an average of only one step on the log scale after receiving shoeprint evidence. This is the way a rational Bayesian would respond if he thought the value of the shoeprint evidence, as measured by likelihood ratio, was only 10. This difference provides additional support for Hypothesis 3—that people give more weight to DNA than shoeprint evidence.

Shifts on the log scale were also sensitive to the expert's characterization of the strength of the forensic evidence. In the

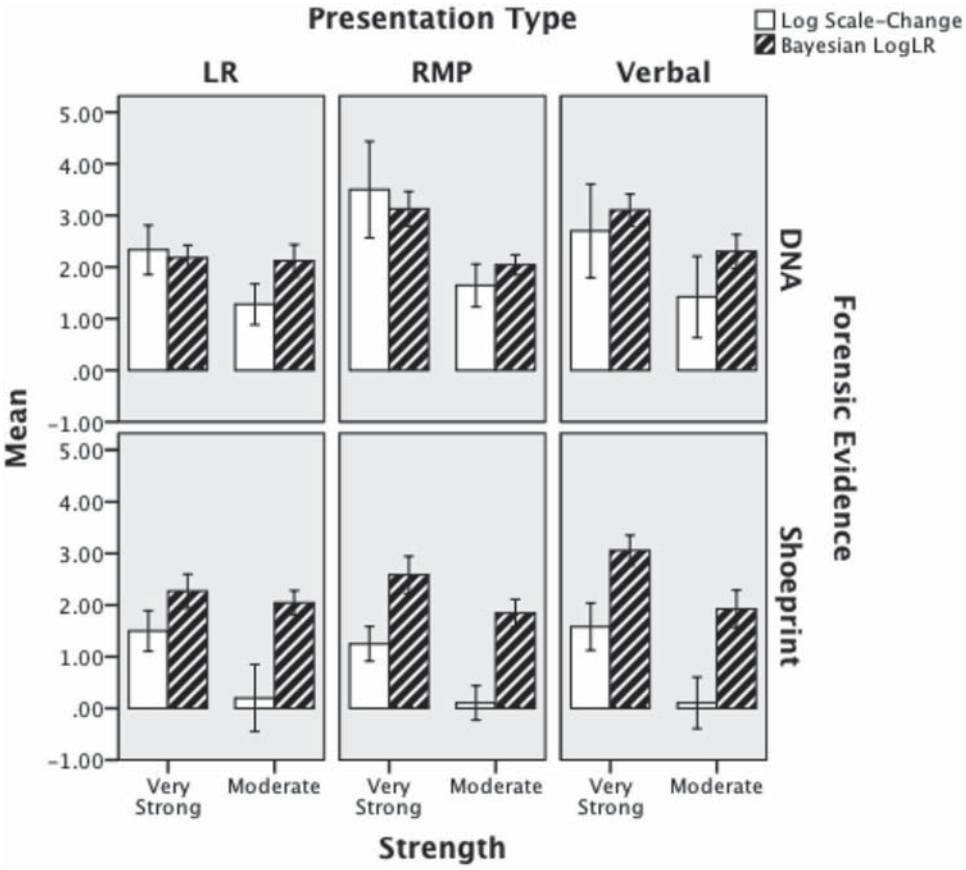


Figure 3. Comparison of log-scale change scores with log likelihood ratios derived (in accordance with a Bayesian model) from estimates of the probability of error. Error bars represent 1 standard error of the mean.

tapraid5/lhb-lhb/lhb-lhb/lhb00215/lhb0211d15z | xppws | S=1 | 4/16/15 | 16:32 | Art: 2014-2030

"very strong" conditions (RMP of 1 in 1 million; LR of 1 million; VE of "very strong") people treated the evidence the way a rational Bayesian would treat evidence with a likelihood ratio of approximately 100; in the "moderate" conditions (RMP of 1 in 100; LR of 100; VE of "moderate"), people treated the evidence the way a rational Bayesian would treat evidence with a likelihood ratio of about 10. Surprisingly, these change scores on the log scale, unlike the verdicts, were not affected by presentation format. Hence, unlike the verdicts, these findings do not support Hypothesis 1 (greater sensitivity to strength of evidence in the RMP condition).

Odds measure. About half of our subjects estimated the odds of guilt before and after receiving the forensic evidence. To measure how much their judgments changed, we computed an implicit likelihood ratio for each subject by dividing their second judgment by their first. To make judgments of those voting guilty and not guilty comparable when doing this calculation, we used the actual estimates of those voting guilty and the reciprocal of the actual estimates for subjects voting not guilty. For example, a subject who initially said the defendant was twice as likely to be not guilty as guilty was coded as expressing odds of 1:2 or 0.5 in favor of guilt, whereas a subject who said he was three times as likely to be guilty as not guilty was coded as expressing odds of 3:1 in favor of guilt. If the subject changed from the former odds to the latter odds as a result of hearing the forensic evidence, the subject had an implicit likelihood ratio of 3/0.5 = 6. Like the log-scale change scores, the implicit likelihood ratios can be understood as indicating the likelihood ratio that would be needed to persuade a rational Bayesian to update an odds judgment in the same way the participant did. A participant with an implicit likelihood ratio of 6, for example, responded to the forensic evidence the way a rational Bayesian would have responded to evidence with a likelihood ratio of 6.

Most subjects had implicit likelihood ratios between 1 and 10, but there were some extreme scores as high as 10,000 and as low as 0.01. Figure 4 displays the median implicit likelihood ratios in each experimental condition (we display medians rather than means because they are less influenced by extreme scores and thereby provide a better picture of the central tendency of the data). To facilitate statistical analysis, we performed a 90% Winsorization of the distribution—restricting the range to 0.67–68.85 before performing statistical tests. F4

An ANOVA on the Winsorized implicit likelihood ratios showed that there was a main effect for type of forensic evidence $F(1, 252) = 10.03$, $p < .01$, $\eta_p^2 = .038$, a main effect for evidence strength, $F(1, 252) = 7.05$, $p = .01$, $\eta_p^2 = .027$, and a main effect for presentation type, $F(2, 252) = 3.14$, $p = .045$, $\eta_p^2 = .024$. None of the two-way interactions was significant, although a presentation type by forensic evidence interaction was marginal, $F(2, 252) = 2.39$, $p = .09$. There was, however, a significant three-way interaction among these variables, $F(2, 252) = 3.85$, $p = .023$, $\eta_p^2 = .030$. The three-way interaction arose because the implicit likelihood ratios of subjects who received DNA evidence were sensitive to the strength variable regardless of presentation format, whereas those who received the shoeprint evidence were sensitive to the strength variable only in the RMP condition.

