2011

If the Shoe Fits They Might Acquit: The Value of Forensic Science Testimony

Jonathan Koehler
Northwestern University School of Law, jay.koehler@northwestern.edu

If the Shoe Fits They Might Acquit:

The Value of Forensic Science Testimony

Jonathan J. Koehler*
Northwestern University School of Law
January 12, 2011

* Beatrice Kuhn Professor of Law, Northwestern University School of Law, 357 E. Chicago Ave., Chicago, IL 60611. 312-503-4469 (w), jay.koehler@northwestern.edu. I thank David Kaye and Molly Mercer for helpful comments.

ABSTRACT

The probative value of forensic science evidence (such as a shoeprint) varies widely depending on how the evidence and hypothesis of interest is characterized. This paper uses a likelihood ratio (LR) approach to identify the probative value of forensic science evidence. It argues that the "evidence" component should be characterized as a "reported match," and that the hypothesis component should be characterized as "the matching person or object is the source of the crime scene sample." This characterization of the LR forces examiners to incorporate risks from sample mix-ups and examiner error into their match statistics. But how will legal decision makers respond to this new characterization of a match statistic? An experiment with 315 jury-eligible jurors who received a shoeprint match statistic in a burglary case finds that, contrary to normative theory, people are *more* persuaded by a statistic that ignores various error risks than by a more probative statistic that expressly takes those risks into account. The experiment also finds that jurors are relatively unresponsive to exposure of those risks by a defense attorney on cross-examination. These results support and extend previous research that finds many people are confused about how to evaluate the risk of error in forensic match statistics.

I. INTRODUCTION

When a match is found between forensic science material recovered from a crime scene (e.g., a fingerprint, a strand of hair, a shoeprint, etc.) and a reference sample provided by a suspect, the match is widely regarded as powerful evidence that the suspect is the source of the recovered material and, perhaps, guilty of the crime as well. In many forensic science subfields, experts use strong, conclusory language to describe their findings. For example, fingerprint examiners commonly claim that a person who matches a latent print recovered from a crime scene is the source of that print to the exclusion of all other people in the world. However, significant questions remain about the underlying probative value of a reported match.

A 2009 report by the distinguished National Research Council Committee on Identifying the Needs of The Forensic Science Community [hereinafter *NRC Report*] concluded that "forensic science professionals have yet to establish either the validity of their approach or the accuracy of their conclusions" (NRC, 2009, p. 1-14). As detailed in the NRC Report, the "science" part of forensic science has not kept pace with the extraordinary claims made on its behalf. As a result, jurors have little idea what the chance is that a forensic scientist's conclusions are wrong, how often different objects share particular characteristics,[1] or how much weight to give to the forensic science evidence as proof of identity.

This paper examines ways to assess the probative value of forensic science evidence and offers an experiment to assess whether jurors appreciate some of the subtleties associated with probative value statistics. Section II introduces some of the key concepts (reliability,

[1] DNA evidence is often the exception to this rule. In many DNA cases, scientific estimates of the frequency with which different genetic profiles occur in various subpopulations are entered into evidence. However, when the profile frequencies are extremely low, some agencies (such as the FBI) bypass frequency statistics in favor of a conclusion that the matchee is the source of the DNA.

diagnosticity, and probative value) that inform subsequent discussion of the value of forensic science evidence. Like many others, I recommend a *likelihood ratio* (LR) approach to probative value. Sections III and IV examine the impact of different characterizations of the forensic science evidence and hypotheses of interest on the corresponding LR. I suggest that the evidence is best described as a reported match between the samples the analyst tested, and that the hypothesis of interest is best described as "the matching person or object is the source of the crime scene sample." Section V reviews the literature on the impact of forensic statistics on jurors. Section VI describes an experiment that tests whether mock jurors appreciate that the probative value of a reported shoeprint match varies depending on whether the match statistic takes account of various sources of error. Section VII is a conclusion.

II. RELIABILITY, DIAGNOSTICITY AND PROBATIVITY

The concepts of reliability, diagnosticity and probativity have distinct meanings and are all important in any discussion of the value of forensic science evidence. My use of these terms below draws on prior commentary (see e.g., Kaasa et al., 2007; Koehler, 2008; Thompson & Cole, 2007).

A. Reliability

The possibility of error plays an important role in the interpretation of forensic science evidence. Even when there is some theoretical and empirical support associated with a particular method or technique, and even when the examiner is well-trained, knowledgeable, and experienced, the risk that the examiner's conclusions is mistaken must be estimated. These error estimates help a legal factfinder identify the *reliability* of the forensic science evidence. If the risk of error is very high, then the forensic science evidence is untrustworthy (i.e., unreliable) and should not be given great weight. If the risk of error is very low, then the forensic science evidence is more trustworthy. Although it is obviously it is important to know whether evidence is reliable or not, nobody knows how often, and under what conditions, mistakes are made in the forensic sciences. The NRC Report repeatedly and forcefully calls for reliability studies that quantify the risk of error across the forensic science disciplines.[2] This call should be heeded. Perhaps the most important scientific task facing the forensic science disciplines today is the systematic study of the risk of error.

B. Diagnosticity

Holding error possibilities aside, the frequency with which the observed characteristics in a print, toolmark, or other marking occurs is another important consideration in the interpretation of forensic science evidence. This frequency, which is described as the random match probability (RMP)[3] in DNA analysis, captures the *diagnosticity* of a true match[4] between some observable features of the evidentiary trace (e.g., a latent print or a toolmark) and a possible source (Kaasa et

[2] Recommendation 5 calls for "research conducted to quantify and characterize the amount of error" in forensic examinations" (NRC, 2009, p. S-18). This Recommendation also calls for "research conducted to quantify and characterize the amount of error" in forensic examinations." Recommendation 12 includes a call for the "quantification of error rates" (p. S-23). A section on error rates concludes that that "the estimation of relevant error rates are key components of the mission of forensic science" (p. 4-9). See also Edwards (2009) criticizing "the paucity of scientific research to … establish quantifiable measures of uncertainty in the conclusions of forensic analyses" (p. 2), and calling for "interdisciplinary, peer-reviewed, scientific research to determine the validity and reliability of existing disciplines" (p. 13).

[3] If a set of characteristics occurs with a frequency of 1 in X in a population, then the chance that a randomly selected member of that population will share or "match" those characteristics is also 1 in X. Hence the label "random match probability" (RMP) describes this frequency.

[4] A "true match" is one in which two labeled items (e.g., a questioned hair recovered from a crime scene and a known hair recovered from a specific suspect) actually do share an identified feature or set of features. It is distinguished from a "reported match" in that a reported match (i.e., a match claim offered by a forensic examiner) may or may not reflect a true match depending, for example, on whether a critical error was made at some point in the collection, labeling, preservation, examination or reporting process.

al., 2007; Koehler, 2008; Schum, 1994).[5] If a large proportion of people or objects share the observable features of the trace, then the diagnosticity of the evidence is low and should not be given great weight. If very few people or objects match the observable trace features, then the evidence is more diagnostic.[6] The NRC Report notes that most forensic sciences have not compiled databases that identify the frequency of various features or characteristics,[7] and calls for quantifiable research to remedy this problem.[8] Research in this area in the non-DNA forensic sciences has just begun (Neumann et al., 2006, 2007, 2010).

C. Probativity

The probative value of forensic science evidence is obtained by combining reliability and diagnosticity into a single metric. The LR is a widely accepted metric for identifying the probative value of evidence in the forensic sciences (Aitken, 1995; Aitken & Taroni, 2004; Bozza et al. 2008; Champod & Evett, 2001; Evett, 1984; Evett & Weir, 1998; Foreman et al., 2003; Neumann et al., 2006; Robertson & Vignaux, 1998) and other areas including psychology

[5] The term "diagnosticity" as used here is approximately identical to the probativity of a *true match*, but it does not necessarily describe the probativity of a *reported match*. The probativity of a reported match takes reliability (i.e., the possibility of error) into account, whereas the probativity of a true match does not.

[6] Evidence is diagnostic with respect to a hypothesis. The implicit hypothesis here is: "The suspect is the source of the crime scene evidence."

[7] Regarding shoeprints and tire tracks, the NRC Report said, "[T]he committee is not aware of any data about the variability of class or individual characteristics or about the validity or reliability of the method. Without such population studies, it is impossible to assess the number of characteristics that must match in order to have any particular degree of confidence about the source of the impression" (NRC, 2009, p. 5-17). The NRC Report offered similar comments about bitemarks ("no large population studies have been conducted," p. 5-37), hair ("[n]o scientifically accepted statistics exist about the frequency with which particular characteristics of hair are distributed in the population," p. 5-25), and fingerprints ("more research is needed regarding the discriminating value of the various ridge formations and clusters of ridge formations," p. 5-13).

[8] Recommendation 3 in the NRC Report calls for research on "[t]he development of quantifiable measures of uncertainty in the conclusion of forensic analyses" (NRC, 2009, p. 6-6). This recommendation was echoed by the National Institute of Standards & Technology Office of Law Enforcement Standards and the National Institute of Justice in their proposal to form an Expert Working Group on Human Factors in Latent Print Analysis. According to the proposal, the Expert Working Group would examine the role of human factors in fingerprint examination and "evaluate various approaches to numerically quantifying measurement uncertainty within forensic science analysis" (Morgan, Stolorow, & Taylor, 2008).

(Edwards, Lindman, & Savage, 1963; Fischhoff, & Beyth-Marom, 1983; Meehl & Rosen, 1955), medicine (Black & Armstrong, 1986; Jaeschke, Guyatt, & Sackett, 1994; Lloyd, Talbot, & Lawson, 1998), and law (Finkelstein & Levin, 2003; Friedman & Park, 2003; Kadane & Schum, 1996; Kaye, 1986; Kaye & Koehler, 2003; Lyon and Koehler; McCormick, 1999; Robertson & Vignaux, 1995, Wagenaar, 1988; but see Allen & Pardo, 2007 for an opposing view).

The LR associated with Evidence E and Hypothesis H is $\frac{P(E \mid H)}{P(E \mid \overline{H})}$. This ratio quantifies the relative probabilities of observing Evidence E under two competing hypotheses, H and $\overline{H}$. The numerator of the LR is the *true positive rate* of the forensic science test, and the denominator of the LR is the *false positive rate*.[9] When the true positive rate is high and the false positive rate is very low the value of the LR is largely controlled by the false positive rate. Small variations in the LR numerator will only have a small effect on the overall LR, whereas small changes in the LR denominator will have much larger effects.

