Jury Understanding of DNA Evidence:
An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Small Random Match Probability

Dale A. Nance & Scott B. Morris*

[Draft of October 17, 2003]

ABSTRACT: *In cases involving scientific evidence in the form of a test result linking the accused to a crime (e.g., DNA match), expert testimony sometimes can also provide a suitably reliable estimate of the chance of a coincidental match (the random match probability). Considerable controversy, however, attends the question of whether to allow testimony reporting that probability and, if so, in what form it should be given. Additional and related controversy concerns the implications of proficiency test results for testimony about the chance of false positive lab error, especially when that figure greatly exceeds the random match probability. This paper reports a large scale empirical study, using members of an Illinois jury pool, designed to contribute to our understanding of the issues involved. Our results confirm earlier research suggesting that jurors, rather than being credulously overwhelmed by the science, tend to undervalue forensic match evidence. On the other hand, our results differ from most prior research in showing that variation in the way the random match probability is presented and explained can reduce the extent of the undervaluation, without at the same time inviting inferential fallacies that would exaggerate the probative value of the match. And contrary to predictions, our results also show that incorporating information about comparatively large lab error rates, when it has any discernible effect, actually increases the jurors' assessed probability of guilt and willingness to convict.*

Some forensic science techniques, such as DNA analysis, allow expert witnesses no only to testify that a marker associated with one sample "matches" that associated

* Dale A. Nance is Professor of Law, Case Western Reserve University. Scott B. Morris is Associate Professor of Psychology, Illinois Institute of Technology. We are grateful for the generous cooperation of the judges, jury commissioners, bailiffs, and citizens of Kane County, Illinois. We also thank Regina Harris for her help in making arrangements for conducting the study and Char Nance for her work in collecting the data. Financial support was provided by Chicago-Kent College of Law, Illinois Institute of Technology and Case Western Reserve University School of Law.

with another, but also to quantify the chance that such a match could occur by coincidence, even though the accused is innocent of the alleged crime. This "coincidental match probability" or "random match probability" (RMP) is determined by applying a statistical model, derived from the science of the technique involved. For example, in the context of DNA evidence, the random match probability is determined by the genetic structures found in matching samples, their frequencies of occurrence in an appropriate reference population, and calculations based on those frequencies. Importantly, the random match probability does not speak to the possibility that the report of the match is the result of laboratory or collection error, police planting of evidence, or perjury. Such other concerns are important reasons that testimony reporting a match is not as probative of guilt as it would seem to be, based only on the random match probability.[1]

Controversy has long surrounded the question of how, if at all, to present the random match probability to a jury. How should it be presented so as to help the jury to understand the probative force of the match report but at the same time not cause the jury to focus excessively on the match evidence, as compared to other, usually less quantified evidence in the case?[2] And how, in particular, should the random match probability be presented so as to avoid a jury's inferring that it equals the probability that the two samples do not match or the probability that the accused is innocent, thereby potentially giving too much weight to the match evidence?[3]

More recently, controversy has attended the question of the rational dominance of the chance of laboratory error over the chance of coincidental match, in contexts where the former is much larger than the latter. A random match probability for DNA evidence that is on the order of one in a billion is considerably less impressive than it might seem if one takes into account a false positive laboratory error rate of, say, one in a hundred. In such a context, the combined chance of a report of a match due to *either* a coincidental match *or* a lab error, assuming the accused is innocent, is essentially one in a hundred; the chance of coincidental match is swamped by the chance of lab error, and the rational trier of fact would, in this sense, ignore the random match probability. If

---

[1] *See generally* DAVID L. FAIGMAN, ET AL., SCIENCE IN THE LAW: FORENSIC SCIENCE ISSUES § 11–2.6.2 (2002).

[2] *Compare* Michael O. Finkelstein & William B. Fairley, *A Bayesian Approach to Identification Evidence*, 83 HARV. L. REV. 489 (1970), *with* Lawrence H. Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329 (1971).

[3] *See, e.g.*, William C. Thompson & Edward L. Schumann, *Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy*, 11 LAW & HUM. BEHAV. 167 (1987).

jurors do not understand this, once again there is a risk of their greatly overvaluing the match evidence.[4] What kind of presentation will allow jurors to understand this point and otherwise make appropriate use of the testimony reporting a match? And, to the extent that jurors tend to err in this way, is this kind of mistake as important as other factors that would cause the jury to *under*value the match report?[5]

These controversies reflect a more general issue. To what extent should the law of trials attempt to regulate the presentation of expert testimony by excluding information that might, in view of perceived cognitive weaknesses of jurors, cause them to reach erroneous verdicts? And to what extent, alternatively, should the law of trials attempt to regulate the presentation of expert testimony by encouraging parties to present the expertise in a way that best educates the jury? The emphasis of the former is on preventing irrational inferential mistakes; the emphasis of the latter is on making sure that the presumptively rational jury gets the information it needs in a way it can use effectively. In view of the importance of scientific evidence in modern trials, these are issues of pressing concern.[6]

The present study is part of a research project designed to contribute to our stock of knowledge for answering such questions. Over a three year period (from Fall,1999 to Spring, 2002), we conducted experiments presenting respondents a hypothetical case, varying the manner in which a DNA match report was explained. Our sample size is by far the largest to be found in such studies to date, and unlike most prior studies of this kind, our respondents were citizens actually called for jury service rather than students. In an earlier paper (Phase I), we reported the results of experiments in which the hypothetical DNA evidence had a random match probability (1 in 25) much *larger* than the laboratory error rate presented to the respondents (1 in 1000).[7] In that context, the chance of coincidental match swamps the risk of lab error, rather than the other way around, and attention was accordingly directed to the question of how to present and explain the 1 in 25 figure, without doing more than informing the respondents of the

---

[4] *See, e.g.*, Jonathan J. Koehler et al., *The Random Match Probability in DNA Evidence: Irrelevant and Prejudicial?*, 35 JURIMETRICS J. 201, 211–16 (1995).

[5] *See* Jason Schklar & Shari Seidman Diamond, *Juror Reactions to DNA Evidence: Errors and Expectancies*, 23 LAW & HUM. BEHAV. 159 (1999).

[6] *See* Dale A. Nance, Reliability and the Admissibility of Experts, 34 Seton Hall L. Rev. 191 (2003) (discussing the significance of these competing views in the context of interpreting the "reliability" requirement imposed by Fed. R. Evid. 702 in the wake of the decisions in Daubert v. Merrell Dow Pharm., Inc., 509 US. 579 (1993), and Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999)).

[7] *See* Dale A. Nance & Scott B. Morris, *An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Large and Quantifiable Random Match Probability*, 42 JURIMETRICS J. 403 (2002).

quantified chance of lab error. The present paper reports the results of Phase II, in which the hypothesized random match probability (1 in 40,000) was much *smaller* than the communicated lab error rate (again, 1 in 1000). For this context–more like the norm in modern cases–greater attention was paid to the issue of communicating an estimate of the lab error rate.

As in Phase I, we focused on two distinct but related inquiries. First, does variation in the presentation format affect the respondents' assessment of the match evidence, as measured by their willingness to convict and by the magnitude of the probability of guilt they are willing to assign? Second, how do the various presentation formats compare in terms of the extent of divergence of the respondents' assigned guilt probabilities from the guilt probabilities that the respondents *ought* to have assigned, if they were assessing the probabilities rationally. Unlike the first inquiry, the second obviously presumes a normative measure of the probability of guilt, something that is certainly problematic. As in Phase I, and in accordance with prior research by others, we use Bayes' Rule to generate normative measures of the probability of guilt in the case, determining the probability of guilt respondents should give after receiving the DNA match evidence (their "posterior probability") given the probability of guilt they would assign in the absence of the DNA match evidence (their "prior probability").[8]

The results of this research shed considerable light on questions about potential jury error in evaluating testimony reporting a match. It will be seen that the Phase II results corroborate those of Phase I in showing (a) that jurors tend to undervalue match

---

[8] According to Bayes' Rule, if O(G)—the "prior odds"—denotes the odds of guilt (for example, "1 to 1" or "3 to 1") that a person assigns before incorporating a particular piece of evidence, E, and if O(G|E)—the "posterior odds"—denotes the odds of guilt the same person assigns *after* taking E into account, then these two odds should be related according to a tautology derived from the axioms of probability:

$$O(G|E) = O(G) \times L(E)$$

where L(E) is called the "likelihood ratio" for E and is defined as:

$$L(E) = \frac{P(E|G)}{P(E|\sim G)}.$$

In these equations, G means guilty in fact, rather than a verdict of guilty. Similarly, ~G means not guilty in fact. The vertical bar "|" denotes conditioning, so that P(E|G) means the probability of the evidence E being presented on the assumption that the defendant is guilty, and P(E|~G) means the probability that E would be presented on the assumption that the defendant is not guilty. *See* C.G.G. AITKEN, STATISTICS AND THE EVALUATION OF EVIDENCE FOR FORENSIC SCIENTISTS § 2.5.1 (1995). The translation from "probabilities" to "odds" or the reverse is done as follows: For any event A, O(A) = P(A) ÷ [1 – P(A)]; conversely, P(A) = O(A) ÷ [1 + O(A)].

evidence with a quantifiable random match probability, as measured by plausible Bayesian norms, and (b) that the extent of the undervaluation can be reduced significantly by testimony explaining the random match probability and the probative value of the match in light thereof. The Phase II results also show that the communication of lab error rate estimates may be much less important than has been suggested, and that when the communication of such estimates does matter, its impact is the opposite of what has been assumed.