To support this interpretation, we separated subjects in the DNA condition from those in the shoeprint condition and conducted follow-up ANOVAs on each group. The ANOVA for the DNA

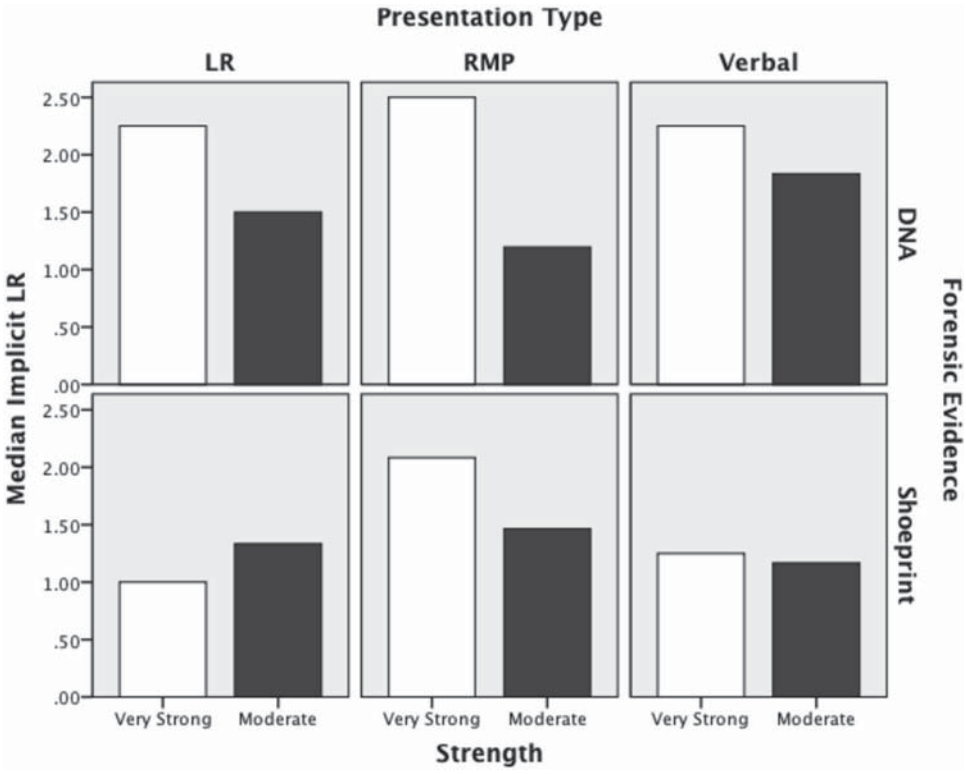


Figure 4. Median implicit likelihood ratios (odds scale).

tapraid5/lhb-lhb/lhb-lhb/lhb00215/lhb0211d15z | xppws | S=1 | 4/16/15 | 16:32 | Art: 2014-2030

condition showed a main effect of evidence strength, $F(1, 116) = 5.25$, $p = .02$, $\eta_p^2 = .04$, (DNA: $M_{moderate} = 5.49$, $SD = 13.96$; $M_{very\ strong} = 13.18$, $SD = .23.18$), mean difference −7.69, 95% CI [−14.46, −.92] and no other significant main effects or interactions (Presentation format $F(2, 116) = 2.50$, $p = .09$; Presentation × Strength, $F(2, 116) = 2.03$, $p = .14$). The ANOVA for the shoeprint condition showed no main effect for evidence strength, $F(1, 136) = 1.34$, $p = .25$, and only a marginal effect for presentation format, $F(1, 136) = 3.00$, $p = .05$, $\eta_p^2 = .04$, but found a significant interaction between presentation type and strength of evidence, $F(2, 136) = 5.61$, $p = <.01$, $\eta_p^2 = .08$. Tukey's HSD post hoc analyses revealed that scores in the RMP condition were sensitive to the strength of the evidence, but scores in the LR and VE conditions were not (RMP: $M_{moderate} = 1.70$, $SD = 1.11$; $M_{very\ strong} = 11.08$, $SD = 22.41$), mean difference −9.38, 95% CI [−16.23, −2.53]; LR: $M_{moderate} = 4.44$, $SD = 14.20$; $M_{very\ strong} = 1.34$, $SD = .60$) mean difference, 3.10, 95% CI [−3.02, 9.22]; VE: $M_{moderate} = 1.97$, $SD = 2.15$; $M_{very\ strong} = 1.39$, $SD = .55$), mean difference .58, 95% CI [−.40, 1.57]). These findings thus support Hypothesis 1 (greater sensitivity to the strength of forensic evidence when it is presented in the RMP format) but only for shoeprint evidence. For DNA evidence, people were sensitive to evidence strength regardless of presentation format.

Log scale versus odds. Subjects who gave their responses as odds appeared to give the forensic evidence less weight than those who gave their responses on the Log scale. As noted earlier, those using the log scale treated the forensic evidence the way a rational Bayesian would treat evidence with a likelihood ratio between 10 and 100. By contrast, those who gave their responses as odds treated the evidence the way a rational Bayesian would treat evidence with a much lower likelihood ratio ($M = 5.87$; $SD = 15.07$; Median = 1.5). This finding supports hypothesis 4—that judgments of the chances of guilt made using the odds measure would be less responsive to the forensic evidence (lesser shifts in judgments) than judgments made using the log scale.

Treating Forensic Evidence as Exculpatory

Martire et al. (2013, 2014) reported that people sometimes give lower estimates of the odds of guilt after receiving forensic evidence that is characterized as weakly supportive of the prosecution (a finding they call the "weak evidence effect"). Our experiment had no conditions in which an expert characterized the forensic evidence as "weak," but we nevertheless found that a small percentage of our participants (7.6% for the Log scale; 6.4% for the Odds measure) judged the defendant less likely to be guilty after receiving the forensic evidence than before—in other words, they treated the forensic evidence as exculpatory. A logistic regression found that exculpatory shifts were more common in the moderate evidence condition (9.9%) than the very strong evidence condition (3.7%), but found no effect of presentation format or type of forensic evidence, pseudo $R^2 = .05$, $\chi^2(4) = 9.19$, $p = .06$ (see Table S3 in the supplementary materials for details).

Estimates of the Probability of Error

Participants estimated the chances of three different events that could falsely incriminate an innocent person: a random (coincidental) match; a false report of a match due to lab error; and match due to a frame-up. They made each estimate on a scale on which the options were labeled as follows: 0—*Impossible*; 1—*1 chance in 1 trillion*; 2— *1 chance in 1 billion*; 3—*1 chance in 1 million*; 4—*1 chance in 100,000*; 5—*1 chance in 10,000*; 6—*1 chance in 1,000*; 7—*1 chance in 100*; 8—*1 chance in 10*. Because this was not an equal interval scale, we treated the data as ordinal and used nonparametric tests to analyze the results. Figure 5 shows how median estimates of the chances of a random match varied across conditions. Overall, estimates were lower for DNA than for shoeprint evidence, $U = 43{,}737.5$, $z = 7.60$, $p < .001$, $r = .34$, and were lower for the "very strong" evidence than the "moderate" evidence, $U = 45{,}484$, $z = 8.75$, $p < .001$, $r = .39$. Presentation format did not have a significant effect overall, $H(2) = 2.65$, $p = .26$, but the variation in strength of evidence produced larger effects in the RMP condition, $U = 6423$, $z = 7.64$, $p < .001$, $r = .57$, than in the LR condition, $U = 4458.5$, $z = 4.61$, $p < .001$, $r = .36$, or the VE condition, $U = 4181$, $z = 2.72$, $p = .006$, $r = .21$. F5

Figure 6 shows the variation across conditions in median estimated chances of a false report of a match attributable to laboratory error. Overall, estimates were lower for the "very strong" evidence than the "moderate" evidence, $U = 38381.5$, $z = 4.28$, $p < .001$, $r = .19$ and for DNA than for shoeprint evidence, $U = 34.821.5$, $z = 2.00$, $p = .045$, $r = .09$. Presentation format did not have a significant effect, $H(2) = 2.59$, $p = .27$. Figure 7 shows the variation across conditions in median estimates of the chances of a frame-up. The nature of the forensic evidence (DNA or shoeprint) did not significantly influence these estimates, $U = 30123$, $z = -0.93$, $p = .36$, but estimates were lower for the "very strong" evidence than the "moderate" evidence, $U = 34809$, $z = 2.03$, $p < .042$, $r = .09$. Presentation format did not have a significant effect overall, $H(2) = 2.81$, $p = .25$. F6 F7

We were surprised that judgments about the probability of a lab error and a frame-up were influenced by the strength of the