To illustrate, consider a LR in which the numerator is .98 and the denominator is .000001. The value of the LR is $\frac{.98}{.000001}$ = 980,000. If the LR numerator increased by an absolute 1% (from .98 to .99) the new LR would be just slightly higher (990,000). However, if the LR denominator (i.e., the false positive rate) increased by an absolute 1% (from .000001 to .010001), the new LR would drop from 980,000 to about 98.[10] In short, the probative value of forensic science evidence depends critically on the denominator of the LR, i.e., the false positive rate.

[9] The numerator and denominator of the LR may also be described as the sensitivity and "1-the specificity" of the forensic test, respectively.

[10] The new LR would be $\frac{.98}{.010001} = 97.99$.

Identifying the false positive rate of interest – and the probative value of the evidence – requires considerable clarity about (a) what evidentiary claim is being made, and (b) which hypothesis is under consideration. Different evidentiary claims and different hypotheses will be associated with different false positive rates and different LRs (Koehler, 1996). Unfortunately, forensic scientists and those who write about forensic science do not always appreciate this point. In the following sections, the "evidence" and "hypothesis" elements that appear in the numerator and denominator of the LR for forensic science evidence are examined. This exercise supports a conclusion that *the value of the LR, hence the probativity of the forensic science evidence itself, depends critically on subtle variations in the descriptions of the evidence and hypotheses.*

III. WHAT IS "THE EVIDENCE?"

Suppose there has been a burglary at an Ace Hardware store and that a shoeprint that allegedly matches a suspect's tennis shoes is recovered at or near the crime scene. Construction of a LR $\frac{P(E\,|\,H)}{P(E\,|\,\overline{H})}$ to identify the strength of the shoeprint evidence[11] requires clarity about what is meant by the shoeprint evidence (E) and what the hypothesis (H) is.

[11] The way forensic match conclusions are offered in court varies by subfield and the guidelines the subfields offer for themselves are in a state of flux. Professional guidelines encourage shoeprint examiners to offer one of seven conclusions: 1. Unsuitable (lacks sufficient detail for a meaningful comparison), 2. Elimination (definite exclusion), 3. Identification (definite conclusion of identity), 4. Probably made (very high degree of association), 5. Could have made (significant association of multiple class characteristics), 6. Inconclusive (limited association of some characteristics), or 7. Probably did not make (very high degree of nonassociation) (Scientific Working Group, 2006). However, some examiners only offer one of four conclusions: 1. Elimination, 2. Identification, 3. Inclusion, or 4. Inconclusive (Hammer, 2010). At present, identifications and various other forms of shoeprint matches are generally offered without any form of quantification. However, this is likely to change in the future as databases are constructed and probabilistic models (including LR approaches) are refined and incorporated into the professional standards.

Consider first the shoeprint evidence (E). At first blush, one might assume that a proper way to characterize E is as follows: E = “The shoeprint recovered from the Ace Hardware burglary scene matches the suspect’s tennis shoes.” But this characterization assumes too much. It assumes that there is no uncertainty about (a) the items in question, or (b) the match itself. I explain these two points below.

Regarding the items in question, the shoeprint examiner may not be a position to know that the recovered print was lifted from the Ace Hardware crime scene or that the matching tennis shoe belongs to the suspect. All the examiner really knows (in most instances) is that the shoeprint and tennis shoe that was examined appear to be a match. But what if a case mix-up or labeling error occurred prior to the shoeprint examination at the forensic laboratory? What if the shoeprint actually came from some place other than the Ace Hardware crime scene, or the examined tennis shoe actually belonged to someone other than the suspect? If such an error occurred prior to the forensic analysis – and these types of errors have been documented in proficiency tests[12] and in casework (see e.g., Thompson, 1995, 1996) – then it would not be correct to say that the forensic analysis showed that the shoeprint recovered from the Ace Hardware crime scene matches the suspect’s tennis shoes.

Regarding the match itself, it is important to distinguish between a “true match” and a “reported match.” A true match occurs when two items truly share a set of common features and are otherwise indistinguishable when examined with the aid of a particular technology or method. A “reported match” occurs when an examiner *says* two items share a set of common features and are otherwise indistinguishable. The difference between a true match and a reported

[12] See various forensic summary reports prepared and administered by Collaborative Testing Services at http://www.ctsforensics.com/reports/main.aspx.

match is structurally identical to the difference between the factual statement “Jim robbed the bank” and the eyewitness’s assertion that Jim robbed the bank. In the latter case, the eyewitness may be mistaken for any number of reasons (poor eyesight, short observation period, Jim looks a lot like the actual robber, etc.). Likewise, a technological error, misreading, recording error, or other error may have occurred, and the tested items may not be a true match after all. This is not to say that all reported matches should be regarded with great suspicion. But the accuracy of the match report is an issue for the factfinder, not the one who offers the report.[13]

Where does this leave us in terms of defining “the evidence” in cases involving forensic science evidence? For reasons stated above, we should reject “The shoeprint recovered from the Ace Hardware burglary scene matches the suspect’s tennis shoes” in favor of a statement along these lines: *“The examiner reports a match between the shoeprint and tennis shoes that he/she tested.”* More generally, forensic science evidence may be characterized as a *report* that the *samples tested* match. Henceforth, I abbreviate this characterization of the evidence as a “*reported match*.”

IV. WHAT IS “THE HYPOTHESIS?”

The probativity of a reported match varies depending on the hypothesis of interest. This point is underappreciated. Consider a case that includes a reported match between a questioned hair and a known hair. Several hypotheses of interest could be identified. Three such hypotheses are

[13] A similar point appears in a footnote in *The New Wigmore: A Treatise on Evidence* by Leonard et al. (2010). In a chapter on forensic science and identity, the authors write “Technically, the evidence is an assertion of what the laboratory has found. . . . This does not mean that the finding is correct …” (Chapter 12, p. 15).

identified in Figure 1.[14] The first two hypotheses concern what truly matches with what, and the third hypothesis is a source claim. For ease of discussion, I refer to H_1, $H_{2,}$ and H_3 below as The Examiner's Hypothesis, The Laboratory Director's Hypothesis, and The Factfinder's Hypothesis, respectively.

INSERT FIGURE 1 ABOUT HERE

In casual conversation or in oral testimony, it is easy to treat the three hypotheses listed in Figure 1 as interchangeable. But this would be a mistake. The probative value of the evidence depends critically on which of these hypotheses is under consideration, because each is located at a different point on the inferential chain (see Figure 2).

Figure 2 shows that once investigators have determined that a pair of tested samples (e.g., a questioned and known hair) reportedly match, this finding may be used to draw inferences about a variety of hypotheses (H_1 – H_5). The hypothesis that is most closely connected with the reported match evidence is the proposition that the tested samples actually do match (H_1). The hypothesis that is further down the inferential chain is the proposition that the suspect is guilty of the crime (H_5). In general, the probative value of the evidence weakens as one moves down the

[14] Forensic science researcher Ian Evett and his colleagues have identified a "hierarchy of propositions" that describes some of the hypotheses that that might be used to identify the significance of forensic science evidence. The three primary propositions are source level (e.g., "the suspect is the source of the semen found at the crime scene"), activity-level (e.g., "the suspect had intercourse with the victim"), and offense-level (e.g., "the suspect raped the complainant") (Champod, Evett, & Jackson, 2004; Cook, Evett, Jackson, Jones, & Lambert, 1998a, 1998b, 1999; Evett, Jackson, Lambert, & McCrossan, 2000; Evett, Jackson, & Lambert, 2000). Activity-level and offense level propositions are higher level propositions. Evett and colleagues have argued that higher level propositions provide "greater added value to the court" (Foreman, Champod, Evett, & Pope, 2003; see also Hicks, 2009). I do not consider these higher level propositions here because LR construction under these hypotheses will commonly incorporate non-forensic considerations (e.g., strength of the defendant's alibi) that belong to the factfinder's domain.

inferential chain. This occurs because, at each inferential stage, additional risks to the validity of the inference must be considered. If those additional risks are non-zero, then the inference becomes less certain than the one that preceded it.[15]

INSERT FIGURE 2 ABOUT HERE

A. The Examiner's Hypothesis (H_1)

From the vantage point of an examiner who is completing an exercise, The Examiner's Hypothesis (*H_1: The tested samples truly match*) will often be the hypothesis of interest. When an examiner reports that a pair of items matches, one benchmark or ground truth criterion for the examiner's scientific judgment is whether the pair of items *that he/she tested* is a true match. To the extent that the examiner reports matches on tested pairs that really do match, and reports exclusions (i.e., non-matches) on tested pairs that really do not match, the examiner's judgment is probative of H_1.

An inference from a reported match between test samples to the hypothesis that the tested samples truly match is limited by the set of risks described above, namely, technological error and human error (see Figure 2). Technology errors could occur, for example, when an item of equipment that the forensic scientist relies on provides faulty information. Human error includes such mistakes as recording errors, coding errors, misinterpretation of test results, contamination,

[15] If the additional risks that are encountered as one moves down the inferential chain are several orders of magnitude less than the preceding risks then, as a practical matter, the inference to the hypothesis in question is about as strong as the one that preceded it. Thus, in cases of a DNA match where the risk of coincidental match is very small, the strength of the evidence of a reported DNA match provides evidence for H_3 ("the suspect is the source of the DNA recovered from the crime scene sample") is about as strong as it is for H_2 ("the DNA recovered from the crime scene and the suspect's DNA truly match").

etc. If the category of human error is defined to include intentional as well unintentional errors, then the risk of fraud should be considered here as well.

Under H_1, the probative value of an examiner's report (as measured by the LR) provides insight into the accuracy with which the examiner has measured or identified the features in question. Large LRs increase our confidence in the examiner's accuracy. However, LRs that are computed under H_1 are unaffected by the frequency of the observed characteristics in a reference population. That is, the LR $\frac{P(Reported\ Match\ |The\ Tested\ Samples\ Truly\ Match)}{P(Reported\ Match\ |\ The\ Tested\ Samples\ Do\ Not\ Truly\ Match)}$ remains the same regardless of whether the matching items are rare (e.g., a multi-locus DNA profile that occurs in fewer than one in one hundred million people) or common (e.g., a partial shoeprint that occurs in about in every ten pairs of tennis shoes). Because a factfinder should consider the possibility that a reported match is merely coincidental, the Examiner's Hypothesis (H_1) has minimal value at trial.