I. EXPERIMENTAL DESIGN

A. Experimental Conditions

A total of 1,520 volunteers from the pool of citizens summoned for jury service by the criminal courts of Kane County, Illinois, were presented a hypothetical rape case, described in writing.[9] Participants were randomly assigned to one of eleven experimental groups, in a 3 x 3 factorial design with two additional comparison groups.[10] Each respondent was asked to estimate the probability that the accused was guilty of the alleged crime and to indicate whether the respondent would convict the accused under the "beyond reasonable doubt" standard applicable in criminal cases.

The first condition ("No Forensics") provides the basis for determining Bayesian measures of the normative probability of guilt. It presents to the respondents all the evidence in the case other than testimony regarding a DNA match. This evidence consisted of modestly probative identification testimony by the alleged victim,

---

[9] Persons released from service because the County had called more than needed were asked to participate in our study before leaving the courthouse. Jurors excused from service after questioning were not eligible to participate. Only a very small proportion of the eligible jurors (certainly less than 5%) chose not to participate. In order to keep the test instruments short, demographic information was not collected for individual respondents. (Previous research has found no significant variations based on demographic categories. *See*, *e.g.*, Schklar & Diamond, *supra* note 5, at 168–69.) The Jury Commission, however, collected the following information for Kane County for those called to service during the period of this study: 48% male, 52% female; 78.9% White Non-Hispanic, 2.6% White Hispanic, 2.4% Black, 0.7% Oriental, 0.2% American Indian, 1.7% Other, and 13.5% Unknown. Communication from Commission dated Aug. 21, 2000 (on file with the author; category names as in original).

[10] Each of the eleven conditions was further subdivided into two groups, for a total of 22 conditions. This last variation was designed to test the effect of response mode, specifically whether it mattered, in reporting the respondent's assessed probability of guilt, whether the respondent was asked to fill in a blank with a probability (expressed as a percentage) or, rather, to mark a point on a supplied probability scale ranging from 0 to 100%. The possibility of an effect on rated probability had been raised by prior research. *See* Baruch Fischhoff & Wändi Bruine De Bruin, *Fifty-Fifty = 50%?*, 12 J. BEHAV. DECISIONMAKING 149 (1999). Because no main effect or interaction involving the response mode was found, the discussion in the text is simplified by collapsing 22 conditions into 11.

corroborating testimony by a police officer who helped the victim to identify a suspect, and a weak alibi offered by the defense.[11] The same evidence was presented in all the remaining conditions, but augmented with evidence concerning a DNA match.[12]

In the second condition ("Mere Match"), the respondents were given testimony that the DNA profile of a sample taken from the accused matched that of a semen sample taken from the victim. The possibilities of coincidental match and lab error were acknowledged in the testimony, but neither risk was quantified. This presentation format is suggested by a rule that would preclude evidence of small random match probabilities or lab error rates out of fear that the jury might infer that one of these small probabilities is the probability that the accused is innocent or that the accused is not the source of the sample taken from the victim.[13]

In the remaining nine conditions, the 3 x 3 design, the chance of a coincidental match was quantified in testimony. One axis of the design consisted of three alternative formats for presenting and explaining this random match probability: (a) some respondents were given the random match probability as a frequency (e.g., 1 in 40,000 chance of a coincidental match with an innocent person)–hereafter called a "Frequency format";[14] (b) some were given as well how that frequency can be stated as a likelihood ratio (e.g., 40,000 times more likely to match if the accused is the source of the crime scene sample than if he is not)–called a "Likelihood Ratio format";[15] and (c) some

---

[11] More precisely, the victim's testimony was that she is "pretty sure" that the accused was the assailant, but she acknowledged on cross-examination that the place where the crime occurred was dark and visibility was impaired. Her in-court identification was corroborated by a police officer's testimony describing the victim's selection of a photograph of the accused at the police station. The accused's alibi was the uncorroborated testimony of his mother that they were together at the time of the assault.

[12] In order to avoid adverse inferences from the absence of expected evidence, respondents in the No Forensics condition were told that physical evidence of the rape that might have been subjected to scientific analysis was accidentally contaminated before analysis could be performed. *See* Nance & Morris, *supra* note 7, at 411 n.39 (discussing options for handling the problem of such adverse inferences in the experimental design).

[13] *Cf.* State v. Bloom, 516 N.W.2d 159, 168-69 (Minn. 1994) (reaffirming the exclusionary rule for RMPs adopted in earlier cases but adopting an exception applicable only to DNA evidence). *See* Faigman, et al., *supra* note 1, 13, § 11–1.2.3[1][a] (discussing arguments regarding the exclusion of the RMP and the rule of non-exclusion followed by most courts).

[14] To be precise, respondents were told that the DNA profile in question can be found "in only 25 ten-thousandths of one percent–.0025%–of the male population. In other words, on average we would expect to see the DNA profile[] . . . in one out of every 40,000 randomly selected men in the population."

[15] For practical purposes, the likelihood ratio is the reciprocal of the proportion of match reports for innocent defendants. *See* Nance & Morris, *supra* note 7, at 405, 421-24. In order to keep the expert within the scope of expertise, however, the testimony given respondents spoke in terms of the hypothesis that the samples come from the same source, rather than the hypothesis of guilt. *See* IAN W. EVETT & BRUCE S. WEIR, INTERPRETING DNA

respondents were given in addition a chart that maps hypothetical prior probabilities to posterior probabilities for the indicated likelihood ratio–called a "Chart format" (see Appendices A and B). While the Frequency format represents the most prevalent conventional practice, each has been used in particular trial contexts, and their advisability has been debated.[16]

The second axis of the 3 x 3 design consisted of three alternative ways of presenting evidence regarding the chance of a laboratory error and its relationship to the random match probability: (a) some respondents were told that there is a chance of lab error, but it was not quantified–called an "unquantified" lab error format; (b) some were given an estimate of the chance of lab error (1 in 1000) based on published proficiency tests for various labs throughout the United States[17]–called a "quantified" lab error format; and (c) some were given the same estimate of lab error risk and also instructed how to combine this risk with the chance of a coincidental match in order to obtain a combined measure of the risk of testimony declaring a match for either of these two reasons when the accused is innocent–called an "aggregated" lab error format.[18] In the aggregated formats, if a likelihood ratio was presented or explained with a chart, the likelihood ratio used was that which combines the RMP with the lab error rate estimate. The appropriateness of presenting estimates of lab error rates derived from proficiency tests and the effect of doing so on the admissibility of the random match probability are matters of some dispute, but conventional practice has not yet seen the use of aggregated formats.[19]

Thus, the 3 x 3 design can be represented as follows, each condition indicated by a cell numbered for future reference:

EVIDENCE: STATISTICAL GENETICS FOR FORENSIC SCIENTISTS 22-23 (1998).

[16] *See* FAIGMAN ET AL., *supra* note 1, 13, §§ 11–1.2.3[1]-[3].

[17] *See* Koehler et al., *supra* note 4, at 206-10 (reviewing proficiency tests and estimating lab error rate at about 1 in 800).

[18] Because a match report due to a coincidentally true match and a match report due to an error in reporting a non-match are mutually exclusive events, the chance of *either* happening is just the sum of the chances of each happening. *See* Nance & Morris, *supra* note 7, at 422 n.80 and accompanying text.