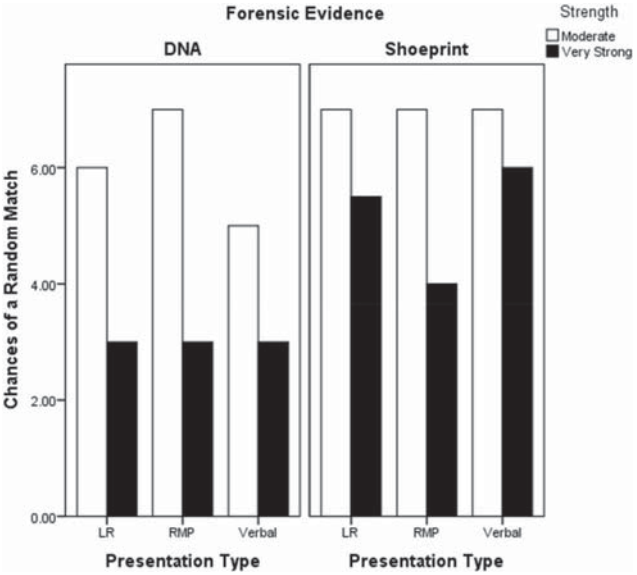


Figure 5. Median estimates of the chances of an innocent defendant being incriminated by a random (coincidental) match. Scale: 0 = *Impossible*; 1 = *1 chance in 1 trillion*; 2 = *1 chance in 1 billion*; 3 = *1 chance in 1 million*; 4 = *1 chance in 100,000*; 5 = *1 chance in 10,000*; 6 = *1 chance in 1,000*; 7 = *1 chance in 100*; 8 = *1 chance in 10.*

tapraid5/lhb-lhb/lhb-lhb/lhb00215/lhb0211d15z | xppws | S=1 | 4/16/15 | 16:32 | Art: 2014-2030

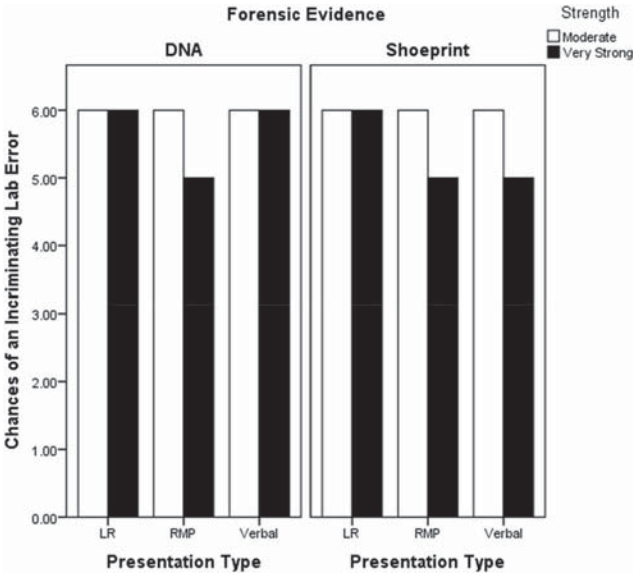


Figure 6. Median estimates of the chances of an innocent defendant being incriminated by a falsely reported match arising from laboratory error. Scale: 0 = *Impossible*; 1 = *1 chance in 1 trillion*; 2 = *1 chance in 1 billion*; 3 = *1 chance in 1 million*; 4 = *1 chance in 100,000*; 5 = *1 chance in 10,000*; 6 = *1 chance in 1,000*; 7 = *1 chance in 100*; 8 = *1 chance in 10.*

forensic evidence when the expert stated clearly that his testimony about its strength was based solely on the probability of a coincidental match. There seems no logical reason that a participant's estimate of the chances of a frame-up or a lab error should be lower when the expert reports a match on a rare DNA profile or shoeprint than when the expert reports a match on a more common profile or print. One possibility is that the effect arose from a form of bidirectional or coherence-based reasoning (Holyoak & Simon, 1999; Read & Simon, 2012) in which people shifted their interpretation of the strength of the evidence in a manner that supports their overall conclusion about the case. But follow-up analyses found that strength of evidence affected the chances of a lab error both for participants who voted guilty, $U = 1164$, $z = 3.37$, $p = .001$, $r = .37$, and those who voted not guilty, $U = 25722.5$, $z = 3.10$, $p = .002$, $r = .15$. A similar analysis on the chances of a frame up found no significant difference among those voting guilty, $U = 918.5$, $z = 0.98$, $p = .329$, but a marginal difference among those who voted not guilty, $U = 24256$, $z = 1.88$, $p = .06$, $r = .09$. So it does not appear that people were simply responding in a way that was consistent with their verdict. Perhaps people assumed that the expert's testimony somehow captured information about the chances of a lab error and a frame up when it did not.

Comparison With Bayesian Norms

To assess the logical coherence of people's judgments about the evidence it was necessary to develop a normative Bayesian model that specified how much each participant should have updated their initial judgment of guilt in light of the forensic evidence. Our model, illustrated in Figure 8, is a Bayesian Network (see Taroni, Aitken, Garbolino, & Biedermann, 2006) with four nodes, each of which has two possible states. It assumes the defendant is either guilty or not guilty, that the forensic evidence either matches or does not match the defendant, and that the forensic expert either reports a match or does not. It also assumes that someone either did or did not attempt to frame the defendant by planting the matching item at the crime scene. The model makes several simplifying assumptions about the conditional probability of various events. Specifically, as shown in Figure 8, it assumes that the defendant is certain to match the forensic evidence if he is either guilty or was framed, and it assumes the expert is certain to report a match if a match occurs. The other conditional probabilities required by the model are those that each participant provided when estimating the random match probability (RMP), false report probability (FRP), and frame-up probability (FUP). By inserting each participant's own estimates of these variables into the model, we computed a normative likelihood ratio that indicates how much weight the participant should give to the forensic evidence (given the participant's perceptions of the probability of the three possible sources of error). The model can also be described with an equation. It assumed that the likelihood ratio describing the strength of the reported forensic science evidence, R, was as follows: $p(R \mid G)/p(R \mid NG) = 1/[FUP + RMP\ (1 - FUP) + FRP(1 - RMP)(1 - FUP)]$. This decomposition of the likelihood ratio is consistent with models of cascaded inference originally described by David Schum and his colleagues (Schum, 1994; Schum & Martin, 1982).

Coherence of log-scale judgments. The base-ten logs of the likelihood ratios derived from this model (which we call the Bayesian LogLRs) indicate how much a rational Bayesian should change his or her estimate of the chances of guilt when using the log scale. Suppose, for example, that we input into the Bayesian network model a participant's estimates of the probability of a coincidental match, lab error, and frame-up, and learn that the