B. The Laboratory Director's Hypothesis (H_2)

Unlike the examiner who is completing an exercise, a Laboratory Director must be concerned about more than whether the samples tested by one of his or her examiners match. Assuming that the person(s) responsible for collecting, labeling, and preserving the samples that will soon be tested is a member of the Laboratory Director's staff, the Laboratory Director must also be concerned with whether the tested samples are what they are presumed to be (*H_2: The sample from the crime scene and the reference sample from the suspected source truly match*). That is, the Laboratory Director (unlike the examiner who is testing the samples) must also be concerned with whether the tested samples are, in fact, from the crime scene in question and from the referent person in question.

The Laboratory Director's Hypothesis (H_2) is distinguished from the Examiner's Hypothesis (H_1) in that its focus is on whether the crime scene sample and reference sample actually match, rather than simply on whether whatever the samples the examiner tested actually match. The Laboratory Director's Hypothesis thus entails moving further out on the inferential chain (see Figure 2).

Potential chinks in the inferential chain at this stage include sample contamination and the possibility that the examiner tested something other than the appropriate crime scene and reference samples. When the wrong samples are tested, the fault may lie either with the examiner or with those who collected, labeled, stored, or retrieved the evidence prior to passing it off to the examiner for testing.

LRs that are constructed under H_2 suffer the same key shortcoming as those constructed under H_1, namely, that they are unaffected by the frequency of the matching characteristics. In cases where the chance of a coincidental match is relatively large, this limitation is important. In cases where the forensic science evidence is highly discriminating (e.g., DNA analyses), this limitation is less important.

C. The Factfinder's Hypothesis (H_3)

The Factfinder's Hypothesis (*H_3: The matching person or object is the source of the crime scene sample)* is usually the forensic science hypothesis of interest for the legal factfinder. H_3 is a *source hypothesis* rather than a match hypothesis of the sort described by H_1 and H_2.

Support for H_3 requires consideration of the rareness of the observed matching characteristics in the potential source population.[16] If the characteristics are relatively common, then the evidence may not be very diagnostic of H_3. Suppose, for example, that a 99.99% reliable hair examiner[17] reports that a questioned hair recovered from a crime scene and a known hair taken from a suspect are both black. Suppose further that, based on other evidence, the source of the questioned hair is known to be a Hawaiian native, and that the frequency of black hair among Hawaiians is 92% (Farabee, 1922, p. 170). Whereas the LRs associated with H_1 and H_2 for this reported hair match are approximately 10,000:1,[18] the LR associated with H_3 is only a smidgen larger than 1:1.[19] In other words, the identical evidence is highly probative under the Examiner's Hypothesis (H_1) and Laboratory Director's Hypothesis (H_2), but has little probative value under the all-important Factfinder's Hypothesis (H_3). Indeed, courts likely should disallow the evidence if offered in support of H_3 on grounds that the evidence fails the basic relevance standard in the Federal Rules of Evidence.[20]

The LR associated with H_3 is very low for the hair evidence due to the rather large chance that the reported match is merely coincidental (92%). The fact that the examiner is an extremely reliable reporter of hair color (99.99%) does not adequately compensate for the fact

[16] The potential source population for a human characteristic is "the group of people who might reasonably be the source of the recovered trace evidence" (Koehler, 1993a, p. 227).

[17] Assume that a 99.99% reliable examiner commits false positive errors about hair color 0.01% of the time and commits false negative errors about hair color 0.01% of the time.

[18] The LR associated with H_1 is $\frac{P(E \mid H_1)}{P(E \mid \overline{H_1})} = \frac{.9999}{.0001} = 9{,}999.$ The LR associated with H_2 will be a bit lower due to the small chance that the hair samples tested may not have been the actual crime scene and references samples.

[19] The LR associated with H_3 is

$$\frac{\text{P(Reported Match | The Matching Person is the Source of the Crime Scene Sample)}}{\text{P(Reported Match | The Matching Person is Not the Source of the Crime Scene Sample)}} = \frac{.9999}{.92} = 1.09.$$

[20] "Relevant evidence means evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence" (Federal Rule of Evidence 401).

that many people have black hair. Put another way, the false positive rate associated with H_3 is very high (92%), and this value severely limits the probative value of the evidence with respect to the source hypothesis that is of interest to the factfinder.

D. Potential Objection

One objection to using the Factfinder's Hypothesis (H_3) rather than one of the other hypotheses when computing the LR is that H_3 may not be the hypothesis that *the examiner* has in mind when he/she produces a report or testifies in court. For example, match reports and testimony from hair examiners often do not entail source claims. So why, then, should the all-important denominator of the LR – the false positive rate for hair comparison – be measured as the rate at which examiners report matches on hairs that came from different sources rather than the rate at which examiners report matches on hairs that do not truly match on the relevant characteristics?

The answer can be found by considering the purpose for which hair comparison evidence – or any forensic science evidence – is generally offered as proof at trial. The evidence is offered to support the hypothesis that the matchee is, in fact, the source of the evidence.[21] This is the hypothesis that *factfinders* do and should use when evaluating the evidence. Factfinders do not consider the evidence of the reported match against the hypothesis that the examiner has accurately identified various hair characteristics, nor is this the purpose for which the evidence is offered. This narrower hypothesis may be of great interest to the examiner and the examiner's manager because they are likely to be more concerned with whether the examiner can perform a hair comparison competently. But this is not the factfinder's concern.

[21] Of course, proof that the defendant is the source of a hair recovered from a crime scene is often important because it provides circumstantial proof of guilt by placing the defendant at the scene.

This point is a source of confusion among those who have commented on the LR approach to forensic science evidence. Returning to our hair example, an examiner who identifies a hair as black[22] may be perfectly trustworthy and reliable in his/her color identification. The LR that identifies the probative value of a *report* that a recovered hair is black with respect to the hypothesis "The tested hairs truly are black" may be 9,999.[23] But such reliability (on color) is an insufficient basis on which to admit forensic science evidence at trial. Evidence of a match (based on color or other characteristics) between a questioned and known hair is ordinarily useful only insofar as it provides evidence about who is the source of that hair.[24] In our admittedly stylized example, the false positive rate is very high when measured against the source hypothesis (.92). Here, evidence of a hair color match has minimal probative value. The fact that the examiner has a 99.99% accuracy rate for typing hair color isn't a particularly useful statistic when nearly every potential source has the same color hair as the sample in question. Indeed, it could be quite misleading even to offer the 99.99% accuracy rate to jurors without detailed cautionary instructions because jurors may mistakenly rely on that impressive-sounding rate as an indicator of the probative value of the evidence with respect to the source hypothesis.

V. IMPACT OF FORENSIC STATISTICS: JUROR RESEARCH

Empirical research on how people react to statistical information gives reason to be concerned about how well jurors understand subtle – but important – differences in the presentation of

[22] Obviously hair comparison covers more than hair color. This example is intentionally simplified.

[23] see fn 17 (indicating that the LR associated with H_1 is 9999).

[24] This discussion assumes that source identity is probative of guilt. In some cases, such as those where the presence of hairs from a defendant is equally consistent with guilt and innocence, hair match evidence would be irrelevant, even if it is highly probative of source.

forensic statistics at trial. Judgment and decision making research indicates that people (a) do not extract as much information from new evidence as the data objectively warrant, (b) are slow to revise incorrect probabilistic hypotheses, (c) misattribute probative value to diagnostically worthless information, (d) underutilize statistical base rates, and (e) confuse likelihoods and posteriors (for reviews and collections, see Gilovich, Griffin, & Kahneman 2002; Kahneman, Slovic & Tversky, 1982; Kahneman & Tversky, 2000; Nisbett & Ross, 1980; Saks & Kidd, 1980; Edwards & von Winterfeldt, 1986; Yates, 1990). A similar set of shortcomings was observed in studies of how jurors react to forensic statistics in the pre-DNA era (for reviews, see Kaye & Koehler, 1991; Thompson, 1989).

When DNA evidence began appearing at trial with some regularity in the 1990s, researchers began studying how mock jurors think about the small RMPs that accompanied reports of DNA matches. A clear pattern of results emerged. The mock jurors (a) fell prey to logical fallacies when reasoning with statistical evidence (Nance & Morris, 2002), (b) failed to appreciate the role that error plays in interpreting the value of a reported match (Koehler, Chia, & Lindsey, 1995; Schklar & Diamond, 1999; but see Lieberman et al., 2008, experiment 3), (c) undervalued DNA evidence (Nance & Morris, 2002, 2005; Kaye et al. 2007; Taroni & Aitken, 1998), and (d) were sensitive to the form in which the statistics and other parts of the forensic science evidence are presented (Koehler, 1996, 2001a, 2001b; Koehler & Macchi, 2004; Lindsey, Hertwig, & Gigerenzer, 2003; McQuiston-Surret & Saks, 2008, 2009). Group (jury) deliberation does not seem to diminish most of these problems substantially (Kaye et al., 2007; Koehler & Thompson, 2006). Even more troubling, some data suggest that the expert witnesses themselves exaggerate and reason poorly with forensic evidence (Koehler, 1993b; Garrett & Neufeld, 2009).

VI. EMPIRICAL STUDY

A. Introduction

There is a growing interest in a quantitative approach to non-DNA forensic science evidence (Neumann et al., 2010). This is a positive development because it represents a movement away from exaggerated "individualization" claims that commonly appear in forensic science reports and courtroom testimony.[25] However, the notion that the quantitative measures introduced at trial should be explicit about assumptions and definitions has not yet caught on.

At present, when a forensic scientist or researcher quantifies the significance of a match, the statistic of choice is typically an RMP or, when presented as a LR, the inverse of a RMP. The focus of this quantification is on identifying the frequency of the matching characteristics in one or more reference populations. This statistic can help jurors understand the plausibility of a defendant's contention that the match is purely coincidental. As Figure 2 indicates, a coincidental match is the central risk to the validity of an inference from the assumption that there exists a true match between a crime scene sample and suspect's sample (H_2) to a conclusion that the suspect is the source of the crime scene sample (H_3)

However, as noted in section III, forensic science evidence should be characterized as a *reported match* between tested samples, rather than as a true match between samples known to come from particular places or people. The true match characterization makes several key assumptions. First, it assumes that that there have not been any mix-ups involving either of the samples. Second it assumes that the examiner did not make a critical error (such as a misreading

[25] For criticism of individualization claims, see Cole (2004, 2009), Koehler and Saks (2010), National Research Council (2009), and Saks and Koehler (2008); see also Kaye (2010) calling for "a more nuanced theory of identification" (p. 1185).

of the data or a recording error) that would lead him to falsely believe that samples from different sources match.