[19] *See* FAIGMAN ET AL., *supra* not 1, 13, §§ 11–1.2.3 [1][c], [d], 11–1.2.5 (presenting arguments and summarizing case law).

RMP format → ---------- Lab error format ↓	Frequency	Likelihood Ratio	Chart
Unquantified	3	6	9
Quantified	4	7	10
Aggregated	5	8	11

The darker the shading of the cell, the more commonly that format is probably encountered in practice, although no precise statistics on such use are available.

In each condition except the first (No Forensics), respondents were also asked to provide three separate estimates of the risks of false positive indications of guilt. Specifically, they were asked to estimate (by filling in the blank in the expression "1 in __") the chances that, were the accused in fact innocent, (a) evidence reporting a match would result from a coincidental matching of the accused's DNA profile with that of the perpetrator, (b) evidence reporting a match would result from a lab error, and (c) evidence reporting a match would result from police misconduct, bribery of a witness, or some other "miscellaneous" cause.[20] Such responses can be compared with the testimonial figures provided to most respondents for the first two risks (i.e., (a) 1 in 40,000, and (b) 1 in 1000) in order to assess the effect of the testimony in altering pre-testimony expectations regarding these risks. Further, these false positive "expectancies" can be used to construct likelihood ratios for the match evidence. With such ratios, the "No Forensics" condition can be mapped to a distribution of normative

[20] As explained in the report on Phase I, we refer to "false positive indications of guilt," as distinguished from false positive reports of a match. In one sense, a match report would be a true positive report if the DNA profiles of the compared samples in fact match but the accused is innocent nonetheless. Because a match report is probative of guilt, however, we refer to (true) matches due to coincidence, police planting of evidence, etc. as false positive indications of guilt. *See* Nance & Morris, *supra* note 7, at 408-09.

posterior probabilities using Bayes' Rule.[21]

B. Hypotheses

In the course of the investigation, the following principal hypotheses were tested:

Hypothesis 1 (Format-Induced Variance): We hypothesized that the variation in presentation format would produce discernible effects on the respondents' assessments of the probability of guilt and their willingness to convict. This would extend the results of Phase I into the context of a relatively small random match probability.[22] We looked for main and interaction effects resulting from the choice of format for random match probability testimony and for lab error rate testimony.

Hypothesis 2 (Undervaluation): We expected that Kane County jurors, like respondents in most prior studies, would undervalue the DNA match evidence, when measured against an appropriate Bayesian norm.[23] This hypothesis is assessed only in terms of the probability of guilt, as there is no Bayesian normative measure of the willingness to convict without introducing a controversial probabilistic interpretation of the "beyond reasonable doubt" standard.

Hypothesis 3 (Education Effects): Each condition in the 3 x 3 design involves providing the respondent with more information about the DNA match and its significance than is true for the Mere Match condition. Further, along each axis of the 3 x 3 design, additional information is provided. Thus, more information is provided in condition 4 than condition 3; more in condition 6 than condition 3, and so on. We hypothesized that the greater the information about the DNA match provided, the closer the rated probability of guilt would come to the Bayesian norm. Willingness to convict rates were expected to track guilt probabilities. Again, this would extend the results in

---

[21] It would be advantageous if a normative posterior probability could be calculated for each respondent, based on that respondent's expectancies. As explained in the report to Phase I of this study, however, this would require eliciting "prior" probability estimates from each respondent (the probability of guilt without regard to the DNA match evidence), and this introduces a substantial risk that eliciting priors might affect the elicited posteriors. Because it unlikely that a court would authorize a procedure that would require jurors explicitly to estimate prior probabilities, the present design avoids this difficulty. *See id.* at 411.

[22] *See id.* at 416-19 (discussing the existence of format-induced variance in the context of a relatively large RMP).

[23] *See id.* at 424-26 (discussing undervaluation in context of a relatively large RMP).

Phase I.[24]

Two additional hypotheses relate to theories that have been suggested with regard to the effect of lab error information on assessments of probability of guilt and verdict preferences. These theories suggest that a failure to communicate an estimate of the lab error rate and its relation to the random match probability might result in overvaluation of evidence with a relatively small random match probability. The present study offered an opportunity to test the following theories:

Hypothesis 4 (Neglect of Lab Error): By withholding quantified estimates of the lab error rate, jurors might be caused to ignore the chance of false positive lab errors, thus implicitly assuming error-free lab tests.[25] (This might be thought especially likely in the presence of quantified random match testimony that might be mistakenly taken to cover both kinds of risk.) If so, one would expect higher assessed probabilities of guilt and higher willingness to convict rates without testimonial estimates of the lab error rate (conditions 3, 6, & 9) than with such estimates (conditions 4, 7, & 10).

Hypothesis 5 (Averaging Fallacy): Failing to inform jurors how to combine a relatively large lab error risk with a relatively small random match risk, when both are quantified in testimony (conditions 4, 7, & 10), might cause them erroneously to average the two risks instead of adding them, thus seriously underestimating the risks of a false indication of guilt.[26] If so, one would expect higher rated probabilities of guilt and willingness to convict rates than if correct combination instructions are given in testimony (conditions 5, 8, & 11).

II. PRINCIPAL RESULTS

A. Probability of Guilt and Verdict Preferences (Hypothesis 1)

Descriptive statistics for each condition are provided in Appendix C and summarized below.

Assessments of the Probability of Guilt. (See Figure 1.) Differences across

---

[24] *See id.* at 437 (noting partial confirmation of this hypothesis in the large RMP context).

[25] *See* Schklar & Diamond, *supra* note 5, at 165 (posing the question).

[26] *See* Koehler et al., *supra* note 4, at 212. The averaging could occur in at least two ways. One might average the risks themselves: $(0.000025 + 0.001)/2 = 0.000513 < 0.001025$. Alternatively, one might average the denominators of the risks when expressed as normalized fractions: $(40{,}000 + 1)^{-1} = 0.000025 << 0.001025$. The latter error is more severe. See Schklar & Diamond, *supra* note 5, at 163 n.5.

conditions in ratings of the probability of guilt were analyzed using Analysis of Variance (ANOVA).[27] An initial analysis was conducted pooling the data for all nine conditions in which the chance of coincidental match was quantified and comparing them to the No Forensics and Mere Match conditions. Not surprisingly, a one-way ANOVA revealed significant differences in the mean probability of guilt across the three groups.[28] The introduction of evidence of a DNA match had significant effects on the rated probability of guilt. Post hoc contrasts confirmed that the rated probabilities of guilt were significantly lower in the No Forensics condition (mean=0.31) than in any of the conditions with DNA match evidence.[29] In addition, the rated probabilities of guilt in the Mere Match condition (mean=0.52) were significantly lower than in the nine conditions with quantified random match evidence (mean=0.66).[30]

Differences among the Frequency, Likelihood Ratio and Chart formats were explored in a second analysis that incorporated the type of testimony regarding lab errors. A factorial analysis of variance was conducted to examine the effect of the format for random match evidence (frequency, likelihood ratio, or chart) and type of lab error testimony (unquantified, quantified, or aggregated) on the probability of guilt (see Table 1). Only the variation in the format of random match evidence produced a significant difference in the rated probability of guilt.[31] Post hoc comparisons indicated a statistically significant difference between the Frequency format (mean=0.62) and the Chart format (mean=0.70).[32] The Likelihood Ratio format generated a mean probability

---

[27] This technique assumes that the outcome variable is normally distributed and has equal variance across groups. Before conducting the analyses, both assumptions were tested. A homogeneity of variance test indicated significant differences in variance across the five types of random match evidence. This was due to less variability in the No Forensics condition relative to the other conditions. Variances did not differ significantly among the other conditions. Because the difference was not large, and because the F-test is robust to violations of the homogeneity assumption, the lack of homogeneous variance was not a concern. *See* WILLIAM L. HAYS, STATISTICS 407 (5th ed. 1994). The distribution of the probability of guilt was not normal: there was a negative skew and a disproportionate number of individuals (24%) who responded 0.5 (50%). Fortunately, when sample size is large, as in the current study, the F-test is robust to violations of the normality assumption. *Id.* at 406. In addition, several normalizing transformations were applied to reduce the degree of skew. The results were essentially the same under all transformations. Therefore, only the results in the original scale are presented.

[28] $F_{(2, 1471)}=102.30$, $p<0.001$. Given the non-normality of the outcome variable, the test was repeated using the non-parametric Median test, which gave the same results, $_{(2, 1474)}=132.62$, $p<0.001$.