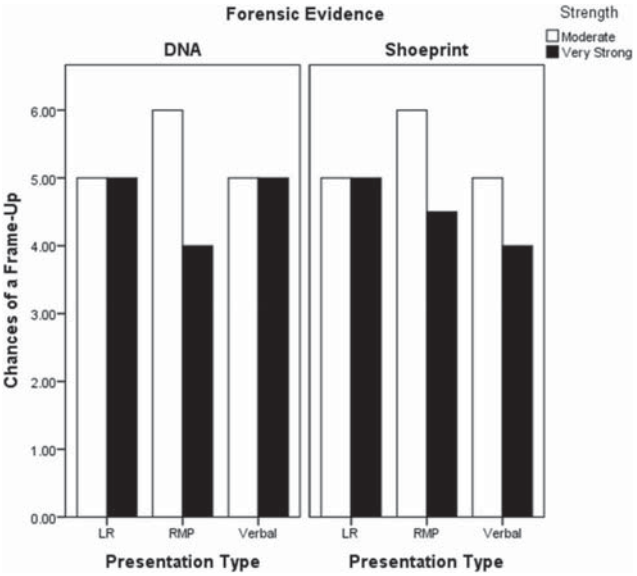
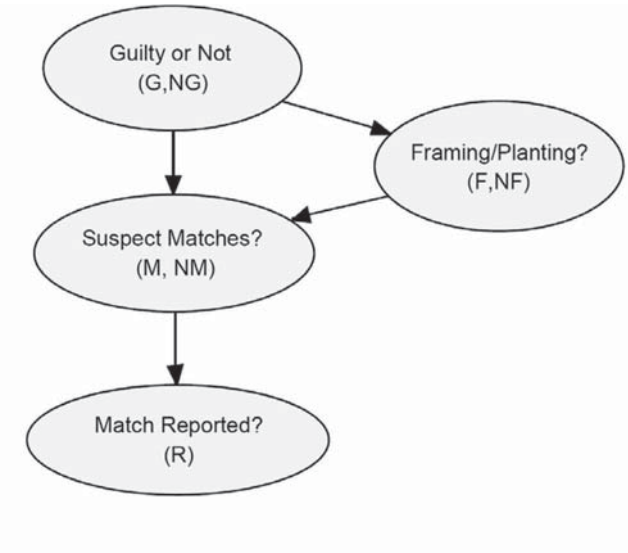


Figure 7. Median estimates of the probability of an innocent defendant being falsely incriminated by evidence planting (a frame-up). Scale: 0 = *Impossible*; 1 = *1 chance in 1 trillion*; 2 = *1 chance in 1 billion*; 3 = *1 chance in 1 million*; 4 = *1 chance in 100,000*; 5 = *1 chance in 10,000*; 6 = *1 chance in 1,000*; 7 = *1 chance in 100*; 8 = *1 chance in 10.*

F8

tapraid5/lhb-lhb/lhb-lhb/lhb00215/lhb0211d15z | xppws | S=1 | 4/16/15 | 16:32 | Art: 2014-2030



	Guilty (G)	Not Guilty (NG)
Framed (F)	FUP	FUP
Not Framed (NF)	1-FUP	1-FUP

	Framed (F)		Not Framed (NF)	
	G	NG	G	NG
Match (M)	1	1	1	RMP
No Match (NM)	0	0	0	1-RMP

	Match (M)	No Match (NM)
Match Reported (R)	1	FRP
Not Reported	0	1-FRP

Figure 8. Bayesian network model for evaluating the probative value of the forensic evidence based on individual perceptions of the RMP, FRP, and FUP.

resulting likelihood ratio is 100. The corresponding Bayesian LogLR would be 2, which indicates that if the participant responds to the evidence in a manner consistent with Bayes' rule, then the participant's log-scale estimate should go up about two steps (a log-scale change score of 2). By comparing the Bayesian LogLRs to log-scale change scores, we can assess the logical coherence of participants' judgments—and specifically, whether they each gave as much weight to the forensic evidence as they should have in light of their estimates of the chances of a false match attributable to coincidence, lab error, or framing.

Figure 3 shows how the mean Bayesian LogLRs and mean log-scale change scores varied across our experimental conditions. To compare the change scores with the Bayesian LogLRs we performed a repeated measures ANOVA in which the within factor was the measure (Change score vs. Bayesian LogLR) and the between factors were evidence strength, presentation format and type of forensic evidence. We found significant main effects for measure, $F(1, 229) = 27.196$, $p < .001$, $\eta_p^2 = .11$, strength of evidence, $F(1, 229) = 24.24$, $p = .001$, $\eta_p^2 = .10$, and forensic evidence, $F(1, 229) = 14.109$, $p = .001$, $\eta_p^2 = .06$, but also a significant two-way interaction between measure and type of forensic evidence, $F(1, 229) = 11.002$, $p = .001$, $\eta_p^2 = .05$. The two-way interaction arose because the Bayesian LogLRs were significantly higher than the change scores for the shoeprint evidence, $t(114) = 6.77$, $p < .01$, mean difference = 1.47, 95% CI [1.04, 1.90] but not for the DNA evidence, $t(125) = 1.22$, $p = .22$, mean difference = .32, 95% CI [−.20, .84]. Compared with Bayesian norms, subjects appear to have underutilized the shoeprint evidence but not the DNA evidence when adjusting their judgments of the likelihood of guilt on the log scale.

Interestingly, presentation format had no significant effects in this analysis. Hence these findings do not support hypothesis 2 (that people respond to forensic evidence in a manner more consistent with Bayesian norms when it is presented in the RMP format, than in the LR or VE format). Surprisingly, the key variable affecting the logical coherence of participants' judgments was the type of forensic evidence (DNA vs. shoeprint) rather than the way it was presented (RMP, LR, or VE).

Coherence of odds estimates. We also wanted to know whether participants who gave initial and final judgments of the chances of guilt using the odds measure updated their estimates in a manner consistent with our Bayesian model. So we compared implicit likelihood ratios (computed by dividing the second odds judgment by the first) to the LRs derived from our Bayesian model. (To make the scales comparable, we examined LRs derived from the Bayesian model, rather than the LogLRs discussed in the previous section). This comparison tells us whether, after receiving the forensic evidence, participants updated their odds estimates as much as our Bayesian model indicates that they should have, given their stated beliefs about the probability that an innocent person could be falsely incriminated. They clearly did not.

A comparison of the overall means is somewhat misleading because both distributions are highly skewed, but the implicit LRs are far smaller than the Bayesian LRs ($M_{\text{ImplicitLR}} = 60.03$; Me-

tapraid5/lhb-lhb/lhb-lhb/lhb00215/lhb0211d15z | xppws | S=1 | 4/16/15 | 16:32 | Art: 2014-2030

dian$_{\text{ImplicitLR}}$ = 1.5; $M_{\text{BayesianLR}}$ = 18,997; Median$_{\text{BayesianLR}}$ = 98.06), paired $t(258) = 4.39$, $p < .01$. This finding supports the notion that the conservatism (relative to Bayesian norms) reported in some earlier studies (e.g., Martire et al., 2013, 2014) may arise partly from scaling effects. People may simply find it easier to give high estimates on the log scale, where they must check a box to indicate their answer, than on the odds scale, where they must generate a number on their own. The great majority of participants gave one or two digit answers when asked to provide odds estimates.

Susceptibility to Fallacious Interpretations

When asked how well they understood the forensic science evidence, subjects gave themselves high ratings—averaging 6.03 (SD = 1.08) on a 7-point scale that ranged from 1 (*I did not understand at all*) to 7 (*I understood completely*). An ANOVA found no significant effects of the experimental variables on these ratings (all Fs <1.94). But nearly two thirds of subjects (63.6%) indicated that at least one of the two statements consistent with the source probability error represented a "correct interpretation" of the expert's testimony (55.2% thought statement E1 was correct; 51.7% thought E2 was correct). Nearly half of subjects (49.1%) indicated that the statement consistent with the defense attorney's fallacy (D1) was "correct."

We used logistic regression to predict whether participants would agree with at least one of the statements representing the source probability error on the basis of experimental condition. There was no significant effect for strength or type of forensic evidence, but as expected (Hypothesis 5) agreement with at least one fallacious statement was higher in the LR condition (85.6%) and in the RMP condition (78.4%) than in the VE condition (26.06%). (For the overall model, pseudo R^2 = .46, $\chi^2(4) = 204.55$, $p < .01$; details may be found in Table S4 in the supplementary materials). When we analyzed the two fallacious statements separately, we found that statement E1 was more likely to be viewed as correct in the LR condition (90.68%) than in the RMP condition (57.46%) or VE condition (18.18%), pseudo R^2 = .43, $\chi^2(4) = 198.4$, $p < .01$ (see supplementary Table S5 for details). By contrast, statement E2 was more likely to be viewed as correct in the RMP condition, (74.57%) than the LR condition (62.11%) or Verbal condition (16.36%), pseudo R^2 = .33, $\chi^2(4) = 145.29$, $p <$ .01 (see supplementary Table S6). This pattern makes sense given that the wording of E1 is similar to the wording of the expert's conclusions in the LR condition, whereas the wording of E2 is similar to the wording of the expert's conclusions in the RMP condition. Participants thought the fallacious statement was a "correct interpretation" when it sounded similar to what the expert had said, even though the statement transposed conditional probabilities in a manner that made it erroneous and potentially misleading.