The disparity between the way forensic science statistics are computed at trial (when they are computed at all), and the fact that such computations exaggerate the probative value of the evidence by assuming no mix-ups or examiner errors raises the following question: *will jurors respond differently to forensic science statistics depending on whether those statistics do or do not incorporate and account for various error risks?* In other words, does it make any difference to jurors what is factored into the statistic they hear, or do jurors simply use that statistic as a proxy for the probative value of the forensic science evidence? Furthermore, how effective is cross-examination about this issue likely to be?

One possibility is that the jurors' judgments about the probativity of the evidence are insensitive to the examiner's implicit characterizations of the evidence and the hypothesis. They may assign weight to forensic evidence based on a match statistic alone, with no regard for which risks that statistic does and does not account for. A second possibility is that jurors will find the forensic evidence to be more compelling when the examiner's statistic ignores certain risks and thereby moves the evidence further down the inferential chain (see Figure 2). Thus, jurors might be more persuaded by a 1 in 1,000 statistic when the mix-up and examiner error risks are ignored than when those forms of error are expressly incorporated into the statistic. A third possibility is that jurors will be most impressed by match statistics that incorporate more of the risks to valid inference. From a normative standpoint, such behavior would be appropriate. A 1 in 1,000 match statistic that incorporates multiple risks (e.g., coincidence and examiner errors) provides *stronger* proof that a matchee is the source than a 1 in 1,000 match statistic that only incorporates, say, the risk of coincidental match.

Based on research noted earlier that shows people struggle with statistical evidence, there is some reason to believe that the first and second possibilities are more likely than the third. However, I offer no strong main effect prediction for this *"risks"* variable.

Regardless of how people respond to 1 in 1,000 match statistics that incorporate varying amounts of risk, *cross-examination* that points out the shortcomings of those statistics presumably will reduce their persuasiveness. For this reason I predict a main effect for cross-examination. And because some of the match statistics will be characterized in ways that include more risk than others, I also predict a risks X cross-examination interaction in which the effects of cross-exam are larger for match statistics that incorporate fewer risks.

B. Method

1. Participants

To investigate these issues, I conducted a controlled behavioral experiment with 315 jury-eligible participants ("jurors"). Jurors were drawn from a panel of internet users affiliated with Survey Sampling International (SSI), a third party vendor that uses opt-in e-mail-based recruiting methods. E-mail invitations were sent to an age-stratified random sample of panel members with the goal of approximating the jury-eligible population. The invitations were spread over a two-week period to reduce the risk of early-respondent bias. Jurors were compensated by SSI in accordance with its policy of providing members with small guaranteed cash incentives and entry in drawings for additional cash prizes. Jurors covered a broad cross-section of the jury-eligible population on race (14% non-white), age (18-86, mean=48), educational level (26% high school graduate or less, 9% graduate degrees), and political beliefs (50% liberal, 50% conservative) and gender (53% women).

2. Materials, Procedure, & Design

Jurors were told that they were about to read a transcript of testimony from a shoeprint expert in a burglary trial. The case and testimony were hypothetical, though the testimony (but not the case) was loosely based on shoeprint testimony provided by a well-regarded expert in the criminal trial of O. J. Simpson.[26]

Prior to reading the shoeprint testimony, jurors were provided with a factual summary of the case against defendant Steven Summers. Mr. Summers was accused of burglarizing an Ace Hardware store in the Phoenix area at 10:15pm on Sunday, January 10, 2010. The case against Mr. Summers included testimony from an eyewitness who heard the store alarm and saw a person fleeing from the store on foot. When presented with a mug shot of Mr. Summers, the eyewitness said that Summers might be the man he saw, though he readily admitted that he was not at all certain. Police investigators recovered three fresh shoeprints from the area of the store where the eyewitness saw the perpetrator. The shoeprints, which matched shoes owned by the defendant, were the primary source of evidence against Mr. Summers. In his defense, Mr. Summers claims that he did not burglarize this store, nor has he ever set foot in or near this Ace Hardware store. He claims that he was home alone, five miles away from the store, at the time of burglary.

Jurors were assigned at random to one of eight experimental groups or to a control group. Six of the experimental groups formed a fully crossed 3 (*risks*: coincidence only, coincidence + mix-up, coincidence + mix-up + examiner error) X 2 (*cross-exam*: no, yes) between-subjects design that comprised the core of the empirical study. Jurors in these groups heard testimony

[26] A transcript of the shoeprint testimony in the Simpson case can be found at http://www.law.umkc.edu/faculty/projects/ftrials/Simpson/Bodziak.html.

from a shoeprint expert who reported a match between the footprints recovered from the crime scene and the soles of the defendant's tennis shoes. The experts qualified the match by stating that the risk of error was 1 in 1,000.

In some cases, the 1 in 1,000 match statistic reflected only the *risk of a coincidental match* between the defendant's shoes and other shoes. In other cases the 1 in 1,000 statistic reflected the aggregated *risks of coincidental match and a mix-up* wherein the tested shoes may have belonged to someone other than the defendant or the examined shoeprint may not have been from the Ace Hardware crime scene. In still other cases, the 1 in 1,000 statistic reflected the aggregated *risks of coincidental match, a mix-up, and examiner error* wherein the examiner himself made a "critical" error of some sort. See Appendix 1.

The experimental conditions also varied in terms of whether there was a thorough cross-examination or none at all. When cross-examination was present, the expert was questioned about the possibility that someone else wore the defendant's shoes. In addition, the expert was grilled about the sources of risk that his 1 in 1,000 match statistic did not take into account. Thus, in the coincidence only risk condition, the expert was grilled about both the possibility of sample mix-ups and examiner error; in the coincidence + mix-up risk condition, the expert was grilled about the possibility of examiner error. In all cases, the expert conceded the validity of the cross-examiner's points.

Jurors assigned to one of the two *individualization* experimental groups (cross-exam: no, yes) read testimony from an expert who did not quantify his opinion with a match probability. In these conditions, the expert simply identified the defendant's shoes as the only ones that could possibly have made the recovered shoeprints (see Appendix 1). In the cross-examination condition, the expert was grilled about *all* of the sources of risks (coincidental match, mix-up, and examiner error). The individualization conditions were included because individualization testimony is common in a number of forensic sciences (e.g., fingerprint evidence). Reactions to individualization testimony also provide a point of comparison with the quantitative conditions.

Jurors in the *control* condition did not read any testimony pertaining to matching shoeprints. They answered questions about the case immediately after reading a minimal case summary that did not mention matching shoeprints. The control condition is useful as a baseline for identifying the overall impact of forensic science testimony on judgments.

After reading the testimony, jurors answered questions pertaining to the strength of the evidence, the source of the shoeprint, and the guilt of the defendant. See Appendix 2. The response type varied by question (e.g., some 7-point Likert-type scales, some numerical values, some forced choice). After answering all of the case-related questions, jurors answered background questions (e.g., age, sex, ethnicity, education level, political views, and prior experience on juries).

C. Results

A 3 (*risks*: coincidence only, coincidence + mix-up, coincidence + mix-up + examiner error) X 2 (*cross-exam*: no, yes) between-subjects multivariate analysis of variance (MANOVA) was performed to examine the effects of risks and cross-examination on evidence strength, source, and guilt. A binary logistic regression was also conducted with risks and cross-examination as predictors and verdict as the dependent variable. The two individualization conditions and control group (no shoeprint testimony) were analyzed separately.[27]

1. Cross Examination

[27] The risks variable may be conceptualized as either a 3- or 4-level variable depending on whether or not individualization is included in the set. For simplicity, I present risk as a 3-level variable but note that the MANOVA results are substantively identical when the risks are treated as a 4-level variable.

Contrary to predictions, none of the source and guilt dependent measures in the main experiment were affected by the introduction of *cross-examination.* There was no effect for cross-examination on source confidence, source probability, guilt confidence, guilty probability, or verdict. Likewise, there was no effect for cross-examination across the two individualization conditions on any of the dependent measures. The only exception was the evidence strength judgments in the main experiment (F(2, 206) = 3.83, p=.023). Jurors who read the exchange between the defense attorney and the expert in which various potential sources of error were raised and conceded regarded the shoeprint evidence to be weaker ($M_{\text{no cross}}$ = 5.6, M_{cross} = 5.2, F(1, 207) = 5.14, p = .024) and less convincing ($M_{\text{no cross}}$ = 5.9, M_{cross} = 5.4, F(1, 207) = 7.70, p = .006) than jurors who did not read cross examination testimony.

2. Risks

A MANOVA in the main experiment detected an effect for *risks* on the guilt dependent measures (F(4, 414) = 3.02, p=.018). When the shoeprint expert was most forthcoming on direct exam about the sources of risk (i.e., when the risks of coincidence, mix-up, and examiner error had all been factored into the expert's 1 in 1,000 match statistic), jurors were *less* confident of the defendant's guilt ($M_{\text{complete revelation}}$ = 4.5, $M_{\text{incomplete revelation}}$ = 5.0, t(211) = 2.40, p = .019), assigned lower probabilities of Guilt (P($\text{Guilt}_{\text{complete revelation}}$) = .654, P($\text{Guilt}_{\text{incomplete revelation}}$) = .761, t(211) = 2.69, p = .008), and were less likely to return a guilty verdict ($P_{\text{complete revelation}}$ = .366, $P_{\text{incomplete revelation}}$ =.570, Wald χ^2 (1, n=213) = 7.74, p = .005) than were jurors in the other two risks conditions (coincidence only, and coincidence + mix-up only). The pattern of results was similar on the source confidence ($M_{\text{complete revelation}}$ = 5.0, $M_{\text{incomplete revelation}}$ = 5.3) and source probability (P($\text{Source}_{\text{complete revelation}}$) = .725, P($\text{Source}_{\text{incomplete revelation}}$) = .789) dependent measures, but the multivariate F did not reach statistical significance (F < 2, p > .05). The relevant means are depicted in Figure 3. There was no effect for risks on evidence strength judgments (F < 1, p > .05).

INSERT FIGURE 3 ABOUT HERE

----

3. Results for the Individualization and Control Conditions

In addition to the main experiment, there were two individualization conditions and a control group. The individualization conditions differed from one another only in terms of whether there was a cross-exam. Contrary to predictions, there was no effect for cross exam on any of the key dependent measures in the two individualization conditions. Furthermore, there were no significant differences between the individualization and non-individualization conditions on any of the key dependent measures.

As expected, jurors in the control group (who were not aware of the forensic science evidence) were much less likely than jurors in the other conditions to think the defendant was guilty. Jurors in the control group convicted 6.1% of the time compared with a 51.4% conviction rate for jurors in the other conditions (χ^2 (1, n=315) < 24.42, p < .001).