[29] Tukey HSD, $p<0.001$.

[30] Tukey HSD, $p<0.001$.

[31] $F_{(2, 1196)}=7.97$, $p<0.001$. The non-parametric Median test also indicated significant differences among the three formats for random match evidence, $^{2}_{(2, 1205)}=17.63$, $p<0.001$.

[32] Tukey HSD, $p<0.001$.

of guilt falling midway between those of the other two (0.04 above the Frequency format, 0.04 below the Chart format).[33]

These results tend to confirm Hypothesis 1 in that differences in ratings of the guilt probability are generated by variation in the format for evidence about the possibility of a coincidental match. This is consistent with the findings in Phase I for relatively large random match probabilities. The data do not, however, confirm main effects on guilt probabilities based on the variation in lab error rate format (a matter not studied in Phase I). Nor were interactions between RMP format and lab error format detected at levels statistically significant by conventional standards.[34] (But see the discussion of verdict preferences, below.)

Verdict Preferences. (See Figure 2.) An initial analysis of verdict preferences was conducted by pooling all nine conditions with quantified RMP testimony and comparing them to the No Forensics and Mere Match conditions. Again not surprisingly, a likelihood ratio chi-square test revealed a significant difference in the proportion of guilty verdicts.[35] In the No Forensics condition, only 4.5% of the participants chose a guilty verdict, which was significantly lower than the rate of guilty verdicts in the Mere Match condition (24.8% guilty).[36] The formats that provided random match rates also produced significantly more guilty verdicts (45.8%) than the Mere Match condition.[37]

Differences among the three quantified random match formats were examined in combination with the differences in lab error formats using a multinomial loglinear model (see Table 2). The likelihood ratio chi-square test indicated a significant main effect for random match format, but not for lab error format.[38] Furthermore, there was a significant interaction between random match format and lab error format.[39] Post hoc tests showed that the choice of random match format was significantly related to the frequency of guilty verdicts when the expert aggregated lab error and random match

---

[33] Tukey HSD: $p=0.15$ and $p=0.08$, respectively.

[34] $F_{(4, 1196)}=1.63$, $p=0.17$.

[35] $G^2_{(2,1483)}=123.20$, $p<0.001$.

[36] $G^2_{(1, 267)}=24.05$, $p<0.001$.

[37] $G^2_{(1, 1349)}=22.69$, $p<0.001$.

[38] Respectively: $G^2_{(2, 1216)}=23.07$, $p<0.001$; $G^2_{(2, 1216)}=0.64$, $p=0.73$.

[39] $G^2_{(4, 1216)}=11.09$, $p=0.03$.

rates, but not when there was no testimony quantifying the rate of lab error, nor when the lab error rate was quantified but not aggregated with the random match rate.[40]

In order to determine whether the main effect for random match format was due entirely to its interaction with the aggregated lab error format, a test was run on the data by omitting aggregated formats. The effect of the type of random match evidence was similar when there was unquantified lab error testimony and when the lab error rate was quantified but not aggregated, as indicated by a non-significant three-way interaction between type of random match format, type of lab error format, and verdict preference.[41] Therefore, unquantified and quantified conditions were combined and tested for the effect of variation in random match format. This test still revealed statistically significant variation in verdict preferences.[42] Moreover, that random match format affects verdict preferences without the indicated interaction is supported by the main effect found with regard to rated probabilities of guilt.

Interestingly, the aggregated lab error format produced a different pattern of results than the other two lab error formats. In the aggregated formats, guilty verdicts were more likely in the Chart format (64% guilty) than either the Frequency format (35% guilty) or the Likelihood Ratio format (41% guilty), whereas the Frequency and Likelihood Ratio formats did not differ significantly.[43] In contrast, when there was unquantified lab error testimony or the lab error and random match rates were only presented separately, there were fewer guilty verdicts with the Frequency format (38.6% guilty) when compared to the Likelihood Ratio format (48.5% guilty) or the Chart format (48.9% guilty), whereas the Likelihood Ratio and Chart formats did not differ significantly.[44]

As with the probability of guilt responses, these results for verdict preferences tend to confirm Hypothesis 1: variation in the format for presenting the random match probability and the lab error rate have significant effects on verdict preferences. There does appear to be a main effect for variations in the random match format, but not for variations in lab error presentation formats. However, significant interactions between

---

[40] Respectively: $G^2_{(2, 408)}=26.62$, $p<0.001$; $G^2_{(2, 402)}=3.63$, $p=0.16$; $G^2_{(2, 406)}=3.91$, $p=0.14$.

[41] $G^2_{(2, 808)}=0.03$, $p=0.99$.

[42] $G^2_{(2, 808)}=7.47$, $p=0.02$.

[43] Respectively: $G^2_{(1, 270)}=24.07$, $p<0.001$; $G^2_{(1, 273)}=14.80$, $p<0.001$; $G^2_{(1, 273)}=1.22$, $p=0.27$.

[44] Respectively: $G^2_{(1, 540)}=5.40$, $p=0.02$; $G^2_{(1, 540)}=5.81$, $p=0.02$; $G^2_{(1, 536)}=0.01$, $p=0.93$.

random match format and lab error format were also found.

Practical Significance. In some respects, the identified effects might seem small. For example, the analysis of variance for the 3 x 3 design indicated that most of the variance in rated probability of guilt occurs within conditions; variation among conditions accounted for only 1% of the total variance. On the other hand, the difference in willingness to convict between, say, the conventional Frequency with Unquantified Lab Error format (rate = 38%) and the unconventional Chart with Aggregated Lab Error format (rate = 64%) is a difference that any prosecutor or defense attorney will appreciate. Moreover, this difference understates the practical significance of the effects, because of the difference between willingness to convict rates and conviction rates. The latter will depend, of course, on the effects of jury deliberation, which was not simulated in the present experiment. We know, however, that jury verdicts tend to coincide with the verdict preference of the majority of the jury at the beginning of deliberation.[45] This fact can be expected to amplify all differences in willingness to convict rates.

To illustrate, we can model the probability of conviction as the probability that a majority of the members of the jury (7 or more in a jury of 12), when presented with evidence in a particular format, will have a pre-deliberation willingness to convict. According to this model, if we assume, plausibly, that a jury is randomly drawn from a population with a given mean pre-deliberation verdict preference, then the probability of conviction (i.e., 7 or more with a pre-deliberation preference for a guilty verdict) can be obtained using the binomial distribution. For the verdict preferences obtained in this study with respect to the particular case presented, the probability of conviction under the Frequency with Unquantified Lab Error format is 0.13, while the probability of conviction under the Chart with Aggregated Lab Error format is 0.77.

B. Expectancies and Bayesian Norms (Hypotheses 2 and 3)

Expectancies. Each respondent that received evidence of a DNA match was asked to estimate the chances that, were the accused innocent, a match might be reported in evidence. This question was asked with respect to the three categories of causes for such "false positive" indications of guilt. Expectancies can be described as "untutored" or "tutored." Untutored expectancies are those elicited from respondents

[45] *See* REID HASTIE, ET AL., INSIDE THE JURY 63-76 (1983).

without quantitative testimony regarding the indicated risk. Thus, in the Mere Match condition, all expectancies were untutored; in the unquantified lab error conditions, the lab error expectancies were untutored; and in all conditions, the expectancies from miscellaneous causes (bribery of a witness, police planting of evidence, etc.) were untutored. Summary descriptive statistics are presented in Table 3.

Expectancies of coincidental match and for lab error were reduced by testimony quantifying these risks at levels lower than the untutored expectancies, but the effects were small and statistically insignificant for both the probability of a random match and the probability of lab error.[46] The mean tutored expectancies of both random match and lab error remained substantially higher than the corresponding rates reported by the expert witness. On the other hand, the median tutored expectancy of random match (0.0001) was fairly close to the value reported in the testimony (0.000025), and the median tutored expectancy of lab error was the same as the value provided by the expert witness (0.001).

Additional analyses were performed to determine whether tutored expectancies were influenced by the type of testimony. The expectancies of coincidental match did not differ significantly across the Frequency, Likelihood Ratio and Chart methods.[47] Expectancies of lab errors, on the other hand, were significantly higher when the random match and lab error probabilities were presented only separately (mean=0.021, c.i.=±0.006, median=0.001) than when the probabilities were also aggregated (mean=0.011, c.i.=±0.004, median=0.001).[48] Perhaps the failure to provide an aggregate risk estimate caused some respondents to exaggerate the component risks.[49] In any event, given the skew of the distributions, medians are perhaps the better measure of central tendency, and this figure was unaffected by aggregation.