We also used logistic regression to predict rate of agreement with the defense attorney's fallacy and found significant effects of type of evidence (55.8% for shoeprint; 42.2% for DNA) and strength of evidence (35.4% for very strong evidence; 61.4% for moderate), but no significant effect of presentation format. (For the overall model, pseudo R^2 = .13, $\chi^2(4) = 52.18$, $p < .01$; see supplementary Table S7 for details).

The final step in our analysis of people's susceptibility to fallacious reasoning was to examine the connection between agreement with fallacious interpretations of the forensic evidence and participants' evaluations of its strength, as reflected in conviction rates and shifts in their estimates of the chances of guilt. Table 1 shows the percentage of participants who endorsed one, both, or neither of the fallacies along with the conviction rate, mean log change score, and mean implicit LR for each of those groups. T1 AQ: 3

As Table 1 shows, falling victim to fallacious reasoning had implications for verdicts. Those who endorsed the source probability error but not the defense attorney's fallacy were most likely to convict (32.29%), which supports Hypothesis 6a. By contrast, those who endorsed the defense attorney's fallacy but not the source probability error were least likely to convict (3.09%), which supports Hypothesis 7a. Interestingly, 28.1% of participants endorsed statements consistent with both fallacies. Among that group, the conviction rate was also low (5.26%), which is consistent with earlier findings suggesting that the defense attorney's fallacy is more influential than the source probability error among people who consider both (Thompson & Schumann, 1987). A logistic regression showed that compared with participants who endorsed neither fallacy, those who endorsed only the source probability error were significantly more likely to convict; while those who endorsed the only the defense attorney's fallacy or both fallacies, were less likely to convict, pseudo R^2 = .20, $\chi^2(3) = 64.87$, $p < .01$ (for details see Table S8 in the supplementary materials).

We found a similar pattern of results when we examined log-scale change scores and implicit LRs (see Table 1). As predicted (Hypothesis 6b), participants who agreed with the source probability error showed larger shifts toward guilt than people who endorsed only the defense attorney's fallacy, both fallacies, or

Table 1
Percentage of Subjects Who Endorsed the Source Probability Error, Defense Attorney's Fallacy, Both Errors, or Neither Error and Conviction Rates, Log Change Scores, and Implicit LRs Within Each Group

Error endorsed	Percentage endorsing fallacy	Conviction rate	Log scale change score	Implicit LR
Source probability error only	35.49% (192)	32.29% (62)	1.93 (3.19)	12.1 (22.18)
Defense attorney's fallacy only	17.93% (97)	3.09% (3)	1.14 (1.84)	3.09 (9.79)
Both errors	28.10% (152)	5.26% (8)	1.26 (2.21)	1.4 (.58)
Neither error	12.20% (66)	15.15% (10)	1.46 (3.25)	4.12 (10.91)

Note. For percentage endorsing fallacy and conviction rate, numbers in parentheses indicate the number of participants falling in each category (n = 507). For Log Scale Change Score and Implicit LR, numbers in parentheses indicate the standard deviation.

tapraid5/lhb-lhb/lhb-lhb/lhb00215/lhb0211d15z | xppws | S=1 | 4/16/15 | 16:32 | Art: 2014-2030

neither fallacy. ANOVAs showed that there was a significant main effect of fallacy for implicit LRs, $F(3, 256) = 9.02$, $p < .01$, $\eta_p^2 = .10$, and post hoc Tukey's HSD analyses show that people who endorsed the source probability error only had the largest shift toward guilt, there were no other significant differences (mean difference$_{\text{source prob only, both}}$, 10.69, 95% CI [6.35, 15.03]; mean difference$_{\text{source prob, only def only}}$, 9.00, 95% CI [3.88, 14.13]; mean difference$_{\text{source prob only, neither}}$, 7.97, 95% CI [2.46, 13.49]). There was a similar pattern for log change scores, $F(3, 240) = 1.27$, $p = .29$, $\eta_p^2 = .02$, that did not reach significance.

Discussion

This experiment examined lay participants' responses to two types of forensic science evidence, a DNA comparison and a shoeprint comparison, when an expert explained the strength of the evidence three different ways—using random match probabilities (RMPs), likelihood ratios (LRs), or verbal equivalents to likelihood ratios (VEs). To assess the appropriateness of people's responses to the evidence we considered three factors: the sensitivity of their judgments to the strength of the evidence, the logical coherence of their judgments, and their susceptibility to fallacious interpretations of the evidence.

Sensitivity to the Strength of Evidence

If people are appropriately sensitive to the strength of the forensic science evidence then conviction rates and estimates of the chances of the defendant's guilt should have been higher among participants who received the very strong evidence (RMP of 1 in 1 million; LR of 1 million; VE of "very strong support") than among those who received the moderate evidence (RMP of 1 in 100; LR of 100; VE of "moderate support"). When we examined verdicts (see Figure 2), we found that participants in the RMP condition were sensitive to the strength of both the DNA evidence and the shoeprint evidence; those in the LR and VE conditions, by contrast, were sensitive to the strength of the DNA evidence, but not the strength of the shoeprint evidence. We found the same pattern of results for implicit likelihood ratios (see Figure 4), which reflect shifts in participants' estimates of the odds of guilt after receiving the forensic evidence. These findings support Hypothesis 1—that people will be more sensitive to the strength of forensic evidence when experts present RMPs—but that conclusion must be qualified in light of our other findings.

Intriguingly, we found a different pattern of results for two other measures of the perceived strength of the forensic evidence that are shown in Figure 3. The log-scale change scores (light bars) show how much participants increased their estimates of the chances of guilt (on the log scale shown in Figure 1) after receiving the forensic evidence. The log likelihood ratios (striped bars) reflect participants' judgments of the chances of three possible sources of a false match: a coincidence, a frame-up, and a lab error (lower estimates of the chances of a false match produce higher log LRs). On both of these measures we found that people were sensitive to the strength of the forensic evidence regardless of presentation format for both DNA evidence and shoeprint evidence, although they gave considerably less weight to shoeprint evidence overall.

The results thus pose an interesting puzzle. If our participants appreciated the difference in strength between the very strong and moderate forensic evidence, regardless of presentation format, as suggested by the results show in Figure 3, why were they sometimes insensitive to this difference when rendering verdicts (see Figure 2) and updating their judgments of the odds of guilt (see Figure 4)? Specifically, why was the difference in strength of evidence not reflected in the verdicts or odds judgments in the conditions where the expert used LRs and VEs to characterize the strength of shoeprint evidence?

Our findings in those conditions were not anomalous. Our findings are entirely consistent with the results reported by Martire et al. (2013), who also found that large variations in the strength of shoeprint evidence (as reflected in the LR or VE) made little difference to the odds judgments and verdicts. Whereas Martire et al. examined only shoeprint evidence, and only the LR and VE presentation formats, our study also examined DNA evidence and the RMP format. The broader perspective afforded by our findings indicates that people's reactions to shoeprint evidence (when characterized with LRs and VEs) may differ from their reactions to other types of forensic evidence, and even from their reactions to shoeprint evidence when characterized with RMPs. We will discuss why that might be later, but first it will be helpful to discuss the logical coherence of our participants' judgments.