4. Other Effects

There were main effects for age, ethnicity, and prior jury service on most of the key dependent variables. Older jurors, white jurors, and jurors with prior experience on a criminal jury thought the evidence was stronger (p’s < .05). These effects did not interact with the primary independent variables of interest. There were no significant effects for gender, education level, or political views.

D. Discussion

This experiment tested the effects of (a) direct exam admissions about shortcomings in a case-specific shoeprint match statistic, and (b) cross-examination on jurors' perceptions about the value of the forensic evidence. Regarding direct exam admissions about the match statistic, the descriptive results do not coincide with normative principles. When the expert offered a match statistic without acknowledging the risks that diminish its probative value (i.e., coincidence, mix-ups, and examiner error), jurors were generally *more* persuaded by the evidence than they were when the expert offered objectively stronger evidence (i.e., evidence that *did* account for the various risks). Jurors reported less confidence in the defendant's guilt, estimated lower probabilities of guilt, and returned fewer guilty verdicts when the expert presented a match statistic that expressly accounted for the various risks versus one that did not.

Apparently, then, the mere acknowledgement and/or expression of the various risks that were taken into account by the shoeprint match statistic by the expert signaled a problem with the statistic. Ignoring those risks on direct examination not only didn't hurt the perceived value of the shoeprint evidence, it enhanced it. Even if one makes the generous assumption that jurors were generally persuaded by the shoeprint match statistic because they *assumed* that the statistic accounted for any unmentioned risks, this assumption does not explain why they were *more* persuaded by a match statistic that said nothing about those risks than by an identical match statistic that explicitly incorporated those risks.

Cross-examination, wherein the defense attorney specifically grilled the examiner about the risks that were not included in his match statistic, was only minimally effective. Whereas the presence of cross about shortcomings in the shoeprint match statistic reduced jurors' beliefs about the strength of the forensic evidence, this effect did not carry through to the various source and guilt measures. That is, jurors who read the detailed cross-examination of the shoeprint expert were no less likely to believe the defendant's shoes made the shoeprint or that the defendant was guilty of the burglary than were jurors who did not read any cross-examination testimony.

Surprisingly, this experiment did not produce evidence that people give more weight to individualization forensic science testimony than they give to probabilistic evidence. Perhaps common sense and experience teach people that shoeprints are not unique, even when an expert suggests otherwise. If so, then jurors may have discounted the expert's individualization claims. Alternatively, those who received the match evidence in statistical form may have treated the evidence as a near-individualization, i.e., near certain proof that the shoeprints were made by the defendant's shoes. However, the fact that the source probability ratings were not extreme ($\overline{X}$ = 76.5%) argues against this interpretation. A third explanation is that the relative lack of corroborating evidence in the case gave jurors pause and made them unwilling to give as much weight to the individualization testimony as they otherwise might. Regardless of the explanation, this result hints that a movement away from individualization claims toward quantifiable match statistics may not diminish the weight that jurors assign to forensic science evidence (Mnookin et al., in press).

Another surprising result pertaining to the individualization conditions is that cross-examination that revealed the many assumptions underlying an individualization claim was ineffective. One possibility is that jurors who were not exposed to cross-examination arguments intuited all of them and discounted the evidence accordingly. A more likely explanation is that jurors didn't credit the arguments, particularly as they came from defendant's counsel.

A note of caution. This Internet-based experiment suffers from many of the same limitations as traditional paper-and-pencil studies with mock jurors. There was no voir dire. There were no opening statements, or live witnesses whose demeanors could be observed. There were no objections, rulings, limiting instructions, closing arguments, detailed judicial instructions, or deliberations with fellow jurors. The testimony was read on a computer screen, the trial was compressed, and there were no consequences associated with jurors' judgments or decisions. Therefore, caution is needed when attempting to extrapolate these results to the courtroom.

VII. CONCLUSION

This paper contains two distinct components. The first component is essentially an argument about the appropriate way to assess the probative value of forensic science evidence. Using a LR approach, the paper argues that the value of a match between, say, a blood evidence recovered from a crime scene and a reference sample is best accomplished by characterizing *the evidence* as "a reported match between the samples the analyst tested," while characterizing *the hypothesis* of interest as "the matching person or object is the source of the crime scene sample." These characterizations of the evidence and the hypothesis differ from the traditional approach.

Regarding the evidence characterization, the traditional approach – which describes the evidence as a *true match* between the crime scene and reference samples – assumes too much. It assumes that there is no possibility that either the actual crime scene or reference samples were mixed up or contaminated in some important way. It also assumes that there is no difference between an examiner's conclusions about whether two samples match and whether those two samples, in fact, match. This approach should be abandoned in favor of the approach recommended in Section III. It is no more appropriate to treat an examiner's *report* of a match as a true match than it is to treat an eyewitness's report of an event as a veridical account of what actually occurred.

Regarding the hypothesis characterization, the paper argues that the perspective of the factfinder, rather than that of the forensic examiner or the laboratory director, belongs in the equation. Whereas laboratory personnel may wish to know whether the tested samples match (examiner's hypothesis) or whether the samples from the crime scene and suspect match (laboratory director's hypothesis), the factfinder must entertain a *source hypothesis*, rather than a match hypothesis.[28]

The arguments about how to characterize a LR for forensic science evidence have implications for the admissibility of error rates and the match evidence itself. Under the standard that Federal courts and many state courts apply when evaluating the admissibility of expert testimony (including forensic science testimony), the reasoning or methodology behind the testimony must be scientifically valid (*Daubert* v. *Merrell Dow Pharmaceuticals*, 1993). According to *Daubert*, a trial court's inquiry into scientific validity may include an assessment of the technique's "known or potential error rate" (p. 580). Because the denominator of a LR is a false positive error rate, these values may be offered to a court as an indicator of its error rate. Such error rates are inappropriate if the LR is constructed solely to capture what is often its least likely source of error (coincidental match) while ignoring a greater threat (human error). The false positive error rates that flow from the LR characterization recommended herein will likely

[28] One could take the factfinder's perspective further to suggest that the factfinder not only needs to know who or what is the source of recovered forensic evidence, but also *when* the evidence was created. But this legitimate question reaches beyond the forensic examination per se into non-forensic features of the case such as whether and when a suspect may have had access to the crime scene or murder weapon, or whether the suspect was framed.

be much larger than those trial judges typically hear and may have a larger impact on their rulings.

One empirical question that these normative arguments beg is whether any of these distinctions will make a different to the factfinder. With this in mind, the second component of this paper is an experiment that examines whether mock jurors are sensitive to variations in the assumptions used to produce a match statistic in a hypothetical burglary case involving shoeprint evidence. Some of the results are surprising. Examiners who took multiple sources of error into account and offered the most probative match statistic were *less* persuasive than those who took fewer sources of error into account and offered a less probative statistic. Contrary to predictions, cross examination was largely futile. Even when examiners expressly conceded on cross that their statistic failed to consider one or more sources of error, jurors continued to assign relatively greater weight to an objectively weaker match statistic. Moreover, jurors failed to differentiate between identification testimony (in which the examiner simply identified the shoeprint as a coming from the defendant's shoe) and probabilistic testimony in which the examiner offered a 1 in 1,000 match that accounted for potential error in varying degrees.

Perhaps examiners who are forthcoming about potential sources of error on direct exam are less persuasive than those who concede those points during cross-exam because the early and voluntary concessions signal an overall lack of reliability to the factfinder. Thus, rather than being impressed with the expert's honesty, factfinders who hear reasons why a reported shoeprint match could be erroneous may associate the references to "error" with the technique and discount the evidence. If such an association exists, future research might consider whether it is best explained by an availability heuristic (Tversky & Kahneman, 1974), affective priming (Murphy & Zajonc, 1993), or some other process.

An obvious and unfortunate practical implication of these results is that prosecutors may wish to advise their forensic experts against being entirely forthcoming about all potential sources of error in direct exam testimony. This unethical strategy contrasts with the well-known strategy of conceding shortcomings on direct exam in order to take the "sting" out of anticipated attacks on cross-exam.

Whether the results observed here are generalizable or not, they lend support to the body of research that finds many people are confused about how to evaluate forensic statistics in general and the risk of error in forensic match statistics in particular. Not only do people fail to appreciate that a match statistic that takes account of fewer sources of potential error is *less* probative than a numerically identical match statistic that does account for those sources of error, but they may be immune to subsequent arguments and concessions related to those undisclosed error risks. Koehler, Chia and Lindsey (1995) observed a similarly disconcerting finding wherein many mock jurors were not concerned about a relatively high DNA error rate when it was accompanied by an RMP of 1 in one billion. In a well-designed study that focused on mock jurors' understand of mitochondrial DNA evidence, Kaye et al. (2007) reported (among other things) that 40% of their jurors thought that the DNA evidence was "completely irrelevant" when more than 1 person could match, and 38% of those who returned *guilty* verdicts agreed with the erroneous statement that "the mtDNA evidence shows only a 1 in 57 chance that the defendant committed the crime" (p. 815). Other studies hint that confused jurors may simply ignore scientific and statistical evidence. Kaasa et al. (2007) found that mock jurors who reported having relatively low confidence in their statistical ability gave *no* weight to bullet lead match evidence regardless of its probative value.

The picture that emerges from mock jury studies is that people are confused about how to evaluate forensic match statistics and forensic error rates. Hopefully, this confusion will motivate forensic scientists, psychologists, statisticians, legal scholars and others not only to identify match statistics that take account of key sources of risk, but to find effective ways to convey the true significance of those statistics and risks to a broad range of laypeople.

REFERENCES

Aitken, Colin G. G. (1995) *Statistics and the Evaluation of Evidence for Forensic Scientists,* Chichester: Wiley.

Aitken, Colin G. G., & Franco Taroni (2004) *Statistics and the Evaluation of Evidence for Forensic Scientists,* 2nd ed. West Sussex, England: John Wiley and Sons Ltd.

Allen, Ronald, & Michael Pardo (2007) "The Problematic Value of Mathematical Models of Evidence," 36 *J. of Legal Studies* 107.

Balding, David. J. (2005) *Weight-of-evidence for Forensic DNA Profiles*, West Sussex, England: John Wiley and Sons Ltd.

Black, W. C., & Armstrong, P. (1986). "Communicating the Significance of Radiologic Test-results: The Likelihood Ratio," 147 *American J. of Roentgenology*, 1313.

Bodziak, William. J. (2000) *Footwear Impression Evidence: Detection, Recovery, and Examination*, 2nd ed. Boca Raton: CRC Press.