Bayesian Norms. The figures in Table 3, together with the testimonial figures for risks of false positive indications of guilt, can be used to construct plausible normative measures of the probability of guilt in the hypothetical case. The first measure assumes that the unimpeached testimonial figures for random match probability and for lab error

---

[46] Respectively: $t_{(1327)}=-1.09$, p=0.27; $t_{(1340)}=-0.85$, p=0.39.

[47] $F_{(2, 1192)}=1.37$, p=0.26.

[48] $F_{(1, 808)}=8.10$, p=0.005.

[49] Notice that the mean tutored expectancy in the absence of aggregation testimony, 0.021, is *higher* than the mean untutored expectancy, 0.019, even though the testimonial estimate of the risk of lab error, 0.001, is *lower* than the mean untutored expectancy.

rates ought to be accepted by a juror. Because there is no testimonial quantification of the miscellaneous risks, we accept the mean untutored expectancy thereof as normative. This yields a total expectancy of false positive guilt indications: 0.000025 + 0.001 + 0.024 = 0.025.[50] Using this figure and Bayes' Rule to transform the distribution for the No Forensics condition into a normative distribution yields a mean normative posterior probability of 0.84 (median=0.94).

Alternatively, we can assume that jurors' expectations regarding the reliability of prosecution experts may properly cause them to assess the false positive risks as higher than the testimonial figures. Thus, one may use the higher tutored expectancies for coincidental match and lab error together with the tutored figure for miscellaneous risks to obtain a more conservative Bayesian norm.[51] This produces a total false positive expectancy of 0.013 + 0.016 + 0.024 = 0.053. Making the indicated transformation with the help of Bayes' Rule, we obtain a second, more conservative mean normative posterior probability of 0.77 (median=0.89).[52]

None of the presentation formats tested produced a mean (or median) probability of guilt in excess of either measure of the Bayesian norm, thus confirming Hypothesis 2. The highest probabilities of guilt were reported in condition 11, the Chart with Aggregated Lab Error format. This format produced a mean rated probability of guilt of 0.75 (median=0.83), just shy of the second, lower Bayesian norm and not differing to a statistically significant degree from the corresponding normative distribution.[53] As general trends, ratings of the probability of guilt tended to increase toward the Bayesian norms as more information was included in the testimony regarding the RMP, whereas the education effect of variation in lab error format is pronounced only when accompanied by a Chart format (see Figure 1). The proportion preferring a guilty verdict

---

[50] To be precise, this is an estimate of the probability that a match would be reported in testimony assuming the accused were innocent. This is the denominator of the relevant likelihood ratio for employing Bayes' Rule to update the odds of guilt by incorporating the DNA match evidence. *See supra* note 8. In this context, the numerator of the likelihood ratio can safely be assumed to be one. *See* Nance & Morris, *supra* note 7, at 421-24.

[51] This point was first clearly articulated with respect to calculating Bayesian norms in Schklar & Diamond, *supra* note 5, at 180-81.

[52] Little difference results from using the anomalously high 0.021 mean for lab error expectancy obtained in the absence of aggregation instructions. The resulting normative distribution has a mean of 0.76 and a median of 0.88.

[53] $t_{(266)}=0.78$, p=0.43. This mean rating of the guilt probability is also 0.01 less than the Bayesian norm computed using the anomalously high 0.021 mean for lab error expectancy obtained in the absence of aggregation instructions. See *supra* note 52 Again, the two distributions do not differ significantly. $t_{(265)}=-0.49$, p=0.63.

showed similar patterns (see Figure 2). Hypothesis 3 is partially confirmed.

Because all the probability of guilt and expectancy distributions were skewed by extreme values, median figures might be better indications of central tendency. Consequently, Bayesian norm calculations were replicated using the (smaller) median expectancies to transform the prior probability distribution into a posterior probability distribution. The same pattern of results was found: the highest median probability of guilt, 0.83 for the Chart with Aggregated Lab Error format, still did not exceed the median-based Bayesian norms (both having mean=0.91 and median>0.99).

III. DISCUSSION

The present results replicate and extend those of Phase I. The undervaluation of DNA match testimony, measured relative to Bayesian norms, is as true when the testimonial lab error rate is 40 times larger than the testimonial RMP (Phase II) as when the testimonial RMP is 40 time larger than the testimonial lab error rate (Phase I). Yet important differences in probability of guilt assessments and conviction rates are associated with the choice among alternative formats for communicating the random match probability. In particular, normative use of coincidental match statistics is facilitated by presenting a likelihood ratio along with a chart mapping prior probabilities to posterior probabilities. This suggests that proponents of match evidence should be willing to consider, and that courts should be willing to entertain, the use of likelihood ratios and appropriate charts, not confining witnesses to the more conventional frequencies formats.[54]

The implications with regard to lab error rate testimony are more complex. On the one hand, inclusion of separate lab error estimates does not appear to significantly affect decisions, as compared to providing no quantitative estimate at all. Perhaps this is because the median untutored expectancy for lab error (0.001) is the same as the testimonial figure used in this study. In other words, perhaps jurors have a pretty decent idea of the risk of lab error going into the trial and are able to incorporate that understanding in assessing the evidence. However, educational effects are conspicuous when the witness tells the jurors something many may not already know, namely how to

[54] This tends to support a ruling like Plemel v. Walter, 735 P.2d 1209, 1219-20 (Or. 1987) (holding in a civil paternity case that, upon request, the expert's testimony must present a chart to explain the effect of the likelihood ratio presented in testimony).

combine the lab error rate with the chance of coincidental match (RMP), at least when the import of the aggregated risk figure is illustrated with an appropriate device, as in our Chart format.

In one important respect, this is a striking result. When the Chart format was employed without aggregating the lab error rate with the RMP (whether or not an estimate of the lab error was quantified), respondents were shown a chart reflecting the import of a likelihood ratio of $(1/40{,}000)^{-1} = 40{,}000$. (See Appendix A.) When, however, the Chart format was employed with an aggregated lab error rate, the respondents were shown a chart reflecting the import of a much smaller likelihood ratio, $(1/40{,}000 + 1/1000)^{-1} = 976$. (See Appendix B.) The former, of course, yields a higher posterior probability than the latter at each level of prior probability. One might expect, therefore, that the Chart format would have produced higher rated probabilities of guilt (and willingness to convict) without aggregation than with. This, indeed, is a major reason that one critic has objected to the failure to include lab error estimates that are aggregated with RMPs.[55] But just the opposite was found: with a Chart format, aggregation testimony produced higher probabilities of guilt and a higher proportion of respondents willing to convict.[56]

It may well be that many respondents discounted the probability of guilt in the non-aggregated formats because they *appreciated*–explicitly or implicitly–that they did not understand well how to integrate the RMP information with the lab error rate information. This would be a *rational* form of discounting, the result of "error" only in the sense that it can be seen by someone with better understanding of the mathematics as giving too little probative value to the match report. In other words, it is not that the respondents erroneously thought they knew how to integrate the information and therefore did it badly; it is that they rationally discounted in favor of the accused on account of their own ignorance or uncertainty. That, presumably, is what we would want them to do. When respondents are given appropriate instruction, however, the extent of the discount was reduced (and the willingness to convict increased) even though the

---

[55] *See* Jonathan J. Koehler, *Why DNA Likelihood Ratios Should Account for Error (Even When A National Research Council Report Says They Should Not)*, 37 JURIMETRICS J. 425, 426 (1997) (arguing that failure to aggregate might cause jurors to give too much weight to the match evidence, but not specifically addressing Chart formats).

[56] The differences are statistically significant. For rated probabilities of guilt, $t_{(396)}=2.34$, $p=0.02$; for willingness to convict rates, $G^2_{(1, 403)}=8.86$, $p=0.003$. The same relationship appears to hold if one compares the Chart with Aggregated Lab Error format (condition 11) with the Chart with Quantified Lab Error format (condition 10): for rated probabilities of guilt, $t_{(262)}=1.81$, $p=0.07$; for guilt preferences, $G^2_{(1, 268)}=5.44$, $p=0.02$.

likelihood ratio used in the chart decreased.