Logical Coherence

We asked participants to estimate the probability that an innocent person could falsely be incriminated by the forensic evidence. They gave separate estimates of the probability an innocent person could be incriminated due to coincidence, a laboratory error, and a frame-up. We used a Bayesian network model (see Figure 8) to combine each participant's estimates of these probabilities to compute a log-likelihood ratio that showed the weight that should be assigned to the forensic evidence by a rational Bayesian who believed the participant's estimates. We then compared these Bayesian Log LRs with the weight that participants actually gave to the forensic evidence, as shown by shifts in their estimates of the chances the defendant was guilty on the log-scale.

Figure 3 shows how well the Log-Scale change scores corresponded to the Bayesian Log LRs across the experimental conditions. For the DNA evidence, participants' change scores on the log scale (top three panels of Figure 3) were quite similar to the log likelihood ratios derived from the Bayesian network model. Although the change scores were slightly lower than the Log LRs, the difference was not statistically significant, which suggests that participants responded to the DNA evidence in a manner that closely tracked Bayesian norms. This finding is consistent with the report by Thompson et al. (2013) that mock jurors respond to DNA evidence in a manner that roughly corresponds with Bayesian norms.

The most fascinating part of Figure 3 is the contrast between the DNA evidence (top panel) and the shoeprint evidence (bottom panel). For DNA evidence, the change scores were similar to the Bayesian Log LRs; for shoeprint evidence, the change scores were significantly lower than the Bayesian Log LRs. That means that judgments about DNA evidence were logically coherent, whereas judgments about shoeprint evidence were not. Our participants appear to have been good Bayesians when evaluating DNA evidence and bad Bayesians when evaluating shoeprint evidence. How can we explain that?

tapraid5/lhb-lhb/lhb-lhb/lhb00215/lhb0211d15z | xppws | S=1 | 4/16/15 | 16:32 | Art: 2014-2030

Our results clearly confirm hypothesis 3—that people will give more weight to DNA evidence than shoeprint evidence. But the comparison to Bayesian norms provides special insight into the reasons for this effect. The relative weakness of shoeprint evidence was reflected partly (but only partly) in participants' estimates of the chances of error. The Log LRs derived from our Bayesian model were significantly lower for the shoeprint evidence than the DNA evidence, which indicates that participants' estimates of the aggregate probability of the three potential sources of error were higher for shoeprint evidence than DNA evidence. The difference arose largely in estimates of the chances of a coincidental match, which were lower for DNA evidence than shoeprint evidence. Participants' general expectations about DNA and shoeprint evidence may have led them to believe that a coincidental DNA match was less likely than a coincidental shoeprint match, notwithstanding the expert's testimony. For example, in the condition where the expert presented a RMP for DNA evidence of 1 in 100, participants' average estimates of the probability of a coincidental match were just below 1 in 1000, suggesting that participants thought the RMP was actually lower than the expert had claimed. In contrast, for shoeprint evidence, participants' estimates of the probability of a coincidental match were higher than what the expert had claimed. It appears, then, that although our participants were influenced by the expert's assessment of the probability of coincidence, they did not simply accept the expert's statements as given. Instead, they made their own estimates that were colored by their general impressions of the quality of the forensic evidence (DNA or shoeprint) as well as by what the expert said.

But participants' differing estimates of the error rates for shoeprint and DNA evidence are only a partial explanation for the perceived weakness of the shoeprint evidence. These differences explain why the LogLRs generated by our Bayesian model are lower for shoeprint evidence than DNA evidence, but they do not explain why the participants deviated strikingly from the Bayesian model in their evaluation of shoeprint evidence, but not DNA evidence. Apparently, people regard shoeprint evidence as weak (relative to DNA evidence) for reasons that are not captured by our Bayesian model. And that suggests that people are considering something besides the chances of error when deciding how much weight to give to forensic evidence. There is something about DNA evidence—something beyond the numbers provided by an expert, and beyond even participants' own estimates of error rates—that makes it more powerful than shoeprint evidence.

As discussed earlier, psychologists have long rejected Bayesian models as descriptions of actual human judgment—so the failure of our model to predict reactions to shoeprint evidence is hardly a surprise. We use the model here not as a description of human judgment, but as a norm against which to compare human judgment. Specifically, the model has helped us detect the logical inconsistency between participants' error rate estimates and the weight that they gave to the shoeprint evidence. It thereby helps us see that judgments about the value of shoeprint evidence depend on something beyond logical extrapolation from error rate estimates.

But if people are not thinking like Bayesians, how are they thinking? One possibility is that they are considering some qualitative aspect of the forensic evidence in addition to their estimates of the chances of error. We propose that they also consider what we will call the credibility of the evidence. They give substantial weight to forensic evidence only if they judge that it has high credibility as well as a low risk of error. We suspect that the perceived credibility of forensic evidence is influenced largely by what epistemologist Susan Haack (2014) has called explanatory integration—that is, by how tightly the evidence fits with other evidence and with the knowledge and presumptions the trier-of-fact brings to the case. Although Haack's theories are normative rather than descriptive—that is, they concern the value the evidence warrants rather than the weight that people choose to give the evidence—we think her notion of explanatory integration captures an element of intuitive psychology that is helpful for explaining our findings. The notion of explanatory integration is consistent with the notion of expectancies (Schklar & Diamond, 1999) and with aspects of the story model of jury decision making (Pennington & Hastie, 1992).

In our view, the DNA evidence was generally given more weight than the shoeprint evidence because it was seen as having higher credibility based on everything our participants knew, or thought they knew, about forensic evidence. DNA evidence is frequently discussed in TV dramas, like the popular CSI series, where it is treated as definitive proof of identity. News reports have recounted its use both to convict the guilty and to exonerate the innocent, which suggests widespread acceptance of its value and importance in criminal justice. Indeed it has been called "the gold standard of forensic science" and a "truth machine" (Lynch, Cole, McNally, & Jordan, 2009; Thompson, 2013). In contrast, shoeprint comparison may be less well known as a forensic technique and, because shoes are mass produced products, may seem inherently less discriminating than comparison of DNA profiles.

The shoeprint evidence was undervalued relative to Bayesian norms because the Bayesian model took account only the risk of error—it did not consider the credibility of shoeprint evidence, which was low enough that participants gave it relatively little weight (particularly in the LR and VE conditions). Our manipulation of the strength of the DNA evidence was effective across all presentation formats because it affected participants' perception of the chances of error. When the expert reported a low RMP (1 in 1 million), a high LR (one million), or simply said the evidence provided "very strong" support for the prosecution's theory, participants were satisfied that the risk of error was low, and because the evidence in question was also of high credibility, they gave it substantial weight. By contrast, when the expert reported a higher RMP (1 in 100), a lower LR (100), or said the evidence provided only "moderate" support for the prosecution's theory, participants must have inferred that there was something amiss with the DNA evidence—that the chances of error for this piece of DNA evidence were higher than normal. Hence, they gave it less weight.

The finding that is most difficult to explain is the effect of presentation format on participants' sensitivity to the strength of the shoeprint evidence. Our manipulation of the strength of the shoeprint evidence had the expected effect only when the expert used the RMP format. In the LR and VE conditions, the strength manipulation did not have the intended effect. We suspect that presenting a low RMP of 1 in 1 million added something to the credibility of the shoeprint evidence. Perhaps the expert's claim that he could estimate the frequency of matching shoes in the

tapraid5/lhb-lhb/lhb-lhb/lhb00215/lhb0211d15z | xppws | S=1 | 4/16/15 | 16:32 | Art: 2014-2030

population and narrow it down to such an extent made the shoeprint evidence seem more scientific, or at least more discriminating. By contrast, the expert's statement in the LR condition that the evidence is "one million times more likely" if the shoeprint was made by the defendant's shoe may have seemed like a conclusion without evidence—hence, less well-grounded in science. The expert's statement in the VE condition that the evidence provided "very strong support" for the hypothesis that the defendant's shoe made the print might have had the same problem. These differences between the RMP format and the other formats may have made no difference for DNA evidence because DNA is already perceived as highly scientific. For shoeprint evidence, however, only the low RMP boosted credibility enough to raise conviction rates and estimates of the odds of guilt in the manner seen in our results.