Bozza, Silvia, Franco Taroni, Raymond Marquis, & Matthieu Schmittbuhl (2008) "Probabilistic Evaluation of Handwriting Evidence: Likelihood Ratio for Authorship," 57 *Applied Statistics* 329.

Champod, Christophe, & Ian W. Evett (2001) "A Probabilistic Approach to Fingerprint Evidence," 51 *J. Forensic. Identification* 101.

Champod, Christophe., Ian W. Evett, & G. Jackson (2004) "Establishing the Most Appropriate Databases for Addressing Source Level Propositions," 44 *Science & Justice* 153.

Cole, Simon A. (2004) "Grandfathering Evidence: Fingerprint Admissibility Rulings from Jennings to Llera Plaza and Back Again," 41 *Am. Crim. L. Rev.* 1189.

Cole, Simon A. (2009) "Forensics Without Uniqueness, Conclusions Without Individualization: The New Epistemology of Forensic Identification," 8 *Law, Probability & Risk,* 233.

Cook, R., Ian W. Evett, G. Jackson, P.J. Jones, & J. A. Lambert (1998a) "A Hierarchy of Propositions: Deciding Which Level to Address in Casework," 38 *Science & Justice* 231.

Cook, R., Ian W. Evett, G. Jackson, P. J. Jones, & J. A. Lambert (1998b) "A Model for Case Assessment and Interpretation," 38 *Science & Justice* 151.

Cook, R., Ian W. Evett, G. Jackson, P. J. Jones, & J. A. Lambert (1999) "Case Pre-assessment and Review of a Two-way Transfer Case," 39 *Science & Justice* 103.

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

Edwards, Harry. T. (2009) "Solving the Problems That Plague the Forensic Science Community," 50 *Jurim. J.* 5.

Edwards, Ward, H. Lindman, Leonard J. Savage (1963). "Bayesian Statistical Inference for Psychological Research," 70 *Psych. Rev.* 193.

Edwards, Ward, & Detlof von Winterfeldt (1986) "Cognitive Illusions and Their Implications for the Law," 59 *S. Cal. L. Rev*. 225.

Evett, Ian W. (1984) "A Quantitative Theory for Interpreting Transfer Evidence in Criminal Cases," 33 *Applied Statistics* 25.

Evett, Ian W., G. Jackson, J. A. Lambert (2000) "More on the Hierarchy of Propositions: Exploring the Distinction Between Explanations and Propositions," 40 *Science & Justice* 3.

Evett, Ian W., G. Jackson, J. A. Lambert, & S. McCrossan (2000) "The Impact of the Principles of Evidence Interpretation on the Structure and Content of Statements," 40 *Science & Justice* 233.

Evett, Ian. W., & Bruce S. Weir (1998) *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*, Sinauer: Sunderland, MA.

Faigman, David L., Michael J. Saks, J. Sanders, Edward K. Cheng (2008) *Modern Scientific Evidence: Forensics*, St. Paul, MN: Thomson West.

Farabee, W.C. (1922) *Indian Tribes of Eastern Peru. Papers of the Peabody Museum of American Archaeology and Ethnology*, Volume X (p. 170), Harvard U. Press: Cambridge, MA.

Federal Rule of Evidence 401

Finkelstein, Michael O., & Bruce Levin (2003) "On the probative value of evidence from a screening search," 43 *Jurim.* 265.

Fischhoff, Baruch, & Ruth Beyth-Marom (1983) "Hypothesis Evaluation From a Bayesian Perspective," 90 *Psych. Rev.* 239.

Foreman, L. A., Christophe Champod, Ian W. Evett, J. A. Lambert, & S. Pope (2003) "Interpreting DNA evidence: A review," 71 *Int'l Statistical Rev.* 473.

Friedman, Richard D., Roger C. Park (2003) "Sometimes What Everybody Thinks They Know is True," 27 *Law & Hum. Behav*. 629.

Garrett, Brandon L., & Peter J. Neufeld (2009) “Invalid Forensic Science Testimony and Wrongful Convictions,” 95 *Virginia Law Rev.* 1.

Gilovich, Thomas, Dale Griffin D, & Daniel Kahneman (2002) *Heuristics and Biases: The Psychology of Intuitive Judgment*, Eds. New York: Cambridge University Press.

Hammer, Lesley (2010) Forensic Footwear Examination Overview. Presentation given at the Workshop on Cognitive Bias and Forensic Science, Northwestern Law School, Chicago, IL.

Hicks, Tacha (2009) Addressing Activity Level When Interpreting Glass or Trace Evidence: What Else? (unpublished manuscript).

Jaeschke, R., G. Guyatt, & D. L. Sackett (1994) “Users’ Guides to the Medical Literature III. How to Use an Article About a Diagnostic Test. A. Are the Results of the Study Valid?” 271 *J. Amer. Med. Assoc.* 389.

Kadane, Jay. B., & David A. Schum (1996) *A Probabilistic Analysis of the Sacco and Vanzetti Evidence*. New York: Wiley.

Kahneman, Daniel, Paul Slovic, & Amos Tversky (1982) *Judgment Under Uncertainty: Heuristics and Biases*, New York: Cambridge University Press.

Kahneman Daniel, & Amos Tversky (2000) *Choices, Values, and Frames*, Eds. New York: Cambridge University Press.

Kaasa, Suzanne O., T. Peterson, E. K. Morris, & William C. Thompson (2007) “Statistical Inference and Forensic Evidence: Evaluating a Bullet Lead Match,” 31 *Law & Hum. Behav.* 433.

Kaye, David H. (1986) “Quantifying probative value,” 66 *Boston U. Law Rev.*, 761.

Kaye, David H. (2010) “Probability, Individualization, and Uniqueness in Forensic Science Evidence: Listening to the Academies,” 75 Brook. L. Rev 1163.

Kaye, David H., Valerie P. Hans, Michael Dann, E. Farley, & S. Albertson (2007) “Statistics in the Jury Box: How Jurors Respond to Mitochondrial DNA Match Probabilities,” 4 *J. of Empirical Legal Studies* 797.

Kaye, David H. & Jonathan J. Koehler (1991) “Can Jurors Understand Probabilistic Evidence?” 154 *J. Royal Statistical Society, Series A* 75.

Kaye, David H. & Jonathan J. Koehler (2003) “The Misquantification of Probative Value” 27 *Law & Human Behav.* 645.

Koehler, Jonathan J. (1993a) “DNA Matches and Statistics: Important Questions, Surprising Answers,” 76 *Judicature* 222.

Koehler, Jonathan J. (1993b) “Error and Exaggeration in the Presentation of DNA Evidence,” 34 *Jurim. J.* 21.

Koehler, Jonathan J. (1996) “On Conveying the Probative Value of DNA Evidence: Frequencies, Likelihood Ratios and Error Rates,” 67 *U. Colorado Law Rev.* 859.

Koehler, Jonathan J. (2001a) “When are People Persuaded By DNA Match Statistics?” 25 *Law & Human Behav.* 493.

Koehler, Jonathan J. (2001b) “The Psychology of Numbers in the Courtroom: How to Make DNA Match Statistics Seem Impressive or Insufficient,” 74 *Southern Calif. Law Rev.* 1275.

Koehler, Jonathan J. (2008) “Fingerprint Error Rates and Proficiency Tests: What They Are and Why They Matter,” 59 *Hastings L. J.* 1077.

Koehler, Jonathan J., Chia, A. & Lindsey, J. S. (1995) “The Random Match Probability (RMP) in DNA Evidence: Irrelevant and Prejudicial?” 35 *Jurim. J.* 201.

Koehler, Jonathan J., & Laura Macchi (2004) “Thinking About Low-probability Events: An Exemplar Cuing Theory,” 15 *Psych. Science* 540.

Koehler, Jonathan J., & Michael J. Saks (2010) “Individualization Claims in Forensic Science: Still Unwarranted” 75 *Brook. L. Rev.* 1187.

Koehler, Jonathan J., & William C. Thompson (2006) “Mock Jurors' Reactions to Selective Presentation of Evidence from Multiple-opportunity Searches,” 30 *Law & Human Behav.* 455.

Leonard, David P., Edward J. Imwinkelried, David H. Kaye, David E. Bernstein, & Jennifer L. Mnookin (2010) *The New Wigmore: A Treatise on Evidence* (Chapter 12: Forensic Science and Identity).

Lieberman, Joel D., Terance D. Miethe, Courtney A. Carrell, Daniel A. Krauss (2008) “Gold Versus Platinum: Do Jurors Recognize the Superiority and Limitation of DNA Evidence Compared to Other Types of Forensic Evidence?” 14 *Psychol. Pub. Pol'y & L.* 27.

Lindsey, Samuel, Ralph Hertwig, & Gerd Gigerenzer (2003) “Communicating Statistical DNA evidence,” 43 *Jurim. J.* 147.

Lloyd, J. J., Talbot, P. R., & Lawson, R. S. (1998) “Quantifying the Value of Diagnostic Tests,” 19 *Nuclear Medicine Communications* 999.

Lyon, T. D., & Jonathan J. Koehler (1996) “The Relevance Ratio: Evaluating the Probative Value of Expert Testimony in Child Sexual Abuse Cases,” 82 *Cornell Law Rev.* 43.

McCormick, C. T. (1999) *Handbook of the Law of Evidence*, 5th ed. St Paul, MN: West Publishing.

McQuiston-Surrett, Dawn, & Michael J. Saks (2008) "Communicating Opinion Evidence in the Forensic Identification Sciences: Accuracy and Impact," 59 *Hastings Law J.* 1159.

McQuiston-Surrett, Dawn & Michael J. Saks (2009) "The Testimony of Forensic Identification Science: What Expert Witnesses Say and What Factfinders Hear," *Law and Human Behav.*, Published online: 4 March 2009.

Meehl, Paul E., & A Rosen (1955) "Antecedent Probability and the Efficiency of Psychometric Signs, Patterns, or Cutting Scores," 52 *Psych. Bull.* 194.

Mnookin, Jennifer L., Simon A. Cole, Itiel E. Dror, Barry A. J. Fisher, Max Houck, Keith Inman, David H. Kaye, Jonathan J. Koehler, Glenn Langenburg, D. Michel Risinger, Norah Rudin, Jay Siegel, and David Stoney (in press) "The need for a research culture in the forensic sciences." *UCLA Law Rev.*

Morgan, John, Mark Stolorow, & Melissa Taylor <mstaylor@nist.gov> (2008, November 8) "Expert Working Group on Human Factors in Latent Print Analysis" [personal email]. (2008, November 8).