There is little value in the Mere Match format. Not only is the juror deprived of important information, pertinent to the assessment of the probative value of the match, but this loss of process-rationality is not compensated by any improvement in accuracy; quite the opposite. To the extent that the use of a Mere Match format can be justified, it must be through an overriding concern to minimize the impact of DNA evidence on verdicts. It is hard to see how that can be justified unless the primary objective is simply to minimize the number of convictions. In Phase I of our study, we found that, with a large random match probability (1 in 25), the Mere Match format produced rated probabilities of guilt not substantially lower than those for formats with quantified testimonial estimates.[57] The results of Phase II now clarify that the Mere Match condition forces jurors to impute some background-informed, albeit conservatively adjusted assumptions about the probative significance of a DNA match, assumptions that cause increasing loss of accuracy as the RMP gets smaller and the rational probative value of the match gets higher.

Two questions are addressed in the following sections. First, what explains the undervaluation that occurs with most, if not all, formats tested? Second, is the seeming "Bayesian improvement" associated with changing, say, from a Frequencies with Unquantified Lab Error Format to a Chart with Aggregated Lab Error Format purchased at the cost of inducing jurors to commit identifiable inferential errors supporting the proponent of the DNA match evidence, here the prosecution? Would this amount to "tricking" the jurors to reach more accurate verdicts? In a third section, we address limitations of the study design that might affect the external validity (real world applicability) of the results.

A. Causes of Undervaluation

Scholars have identified several fallacies or inferential errors that favor the opponent of quantified forensic match evidence, here (as usual) the accused. The data allow us to say a few things about these potential causes of undervaluation.

---

[57] *See* Nance & Morris, supra note 7, at 416-418 , 442-444 (discussing the Mere Match format in the context of an RMP of 1 in 25).

Defense Attorney's Fallacy. This fallacy arises when a juror infers that the match testimony is *irrelevant* because there are too many people in the population who share the DNA profile of the accused and the perpetrator. Thus, with a 1 in 40,000 RMP, a suspect population of, say, 4,000,000 might cause a juror to reason that about 100 other people could be the perpetrator and *therefore* the match evidence is irrelevant. The mistake, of course, is in thinking that a single item of evidence is irrelevant unless it, by itself, narrows the suspect population down to one person, or at least a very small number of persons.[58] The incidence of this fallacy could not be measured without being able to identify respondents whose probability of guilt was not affected by receiving the DNA match evidence.[59] In the present context, however, one can infer something from the fact that the hypothetical was set in a "small Illinois town." If respondents assumed that the perpetrator is likely to be local, it is unlikely that the undervaluation of the DNA match evidence was significantly attributable to the Defense Attorney's Fallacy.[60]

Inversion Fallacy. Another mistake that favors the opponent of the match evidence, here the defense, is to equate the RMP with the probability of guilt. This mistake causes the rated probability of guilt to be inversely related to the rational probative value of the match because the smaller the RMP the greater the rational probative value, but the smaller the probability of guilt attributed by someone falling into the trap. In Phase I, this error was found to be surprisingly prevalent in the context of a relatively high RMP of 1 in 25.[61] In the present context, however, only one response among the 1,205 responses in the nine conditions with quantified random match testimony can be clearly identified as probably reflecting this mistake–by giving a probability of guilt of 0.000025. To be sure, this might understate the incidence of this mistake, because someone making it might simply report the probability of guilt as

---

[58] *See* Thompson & Schumann, *supra* note 3, at 171. A more extreme form of this fallacy occurs when the evidence is considered relevant but in fact exculpatory because, continuing the example in the text, an inference is made that the probability is only 1 in 100 that the defendant is guilty. This is fallacious because it ignores all other evidence in the case. *See* Nance & Morris, *supra* note 7, at 427 n.99.

[59] *See supra* note 21.

[60] The same hypothetical case was used in Phase I, where we found an overall incidence of the Defense Attorney's Fallacy of about 7%, but that was in the context of a random match probability of 1 in 25. *See* Nance & Morris, *supra* note 7, at 433-34. A small town of, say, 2,000 adult male residents would be expected to have about 80 such residents with matching DNA profiles.

[61] *See* Nance & Morris, *supra* note 7, at 429 (reporting an overall incidence rate of 5%).

0%.[62] In fact, 35 respondents (3% of the responses) gave probability of guilt ratings of 0%. This figure, however, likely overstates the incidence of this error, because more than 7% of respondents in the Mere Match condition (who did not receive a quantified RMP from which this error could occur) also rated the probability of guilt as 0%. Although a significant number of respondents gave very little weight to the evidence of a DNA match, it cannot be said with confidence that this was due to the inversion fallacy.

Misaggregation. The cumulative import of several studies, including this one, is that people have difficulty giving statistical evidence as much weight as it deserves, a problem called "misaggregation."[63] The significant improvements, relative to Bayesian norms, derived from employing Likelihood Ratio and Chart formats, as well as the even more pronounced improvement from employing the Chart with Aggregated Lab Error Rate format, suggest that misaggregation is a principal source of the undervaluation in the respondents' assessments in this context, and that education effects offset to varying degrees the tendency to undervalue the match report.

B. Cautionary Considerations (including Hypotheses 4 and 5)

Scholars have also identified a number of fallacies of inference that tend to favor the proponent of quantified random match evidence, here the prosecution, at least when the RMP is small. If the apparent superiority of one format relative to another is accompanied by significant incidence of an identifiable fallacy in the ostensibly superior format, then we should be cautious in recommending or allowing its use. And if the apparent superiority of one format relative to another is the result of significantly *increased* incidence of identifiable fallacy in the ostensibly superior format, then our reticence should be even greater. The following paragraphs consider identifiable fallacies that relate to the present study.

Prosecutor's Fallacy. With regard to pro-prosecution fallacies, perhaps the most dangerous is the aptly named "Prosecutor's Fallacy," in which the juror interprets the

---

[62] *Cf.* Jonathan J. Koehler, *The Psychology of Numbers in the Courtroom: How to Make DNA Match Statistics Seem Impressive or Insufficient*, 74 S. CAL. L. REV.1275, 1296 (2001) (suggesting that, in the context of an RMP of 1 in a billion, the incidence of this mistake can be estimated by the percentage of respondents giving any probability of guilt less than 1%, or even less than 10%).

[63] *See* Schklar & Diamond, *supra* note 5, at 163 (distinguishing "misperception error" from "misaggregation error").

random match probability as the probability of innocence (or as the probability that the accused is not the source of the crime scene DNA sample).[64] That an error is involved can be seen from the fact that the probability of innocence must take into account all the other evidence in the case, whereas the random match probability speaks only to the probative value of the evidence reporting the match. Similarly, the probability that the accused is not the source of the crime scene DNA sample depends not only on the random match probability but also things like the chance of lab error and the chance that the prosecution's expert might lie in reporting a match.

In the present context, the only reasonably clear indication that a respondent has made the indicated mistake would be a response that gave the probability of guilt as 1 – RMP = 99.9975%. In fact, only two such responses were found among the 1,205 responses in the nine conditions with quantified random match testimony (one in a Frequency format, one in a Likelihood Ratio format), or about two tenths of one percent of all quantitative responses.[65] This incidence rate is not practically significant.[66]

A similar phenomenon (not previously identified as a form of the Prosecutor's Fallacy) would involve the juror treating the quantified lab error rate as the same as the probability of innocence, or the same as the probability that the samples came from the same source. The inferential error is analogous. With a testimonial lab error rate of 0.001, one would expect such error to appear as probability of guilt responses of 99.9%. Only 7 such responses appear in the data (two in a Frequencies format, two in a Likelihood Ratio format, and three in a Chart format). These responses do not

---

[64] *See* Thompson & Schumann, *supra* note 3, at 170-71.

[65] Scaled responses (*see supra* note 10) did not allow respondents to communicate with the necessary precision, and both "99.9975%" responses were given in the fill-in-the-blank response mode. There were 613 fill-in-the-blank responses in conditions with testimony quantifying the random match probability. If we ignore the scaled responses for this purpose, the incidence rate is still only 0.3%.