To test our theory about the credibility of forensic evidence in future research, it will be necessary to find a way to measure credibility, perhaps by asking people directly how much confidence that they would generally have in evidence of a particular type. Researchers could then test whether the weight given to evidence, relative to Bayesian norms correlates with credibility. Better yet, researchers could seek to manipulate the credibility of forensic evidence experimentally, by varying the way it is described, in order to test whether that manipulation affects the weight given to the evidence in ways that go beyond Bayesian norms.

Fallacious Interpretation

A surprisingly high percentage of our participants (about two thirds) indicated that one or both statements consistent with the source probability error were a "correct interpretation" of what the expert said. These fallacious interpretations were not merely matters of semantics—we found that they were strongly associated with verdicts and estimates of the probability the defendant was guilty. As expected (Hypotheses 6a, 6b, 7a, and 7b), and as shown in Table 1, conviction rates and estimates of the chances of guilt were highest among participants who agreed with the source probability error (but not the defense attorney's fallacy), and were lowest among participant's who agreed with the defense attorney's fallacy. These findings support the conclusion that fallacious interpretation of forensic science evidence may play a significant role in the decision to convict.

Limitations and Future Directions

Before making policy recommendation on the basis of this research, we would like to see additional studies to test the generalizability of our findings. Our participants responded to a relatively brief written summary of evidence in a single hypothetical case. It remains to be seen how well our findings will replicate across a broader range of cases and types of scientific evidence. It will be particularly important for future studies to explore whether people's performance can be improved with more detailed and complete explanations of the scientific evidence, either by experts, lawyers or the judge. It seems possible, for example, that better explanations of forensic statistics, perhaps with the use of visual aids, might improve sensitivity to the strength of forensic evidence, increase the coherence of judgments, and reduce susceptibility to fallacious misinterpretations.

A key issue for future researchers will be how to elicit probability judgments. We found that people updated their estimates of the likelihood of guilt far less when estimating odds than when using the log scale, and consequently that their judgments appeared more coherent (i.e., consistent with Bayesian norms) when they used the log scale. We suspect that the odds measure restricts participants' range of responses in ways that make their judgments appear less coherent than they actually are and, accordingly, we believe that the log scale is a better—that is, more accurate—way to elicit probability judgments, although our findings do not allow us to rule out the alternative interpretation that the log scale induces exaggerated estimates that make people's judgments appear more coherent than they are.

In light of uncertainty about how best to elicit probability judgments, researchers should consider using multiple measures (as we did in this experiment) to avoid mistakenly concluding that a phenomenon is an inherent property of human judgment when it is actually an artifact of a particular method of measurement. Our uncertainty about which elicitation method is best should not deter us from using such methods to assess the relative weight that people give to different pieces of evidence. One yardstick may be too long, another may be too short, but if both yardsticks tell us that item A is longer than item B, we can be confident that item A is indeed longer. For example, in this experiment the odds measure and the log scale both told us that people gave more weight to DNA than shoeprint evidence, making us very confident in that finding.

Another methodological issue is how to measure changes in probability judgments. In this experiment we used a within-subjects method—asking participants to judge the chances of the defendant's guilt before and after receiving the forensic evidence. Martire et al. (2013, 2014) took the same approach, whereas Thompson et al. (2013) used a between-subjects method in which the judgments of participants who received the forensic evidence were compared with the judgments of another group who did not (although both groups received the same nonforensic evidence). The within-subjects method has the advantage of allowing assessment of how each participant responded to the evidence, whereas the between-subjects method allows only group-level comparisons. But the within-subjects method requires each participant to make and then update a preliminary judgment, which conceivably might affect their subsequent judgments through anchoring effects or other mechanisms. We think it would be wise for future researchers to pursue both within and between-subjects methods to address this issue.

The complexity of our findings suggests that the problem of how "best" to present forensic evidence to lay audiences may not have a single, simple solution. The presentation format may interact in unexpected ways with people's expectations or perceptions of the evidence, such that a presentation format that appears to work well for one type of evidence (e.g., use of LRs and VEs to describe the strength of DNA in this study) may work poorly with another type of evidence (e.g., use of LRs and VEs to describe shoeprint evidence in this study). Before making policy recommendations, researchers should test the generalizability of their findings in a variety of ways across a variety of contexts.

References

Association of Forensic Science Providers. (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice, 49,* 161–164. http://dx.doi.org/10.1016/j.scijus.2009.07.004

Balding, D. J., & Donnelly, P. (1994, October). The prosecutor's fallacy and DNA evidence. *Criminal Law Review,* 711–721.

Berger, C. (2010). Criminalistics is reasoning backwards: Logically correct reasoning in forensic reports and in the courtroom. *Nederlands Juristenblad,* (Feb 4, 2010), 784–789.

Buckleton, J. (2005). A framework for Interpreting Evidence. In J. Buckleton, C. M. Triggs, & S. J. Walsh (Eds.), *Forensic DNA evidence interpretation* (pp. 27–63). Boca Raton, FL: CRC Press.

Cole, S. A. (2014). Individualization is dead, long live individualization! Reforms of reporting practices for fingerprint analysis in the United States. *Law Probability and Risk, 13,* 117–150. http://dx.doi.org/10.1093/lpr/mgt014

Cook, R., Evett, I. W., Jackson, G., Jones, P. J., & Lambert, J. A. (1998). A model for case assessment and interpretation. *Science & Justice, 38,* 151–156. http://dx.doi.org/10.1016/S1355-0306(98)72099-4

de Keijser, J., & Elffers, H. (2012). Understanding of forensic expert reports by judges, defense lawyers and forensic professionals. *Psychology, Crime & Law, 18,* 191–207. http://dx.doi.org/10.1080/10683161003736744

Evett, I. W. (1995). Avoiding the transposed conditional. *Science & Justice, 35,* 127–131. http://dx.doi.org/10.1016/S1355-0306(95)72645-4

Evett, I. W. (1998). Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice, 38,* 198–202. http://dx.doi.org/10.1016/S1355-0306(98)72105-7

Faigman, S., & Baglioni, A. (1988). Bayes' theorem in the trial process: Instructing jurors on the value of statistical evidence. *Law and Human Behavior, 12,* 1–17. http://dx.doi.org/10.1007/BF01064271

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684–704. http://dx.doi.org/10.1037/0033-295X.102.4.684

Goodman, J. (1992). Jurors' comprehension and assessment of probabilistic evidence. *The American Journal of Trial Advocacy, 16,* 361.

Haack, S. (2014). *Evidence matters: Science, proof and truth in law*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139626866

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Medicine. Communicating statistical information. *Science, 290,* 2261–2262. http://dx.doi.org/10.1126/science.290.5500.2261

Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General, 128,* 3–31. http://dx.doi.org/10.1037/0096-3445.128.1.3

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511809477

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3,* 430–454. http://dx.doi.org/10.1016/0010-0285(72)90016-3

Kaye, D. H., Hans, V. P., Dann, M. B., Farley, E., & Albertson, S. (2007). Statistics in the jury box: How jurors respond to mitochondrial DNA match probabilities. *Journal of Empirical Legal Studies, 4,* 797–834. http://dx.doi.org/10.1111/j.1740-1461.2007.00107.x

Kaye, D. H. & Koehler, J. J. (1991). Can jurors understand probabilistic evidence? *Journal of the Royal Statistical Society, Series A, 154,* 75–81. http://dx.doi.org/10.2307/2982696

Kaye, D. H., & Sensabaugh, G. F. (2011). Reference guide on DNA evidence. In J. Cecil (Ed.), *Reference manual on scientific evidence* (pp. 485–576). Washington, DC: Federal Judicial Center.