Murphy, Shiela T., & Robert B. Zajonc (1993) "Affect, cognition, and awareness: Affective priming with optimal and suboptimal stimulus exposures," 64 *J. of Personality & Social Psych.* 723.

Nance, Dale. A. & Scott B. Morris (2002) "An empirical assessment of presentation formats for trace evidence with a relatively large and quantifiable random match probability," 42 *Jurim. J.* 403.

Nance, Dale A. & Scott B. Morris (2005) "Juror understanding of DNA evidence: An empirical assessment of presentation formats for trace evidence with a relatively small random-match probability," 34 *J. of Legal Studies* 395.

National Research Council (2004) *Committee on Scientific Assessment of Bullet Lead Elemental Composition Comparison, Forensic Analysis: Weighing Bullet Lead Evidence.* Washington D.C.: The National Academies Press.

National Research Council (2009) *Committee on Identifying the Needs of The Forensic Science Community, Strengthening Forensic Science in The United States: A Path Forward.* Washington D.C.: The National Academies Press.

Neumann, Cedric, Christophe Champod, Roberto Puch-Solis, Nicole Egli, Alexandre Anthonioz, Didier Meuwly, & Andie Bromage-Griffiths (2006) "Computation of likelihood ratios in fingerprint identification for configurations of three minutiae," 51 *J. Forensic Sci.* 1255.

Neumann, Cedric, Christophe Champod, Roberto Puch-Solis, Nicole Egli, Alexandre Anthonioz, Andie Bromage-Griffiths (2007) “Computation of Likelihood Ratios in Fingerprint Identification for Configurations of Any Number of Minutiae,” 52 *J. Forensic Sci*. 54.

Neumann, Cedric, Ian W. Evett, James E. Skerrett, & Ismael Mateos-Garcia (2010) “Quantitative Assessment of Evidential Weight for a Fingerprint Comparison I. Generalisation to the Comparison of a Mark With Set of Ten Prints From a Suspect,” *Forensic Sci. Int’l* (Available online 20 October 2010).

Nisbett, Richard. E., & Lee Ross (1980) *Human inference: Strategies and Shortcomings of Social Judgment*, Englewood Cliffs, NJ: Prentice Hall.

Robertson, Bernard, & G. Anthony Vignaux (1995 or 1998) *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*, Chichester, England: John Wiley and Sons Limited.

Saks, Michael J. and Robert F. Kidd (1981) “Human information processing and adjudication: trial by heuristics,” 15 *Law & Soc’y Rev*. 124.

Saks, M. J. & Jonathan J. Koehler (2008) „ The individualization fallacy in forensic science evidence,” 61 *Vanderbilt Law Rev*. 199.

Schklar, Jason & Shari Diamond (1999) “Juror reactions to DNA evidence: Errors and expectancies,” 23 *Law and Human Behavior* 159.

Schum, David A. (1994) *Evidential Foundations of Probabilistic Reasoning*, New York: Wiley.

Scientific Working Group for Shoeprint and Tire Tread Evidence (2009) “Standard terminology for expressing conclusions of forensic footwear and tire impression examinations.” http://www.swgtread.org/images/guidelines/published/10_terminology_expressing_conclusions.pdf

Taroni, Franco & Colin G. G. Aitken (1998) “Probabilistic reasoning in the law, part I: Assessment of probabilistic and explanation of the value of DNA evidence,” 38 *Science & Justice* 165.

Thompson, W. C. (1989) “Are juries competent to evaluate statistical evidence?” 52 *Law and Contemporary Problems* 9.

Thompson, William C. (1995) “Subjective interpretation, laboratory error and the value of DNA evidence: Three case studies,” 96 *Genetica* 153.

Thompson, William C. (1996) “DNA Evidence in the Trial of O.J. Simpson,” 67 *Col. L. Rev.* 845.

Thompson, William C., & Simon A. Cole (2007) “Psychological aspects of forensic identification evidence,” In M. Costanzo, D. Krauss, & K. Pezdek (Eds.), *Expert Testimony for the Courts* (pp. 31-68). Mahwah, NJ: Erlbaum.

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

Wagenaar, Willem A. (1988) “The proper seat: A Bayesian discussion of the position of expert witnesses,” 12 *Law & Human Behav*. 499.

Yates, J. Frank (1990) *Judgment and Decision Making*, New York: John Wiley & Sons.

Figure 1
Forensic Science Hypotheses

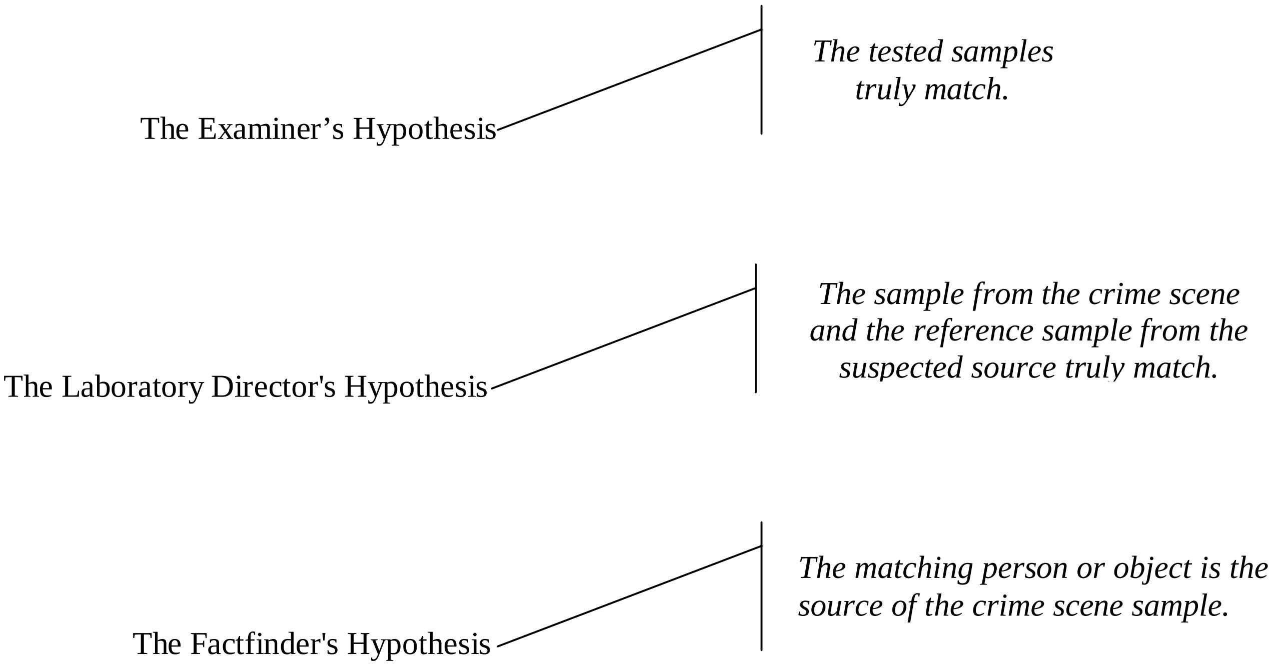
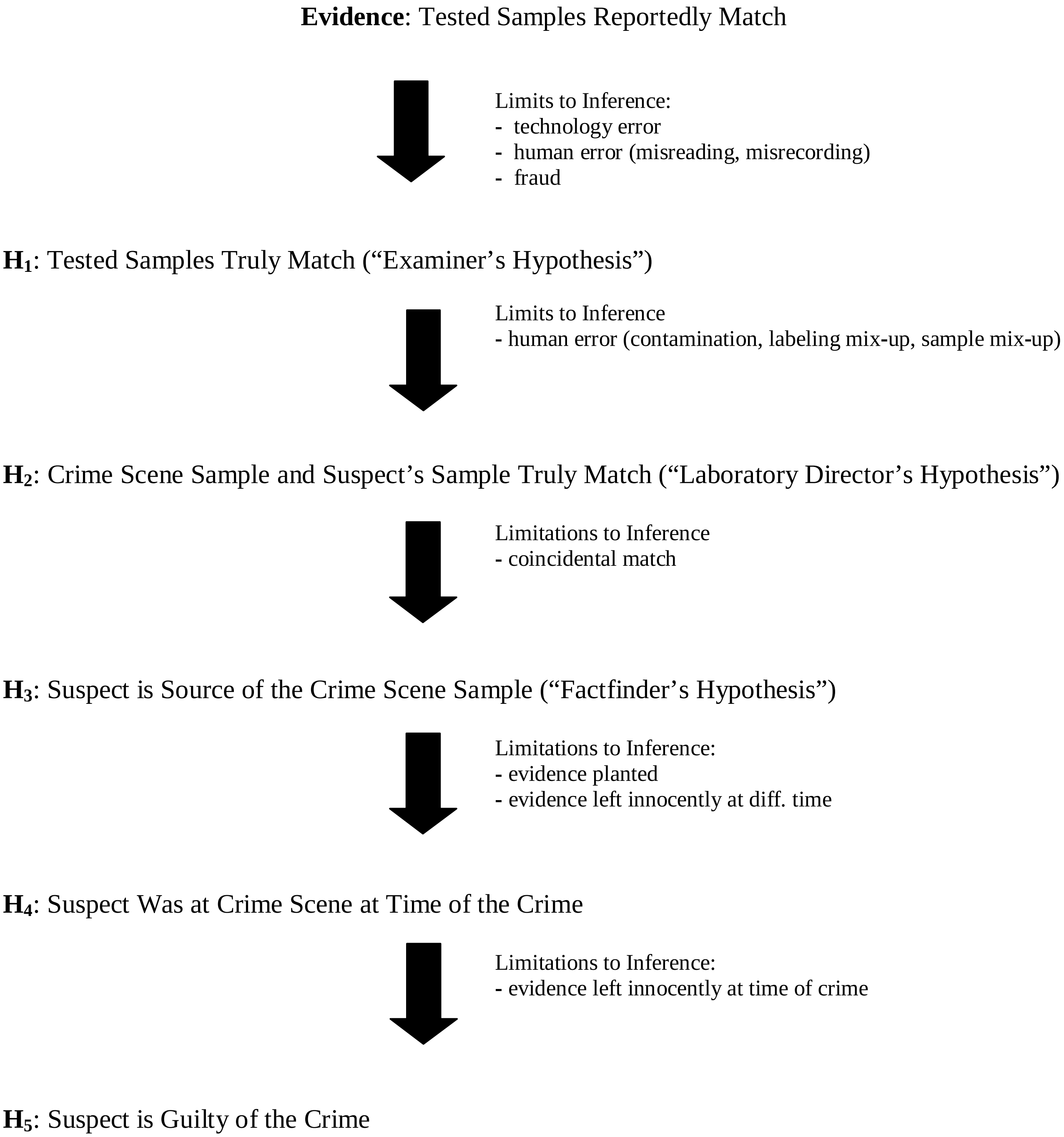

The Examiner's Hypothesis
The tested samples truly match.
The Laboratory Director's Hypothesis
The sample from the crime scene and the reference sample from the suspected source truly match.
The Factfinder's Hypothesis
The matching person or object is the source of the crime scene sample.