[66] To be sure, the figure in the text might understate the incidence of this fallacy. Respondents might have made the fallacious inference but then reported their probability of guilt by rounding to something like 99.9%. Alternatively, respondents might have fallaciously inferred that the probability that the samples came from the same source was .999975 but then (perhaps appropriately) adjusted this figure to account for the chance that, if the samples did come from the same source, it was only because the police somehow planted the evidence that would yield this result. It is not possible to discern whether such errors occurred, because such responses might well *not* reflect this fallacy. Moreover, the figure in the text might also overstate the incidence, because even responses of 99.9975% cannot be considered fallacious if the respondent's probability of guilt without regard to the match evidence was about 50% (see Appendix A, entry for 50% prior probability); in such a case the probability of guilt is plausibly consistent with Bayesian inference. *See* Nance & Morris, *supra* note 7, at 428, n.101. For good reason, the respondents' "prior" probability of guilt was not elicited in conditions 2 through 11. *See supra* note 21. Nevertheless, 24% of responses in the No Forensics condition (which represents the prior probability distribution for the case) were responses of "50%."

necessarily reflect fallacious inference,[67] but even if these are added to those discussed in the previous paragraph, it still gives an incidence rate of less than two percent.[68] In view of the probable effect of deliberation in defusing such fallacies, this is still not a practically significant risk.[69]

Some earlier studies have found somewhat larger incidence of the Prosecutor's Fallacy.[70] In part this is due to the sensitivity of this fallacy to the exact way of communicating the frequency of the DNA profile in the relevant population (or, analogously, the frequency of lab errors). In all nine conditions with quantified random match testimony, these frequencies were communicated in terms that have been found to reduce the incidence of that fallacy.[71] A contributing cause of the difference may be the size of the RMP. In Phase I of this study, for example, an RMP of 0.04 (1 in 25) generated an incidence of this fallacy (using the same approximate measure) of about 6%.[72] In any event, risk of the Prosecutor's Fallacy does not seriously qualify the conclusions to be drawn from our results concerning the comparative utility of the various presentation formats, at least so long as testimony and legal argument at trial do not explicitly encourage this fallacy.[73]

Neglect of Lab Error. As noted above, it might be supposed that if the

---

[67] Again, the figure may understate or overstate the incidence of fallacy. Overstatement is particularly likely here in that responses of 99.9% are entirely understandable without reference to the idea of fallacy, as a way of saying "very nearly certain." Note, in particular, that the Chart formats (in which 3 of these responses were found) use illustrations that suggest such figures in accordance with Bayes' Rule (see Appendices A and B).

[68] This figure ignores the uninformative scaled responses. *See supra* note 65.

[69] *See* Nance & Morris, *supra* note 7, at 438-39 (arguing that incidence rates of as high as 4.6% pose no serious threat to the integrity of a post-deliberation verdict).

[70] *See* Thompson & Schumann, *supra* note 3, at 173-74 (reporting overall incidence rate of 13.2%, counting only the RMP-based fallacy); Jane Goodman, *Jurors' Comprehension and Assessment of Probabilistic Evidence*, 16 AM. J. TRIAL ADVOC. 361, 375 (1992) (reporting incidence of less than 2% counting only the RMP-based fallacy).

[71] Respondents were told that the DNA profile in question can be found "in only 25 ten-thousandths of one percent–.0025%–of the male population. In other words, on average we would expect to see the DNA profile[] . . . in one out of every 40,000 randomly selected men in the population." On the basis of earlier research, this was expected to minimize the potential incidence of the fallacy. *See* Thompson & Schumann, *supra* note 3 (reporting smaller incidence of the fallacy when the RMP is communicated as a frequency of incidence than when it is communicated as a probability of coincidental match); Jonathan J. Koehler, *When Are People Persuaded By DNA Match Statistics*, 25 LAW & HUM. BEHAV. 493 (2001) (expanding on this point using an "exemplar queing" theory).

[72] *See* Nance & Morris, *supra* note 7, at 429. Notice, however, that further refinement of the measure indicated that the 6% figure likely overstated the incidence. Id. at 431-34.

[73] *See* Thompson & Schumann, *supra* note 3, at 177-81 (Experiment 2: addressing effects of arguments of counsel).

possibility of lab errors is not quantified, while the risk of coincidental match is, a juror might act upon the premise that the lab error rate is essentially zero, inappropriately raising the respondent's rated probability of guilt and willingness to convict (Hypothesis 4). If so, then unquantified lab error formats (as in conditions 3, 6, and 9) might be thought to cause overvaluation of match evidence relative to, say, the Mere Match format.[74]

Our data, however, do not support this idea. In the first place, untutored expectancies of lab error implicating an innocent defendant are significant (mean=0.019; median=0.001). More importantly, as noted above, no main effect of lab error format on probabilities of guilt or verdict preferences was discovered. Further, when the Frequency, Likelihood Ratio, and Chart formats were combined, there was no significant difference among the three types of lab error testimony (i.e., unquantified, quantified, and aggregated) on the rated probability of guilt.[75] And most importantly, a planned comparison yielded no significant difference between the probability of guilt when there was an estimate of lab error (i.e., in the "quantified" lab error formats) and the probability of guilt when there was no estimate of the lab error rate (the "unquantified" formats).[76] Similarly, when the Frequency, Likelihood Ratio, and Chart formats were combined, there was no significant difference among the three types of lab error testimony on the frequency of guilty verdicts.[77] And a planned comparison yielded no significant difference between the frequency of guilty verdicts when there was an estimate of lab error (i.e., "quantified" lab error formats) and the frequency of guilty verdicts when there was no estimate of the lab error rate ("unquantified" formats).[78] In sum, we found no evidence that jurors will (in any systematic way) ignore the risk of lab error under any of these presentation formats.

Improper Averaging. As reflected in Hypothesis 5, another claim that has been

---

[74] The argument might run like this. Suppose the prosecution convinces the trial judge not to admit lab error estimates because those estimates are based on average performance of other labs; although no proficiency test data is available for the lab used in the present case, the prosecution's witness is confident that his lab is better than the average. *See* Koehler, *supra* note 55, at 429-37 (criticizing this and other arguments for excluding lab error estimates based on proficiency tests). The defense then replies by arguing that the statistical measure of the RMP should be excluded because introducing it might cause the jury to ignore the chance of lab error.

[75] $F_{(2, 1196)}=0.28$, p=0.76.

[76] $t_{(802)}=0.17$, p=0.86.

[77] $G^2_{(2, 1216)}=0.64$, p=0.73.

[78] $G^2_{(1, 808)}=0.34$, p=0.56.

made is that, without instructions on how to combine random match risk with lab error risk, jurors might average them rather than adding them in determining the total risk (from these sources) of a match report when the accused is innocent. This would have the effect of underestimating the combined risk and, therefore, overestimating the probative value of the match. A planned comparison of rated probabilities of guilt, however, indicated no significant difference between quantified and aggregated lab error formats.[79] Similarly, the averaging hypothesis predicted a higher frequency of guilty verdicts with only a separately quantified lab error rate than with the addition of an aggregated risk rate. A planned comparison, however, indicated no significant difference between these groups.[80] Overall, the averaging hypothesis is not supported by our data.

To be sure, the incidence of erroneous averaging could simply be undetectable because we did not attempt to elicit probabilities that would reveal them directly. Instead, we have looked for effects of averaging on the probability, if any, on which respondents would focus–the probability of guilt–and on verdict preferences. For these measures, the total expectancy of a false positive indication of guilt, and thus the likelihood ratio relative to the hypothesis of guilt, tends to be dominated by the relatively large risk of miscellaneous causes of a falsely incriminating match report. Moreover, it remains possible that some degree of averaging–or a similar combination strategy that leads to underestimation of the combined risk–might account for the slightly higher probability of guilt assessments and more noticeably higher guilt preferences in the quantified lab error formats than in the aggregated formats *when explanatory charts are not employed*.[81] Though not statistically significant even at these large sample sizes,[82] these differences might indicate an averaging effect that is reduced by giving aggregated rates but is also swamped (in the opposite direction) by the education effect of using the Chart format, as discussed above.[83]

Vividness Heuristic. It has been suggested that the presentation of a very small

---

[79] $t_{(797)}=0.73$, p=0.46. Median tests also failed to yield statistically significant differences.

[80] $G^2_{(1, 814)}=0.01$, p=0.91.

[81] *See* Schklar & Diamond, *supra* note 5, at 175-76 (reporting some averaging among undergraduates when presented the RMP and lab error rate estimate in what we have here called a Frequency format). But see id. at 174 (finding no significant effect of combination instruction on verdict preferences).

[82] Respectively: $t_{(533)}=0.38$, p=0.70; $G^2_{(1, 546)}=2.17$, p=0.14.

[83] *See supra* notes 55-56 and accompanying text.