Koehler, J. J. (1993). Error and exaggeration in the presentation of DNA evidence at trial. *Jurimetrics Journal, 34,* 21–39.

Koehler, J. J. (2001). When are people persuaded by DNA match statistics? *Law and Human Behavior, 25,* 493–513. http://dx.doi.org/10.1023/A:1012892815916

Koehler, J. J., Chia, A., & Lindsey, S. (1995). The random match probability (RMP) in DNA evidence: Irrelevant and prejudicial? *Jurimetrics Journal, 35,* 201.

Koehler, J. J., & Macchi, L. (2004). Thinking about low-probability events. An Exemplar-Cuing theory. *Psychological Science, 15,* 540–546. http://dx.doi.org/10.1111/j.0956-7976.2004.00716.x

Koehler, J., & Saks, M. (2010). Individualization claims in forensic science: Still unwarranted. *Brooklyn Law Review, 75,* 1187.

Lempert, R. O. (1977). Modeling relevance. *Michigan Law Review, 75,* 1021–1057. http://dx.doi.org/10.2307/1288024

Lieberman, J., Carrell, C., Miethe, T., & Krauss, D. (2008). Gold vs. platinum: Do jurors recognize the superiority and limitations of DNA evidence compared to other types of forensic evidence? *Psychology Public Policy and Law, 14,* 27–62. http://dx.doi.org/10.1037/1076-8971.14.1.27

Lynch, M., Cole, S., McNally, R., & Jordan, K. (2009). *Truth machine: The contentious history of DNA fingerprinting*. Chicago, IL: University of Chicago Press. ISBN: 978022649806.

Martire, K. A., Kemp, R. I., Sayle, M., & Newell, B. R. (2014). On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect. *Forensic Science International, 240,* 61–68. http://dx.doi.org/10.1016/j.forsciint.2014.04.005

Martire, K. A., Kemp, R. I., Watkins, I., Sayle, M. A., & Newell, B. R. (2013). The expression and interpretation of uncertain forensic science evidence: Verbal equivalence, evidence strength, and the weak evidence effect. *Law and Human Behavior, 37,* 197–207. http://dx.doi.org/10.1037/lhb0000027

McQuiston-Surrett, D., & Saks, M. J. (2008). Communicating opinion evidence in the forensic identification sciences: Accuracy and impact. *The Hastings Law Journal, 59,* 1159–1190.

McQuiston-Surrett, D., & Saks, M. J. (2009). The testimony of forensic identification science: What expert witnesses say and what factfinders hear. *Law and Human Behavior, 33,* 436–453. http://dx.doi.org/10.1007/s10979-008-9169-1

Morrison, G. S. (2011). The likelihood-ratio framework and forensic evidence in court: A response to R v T. *International Journal of Evidence and Proof, 15,* 1–29.

Murphy, E., & Thompson, W. C. (2010). Understanding potential errors and fallacies in forensic DNA statistics: An amicus brief in McDaniel v. Brown. *Criminal Law Bulletin, 46,* 709–757.

Nance, D. A., & Morris, S. B. (2002). An empirical assessment of presentation formats for trace evidence with a relatively large and quantifiable random match probability. *Jurimetrics Journal, 42,* 403–438.

Nance, D. A., & Morris, S. B. (2005). Juror understanding of DNA evidence: An empirical assessment of presentation formats for trace evidence with a relatively small random match probability. *The Journal of Legal Studies, 34,* 395–444. http://dx.doi.org/10.1086/428020

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

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the Story Model for juror decision making. *Journal of Personality and Social Psychology, 62,* 189–206. http://dx.doi.org/10.1037/0022-3514.62.2.189

Read, S. J., & Simon, D. (2012). Parallel constraint satisfaction as a mechanism for cognitive consistency. In B. Gawronsky & F. Strack (Eds.), *Cognitive consistency: A fundamental principle in social cognition* (pp. 66–87). New York, NY: Guilford Press.

Redmayne, M., Roberts, P., Aitken, C., & Jackson, G. (2011). Forensic science evidence in question. *Criminal Law Review, 5,* 347–356.

tapraid5/lhb-lhb/lhb-lhb/lhb00215/lhb0211d15z | xppws | S=1 | 4/16/15 | 16:32 | Art: 2014-2030

Robertson, B., & Vignaux, G. A. (1995). *Interpreting evidence: Evaluating forensic science in the courtroom*. New York, NY: Wiley.

Robertson, B., Vignaux, G. A., & Berger, C. E. H. (2011). Extending the confusion about Bayes. *The Modern Law Review, 74,* 444–455. http://dx.doi.org/10.1111/j.1468-2230.2011.00857.x

Schklar, J., & Diamond, S. S. (1999). Juror reactions to DNA evidence: Errors and expectancies. *Law and Human Behavior, 23,* 159–184. http://dx.doi.org/10.1023/A:1022368801333

Schum, D. (1994). *The evidential foundations of probabilistic reasoning*. New York, NY: Wiley.

Schum, D., & Martin, A. (1982). Formal and empirical research on cascaded inference in jurisprudence: A summary. *Law & Society Review, 17,* 105–142. http://dx.doi.org/10.2307/3053534

Smith, B. C., Penrod, S. D., Otto, A. L., & Park, R. C. (1996). Jurors' use of probabilistic evidence. *Law and Human Behavior, 20,* 49–82. http://dx.doi.org/10.1007/BF01499132

Taroni, F., Aitken, C., Garbolino, P., & Biedermann, A. (2006). *Bayesian networks and probabilistic inference in forensic science*. West Sussex, UK: Wiley. http://dx.doi.org/10.1002/0470091754

Thompson, W. C. (1989). Are juries competent to evaluate statistical evidence? *Law and Contemporary Problems, 52,* 9–41. http://dx.doi.org/10.2307/1191906

Thompson, W. C. (2012). Discussion paper: Hard cases make bad law: Reactions to R v. T. *Law Probability and Risk, 11,* 347–359. http://dx.doi.org/10.1093/lpr/mgs020

Thompson, W. C. (2013). Forensic DNA evidence: The myth of infallibility. In S. Krimsky & J. Gruber (Eds.), *Genetic explanations: Sense and nonsense* (pp. 227–255). Boston, MA: Harvard University Press.

Thompson, W. C., & Cole, S. A. (2007). Psychological aspects of forensic identification evidence. In M. Costanzo, D. Krauss, & K. Pezdek (Eds.), *Expert psychological testimony for the courts* (pp. 31–68). Mahwah, NJ: Erlbaum.

Thompson, W. C., Kaasa, S. O., & Peterson, T. (2013). Do jurors give appropriate weight to forensic identification evidence? *Journal of Empirical Legal Studies, 10,* 359–397. http://dx.doi.org/10.1111/jels.12013

Thompson, W. C., & Schumann, E. (1987). Interpretation of statistical evidence in criminal trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy. *Law and Human Behavior, 11,* 167–187. http://dx.doi.org/10.1007/BF01044641

Thompson, W. C., Taroni, F., & Aitken, C. G. (2003). How the probability of a false positive affects the value of DNA evidence. *Journal of Forensic Sciences, 48,* 47–54.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124.

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511809477.011

Received July 14, 2014
Revision received March 30, 2015
Accepted March 30, 2015 ■