Figure 2
The Inferential Chain: From Evidence Through Guilt Hypothesis



Evidence: Tested Samples Reportedly Match

Limits to Inference:
- technology error
- human error (misreading, misrecording)
- fraud

H_1: Tested Samples Truly Match (“Examiner’s Hypothesis”)

Limits to Inference
- human error (contamination, labeling mix-up, sample mix-up)

H_2: Crime Scene Sample and Suspect’s Sample Truly Match (“Laboratory Director’s Hypothesis”)

Limitations to Inference
- coincidental match

H_3: Suspect is Source of the Crime Scene Sample (“Factfinder’s Hypothesis”)

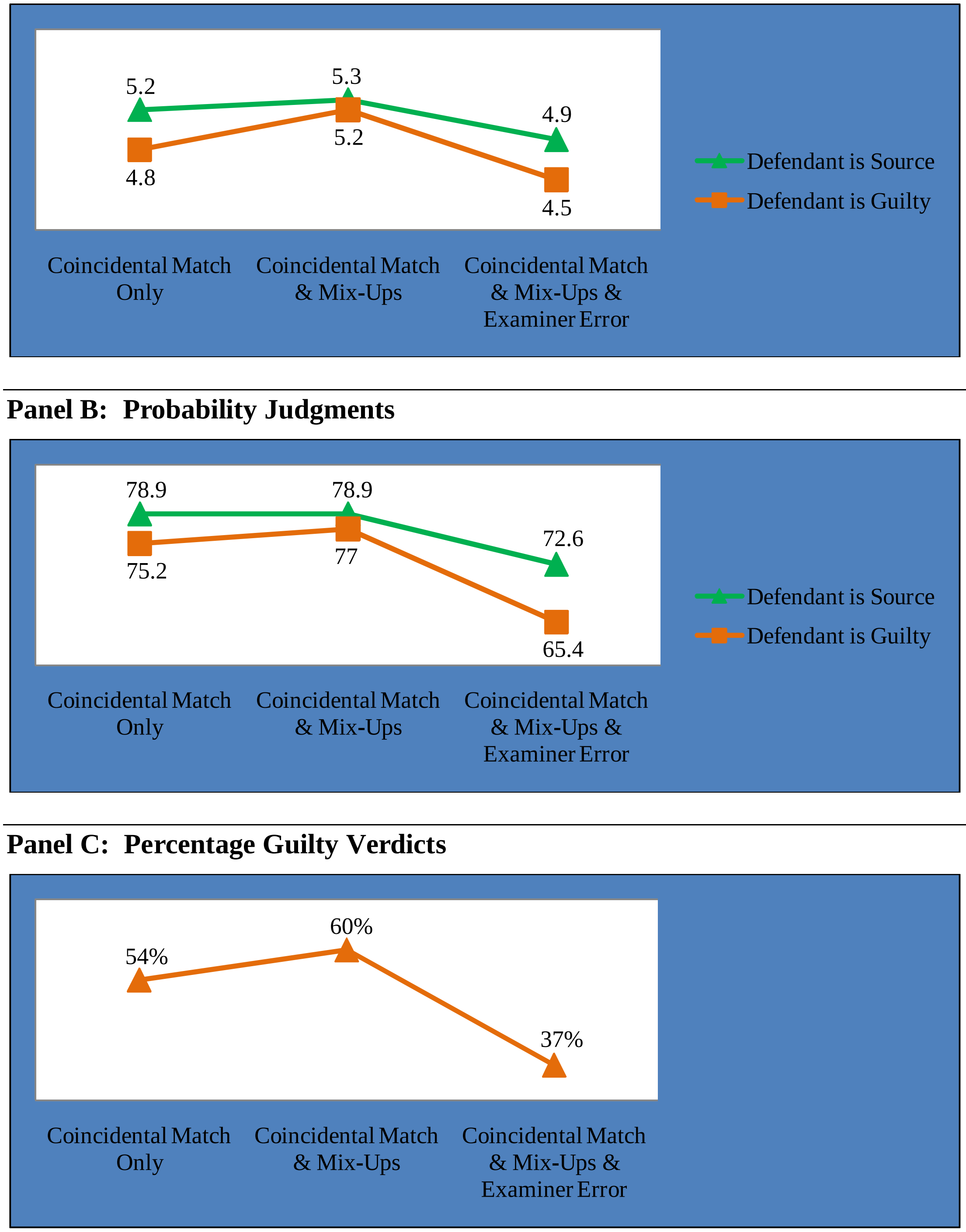
Limitations to Inference:
- evidence planted
- evidence left innocently at diff. time

H_4: Suspect Was at Crime Scene at Time of the Crime

Limitations to Inference:
- evidence left innocently at time of crime

H_5: Suspect is Guilty of the Crime

Figure 3
Confidence Judgments, Probability Judgments and Verdicts as a Function of Risk in the Main Experiment



Panel A: Confidence Judgments (on a 7-point scale)

Panel B: Probability Judgments

Panel C: Percentage Guilty Verdicts

APPENDIX 1: Key Testimonial Differences Across Risk Conditions

Risk: Coincidental Match Only

My tests indicate that the Servis-Cheetah shoes that the defendant was wearing when he was arrested match the shoeprints that came from the Ace Hardware crime scene. In other words, there is a match. If the defendant's Servis-Cheetah shoes did *not* produce the shoeprints left at Ace Hardware, the probability that we would have seen this match by chance is about 1 in 1,000.

… The one in 1,000 statistic is known as a random match probability. In this case, it describes the chance that a randomly selected pair of shoes would match up with the shoeprints found at Ace Hardware as well as the particular pair of shoes the defendant was wearing when he was arrested. In other words, the random match probability gives us an idea of the chance that our finding is purely coincidental.

Risk: Coincidence Match & Mix-ups

My tests indicate that the Servis-Cheetah shoes that we were told the defendant was wearing when he was arrested match the shoeprints that we were told came from the Ace Hardware crime scene. In other words, there is a match. If the defendant's Servis-Cheetah shoes did *not* produce the shoeprints left at Ace hardware, the probability that we would have seen this match either because the match was purely coincidental, or because there was a mix-up wherein the shoes I examined weren't actually the defendant's, or because the shoeprints that I examined weren't actually from the Ace Hardware crime scene, is about 1 in 1,000.

… The 1 in 1,000 statistic is a probability that I derived from considering three ways in which the defendant's shoes might match if, in fact, some other pair of shoes actually left the shoeprints. One possibility is that the match is purely coincidental. That is, maybe some other pair of shoes that just happens to match the defendant's size 12 Servis-Cheetah shoes actually left the shoeprints. A second possibility is that there was a case mix-up where the shoes I was given or the shoes I tested weren't even the defendant's shoes. A third possibility is that the shoeprints were taken from some place other than the Ace Hardware crime scene. Now, none of these are very likely possibilities. But I included them all in my calculations and arrived at 1 in 1,000 in this way. So the 1 in 1,000 statistic is the chance that I would have seen this particular match because the match was purely coincidental, because there was some sort of shoe mix-up, or because there was a shoeprint mix-up.

Risk: Coincidental Match & Mix-ups & Examiner Error

My tests indicate that the Servis-Cheetah shoes that we were told the defendant was wearing when he was arrested match the shoeprints that we were told came from the Ace Hardware crime scene. In other words, we are reporting a match. If the defendant's Servis-Cheetah shoes did *not* produce the shoeprints left at Ace hardware, the probability that we would report this match either because the match was purely coincidental, or because there was a mix-up wherein the shoes that I examined weren't actually the defendant's, or because the shoeprints that I examined weren’t actually from the Ace Hardware crime scene, or because I made some sort of critical error in my analyses, is about 1 in 1,000.

… The 1 in 1,000 statistic is a probability that I derived from considering four ways in which I might report that the defendant’s shoes match if, in fact, some other pair of shoes actually left the shoeprints. One possibility is that the match is purely coincidental. That is, maybe some other pair of shoes that just happens to match the defendant’s size 12 Servis-Cheetah shoes actually left the shoeprints. A second possibility is that there was a case mix-up where the shoes I was given or the shoes I tested weren’t even the defendant’s shoes. A third possibility is that the shoeprints were taken from some place other than the Ace Hardware crime scene. A fourth possibility is that I made a critical error in my analyses. Maybe the shoes I tested and the shoeprints that I observed really do not match. Maybe I misperceived something or made some other crucial error. Now, none of these are very likely possibilities. But I included them all in my calculations and arrived at 1 in 1,000 in this way. So the 1 in 1,000 statistic is the chance that I would have seen this particular match because the match was purely coincidental, or because there was some sort of shoe mix-up, or because there was a shoeprint mix-up, or because I made a crucial error in my analyses.

Individualization

My tests indicate that the Servis-Cheetah shoes that the defendant was wearing when he was arrested were the one and only pair of shoes that produced the shoeprints recovered from the Ace Hardware crime scene. In other words, there is a match. If the defendant’s Servis-Cheetah shoes did *not* produce the shoeprints left at Ace Hardware, it is simply impossible that we would have seen this match.

APPENDIX 2: Key Dependent Measures

Evidence Strength

In comparison with other possible kinds of evidence, how strong would you say the shoeprint evidence is in this case?

How convincing is the testimony from the shoeprint expert?

Source

How confident are you that the shoeprints recovered at the crime scene were made by the defendant's shoes?

What would you say is the probability that the shoeprints recovered at the crime scene were made by the defendant's shoes? (Please provide a number between 0% and 100%)

Guilt

How confident are you that the defendant committed the burglary?

What would you say is the probability that the defendant committed the burglary? (Please provide a number between 0% and 100%)

As a juror, a judge would instruct you to consider all of the evidence and arguments in this case carefully. You are to find against the defendant if and only if the evidence convinces you "beyond a reasonable doubt" that he is guilty of this burglary. You have a reasonable doubt if you cannot say that you are firmly convinced that the charge is true. Reasonable doubt is a doubt based on reason and common sense. It is not a possible or imaginary doubt. Proof beyond a reasonable doubt is the sort of proof that you would be willing to rely and act upon in your most important affairs. What verdict would you return?