RMP in testimony might make this the particularly memorable piece of information, as compared for example to larger, more pallid estimates of lab error rate, and that this effect could artificially increase rated probabilities of guilt.[84] This effect was not tested directly, because all conditions that employed information about the quantified random match probability involved testimony conveying that probability separately; even aggregated formats included not only the aggregated risk rate of 1 in 976, but also the two components used to calculate that rate (i.e., 1 in 40,000 and 1 in 1000).[85] To be sure, the higher probabilities of guilt in conditions 3 through 11, as compared to those for the Mere Match condition, are consistent with a vividness hypothesis. They are also, however, consistent with an obvious education effect: Jurors may conservatively attribute low weight to the match when not provided with quantification of the RMP because they are aware of the absence of information important for assessing the weight of the match evidence. And as already noted, an education effect tends to explain the pattern of variation within conditions 3 through 11, an effect that was at least not dominated by any constant vividness effect. Other research, moreover, has provided strong reasons to doubt that vividness is irrationally contributing to the higher probabilities of guilt and conviction rates obtained here in those nine conditions with testimony quantifying the chance of a coincidental match.[86]

C. Limitations of the Study

Unrealistic Stimuli. Some readers have commented that it might have been better if the testimony and jury instructions in the case were provided by video-taped enactments rather than in written form. This would certainly contribute to the realism of the stimuli to which the subjects responded. On the other hand, it would also introduce variables not easily evaluated. The results might be dependent on the particular traits of

---

[84] *See* Koehler et al., *supra* note 4, at 212.

[85] In arguing against the separate admission of very small RMPs, Professor Koehler reports findings consistent with vividness effects, but these appear by comparing quantified RMP formats with formats presenting aggregated error risks *but not* the components of the aggregation. *See id.* at 213-14. That the former would produce higher rates of conviction than the latter need not necessarily mean that the former reflects cognitive error or that the latter represents the more accurate result. Depriving subjects of the component information might decrease their understanding of, and thus confidence in, the communicated aggregate risk figure, which could suppress the weight they attribute to the match quite apart from any vividness effect. Indeed, the low probability of guilt ratings and willingness to convict rates for the Mere Match condition seem to reflect that kind of conservatism.

[86] *See* Schklar & Diamond, *supra* note 5, at 171-78 (reporting several specific results inconsistent with the vividness hypothesis and referring to other studies reaching similar conclusions).

the actors used, for example. Written testimony has the advantage of requiring the respondents to think in terms of an abstracted and typical presentation, to the extent that they envisioned the testimony and instructions that they read.

Absence of Deliberations. We sought responses from individuals, not the corporate responses of a group of deliberating individuals. While this obviously does not mirror actual trials, it nonetheless allows more specific information about individual respondents. Some modeling of the corporate response is possible and has been noted above.[87] It would, however, be desirable to see the effects found here replicated in experiments involving deliberation.

Neglect of Cases with Infinitesimal Random Match Probabilities. The present study used an RMP that was quite small (1 in 40,000), and much smaller than the testimonial lab error rate (1 in 1000). It was, however, still large enough plausibly to allow the detection and measurement of the incidence of the Prosecutor's Fallacy and the Inversion Fallacy. Many DNA cases today, though certainly not all, involve testimony about random match probabilities on the order of one in a million, one in a billion, or even smaller. Although one cannot assume that this difference matters, it must be acknowledged that such extreme figures might introduce effects that cannot be extrapolated from the present context.

Use of Non-Specific Lab Error Estimates. To the extent that lab error rates were quantified in the testimony, our respondents were told that the estimates were based on proficiency studies conducted nation-wide. Use of lab-specific or even analyst-specific proficiency test data, when such are available, might affect the results, although it is very difficult to discern *a priori* what those effects would be.[88]

Type of Case. The experiment used a hypothetical rape case. It is possible that the context of rape invokes social norms that could generate atypical responses. To be sure, the hypothetical used was a stranger assault, less likely to invoke such obfuscating norms than, perhaps, a date rape case or a case in which consent was asserted by the

---

[87] *See supra* note 45 and accompanying text.

[88] Although regular proficiency testing has been authoritatively recommended and is becoming the norm, there is disagreement about the feasibility of obtaining useful lab-specific or analyst-specific error rates for use at trial. *Compare* COMMITTEE ON DNA TECHNOLOGY IN FORENSIC SCIENCE, NATIONAL RESEARCH COUNCIL, DNA TECHNOLOGY IN FORENSIC SCIENCE 88-89 (1992) (stating that the results of regular blind proficiency tests should be disclosed to juries), *with* COMMITTEE ON DNA FORENSIC SCIENCE: AN UPDATE, NATIONAL RESEARCH COUNCIL, THE EVALUATION OF FORENSIC DNA EVIDENCE 85-87 (1996) (arguing that meaningful and up to date error rates for specific labs would require impractically large numbers of proficiency trials).

accused. Nonetheless, it would be useful to see some replication of these results in other kinds of cases.

CONCLUSION

Prior to the present research project, efforts to reduce the "misaggregation" of statistical match evidence with other evidence in the case, and thus to reduce the undervaluation of the match evidence, had not been able to report much success.[89] With the benefits of large sample sizes, in the context of typical citizens called for jury service, the two phases of this project has demonstrated that pessimism on this score is not as warranted as it may have seemed. Potential jurors' assessments of the probative value of DNA match evidence can be affected by variations in the format by which the chance of coincidental match is presented. Measured by Bayesian norms, juror assessments can be improved by providing appropriate instruction that offsets the otherwise extant, but reasonable–indeed laudable–tendency of jurors to discount the probative value of a scientific technique the results of which they do not adequately understand.

The incorporation of lab error estimates, based on proficiency tests, has comparatively little impact on jury assessments and decisions. It is a distinctly "second order" phenomenon, at least when, as here, the estimate that would be presented to the jury is not too far from what jurors are likely to expect anyway. Moreover, at least for cases in which the chance of coincidental match is dominated by the chance of lab error, our results show that incorporating lab error estimates in the testimony can actually increase the probative value that jurors attribute to the match testimony, causing their assessments to approach Bayesian norms. This occurs when three conditions are met: both the lab error estimate and the random match probability are communicated (as frequencies) to the jury; the testimony then combines the lab error rate with the random match probability to obtain a measure of the combined risk; and, the probative value associated with the combined measure is illustrated for the jury. This indicates the considerable importance that completeness of explanation plays in convincing a jury to accept what the science is telling them.

On the whole, our results tend to support a view of trial law as more urgently

---

[89] *See* Schklar & Diamond, *supra* note 5, at 164.

concerned with assisting the jury to understand the evidence, in this case a DNA match, than with shielding the jury from evidence that might lead them astray by triggering irrational mistakes or submissive credulity favoring the prosecution. Although fallacies of inference favoring the prosecution can occur, cognitive error favoring the opponent can also occur. Indeed, even in the context of a case, like one used in this experiment, that renders most of the pro-defense fallacies unlikely, for the most part jurors' innate skepticism and need to be convinced creates a dominating undervaluation of the evidence. This undervaluation can, however, be addressed by allowing expert witnesses to explain the significance of the random match probability using Bayesian methods.

Appendix A: Chart Used in Conditions 9 and 10

Effect of Likelihood Ratio of 40,000 to 1

(posterior probabilities rounded to nearest ten-thousandths of one percent)

Prior Probability		Posterior Probability
0 %	→	0 %
1 in 1 million	→	3.8463 %
1 in 500,000	→	7.4074 %
1 in 100,000	→	28.5716 %
1 in 10,000	→	80.0016 %
1 in 1,000	→	97.5634 %
1 in 100 (1 %)	→	99.7531 %
1 in 10 (10 %)	→	99.9775 %
20 %	→	99.9900 %
30 %	→	99.9942 %
40 %	→	99.9963 %
50 %	→	99.9975 %
60 %	→	99.9983 %
70 %	→	99.9989 %
80 %	→	99.9994 %
90 %	→	99.9997 %
100 %	→	100%

Appendix B: Chart Used in Condition 11

Effect of Likelihood Ratio of 976 to 1

(posterior probabilities rounded to nearest ten-thousandths of one percent)

Prior Probability		Posterior Probability
0 %	→	0 %
1 in 100,000	→	.9666 % (slightly less than 1%)
1 in 10,000	→	8.8929 %
1 in 1,000	→	49.4177 %
1 in 100 (or 1%)	→	90.7907 %
1 in 10 (or 10%)	→	99.0863 %
20 %	→	99.5918 %
30 %	→	99.7615 %
40 %	→	99.8465 %
50 %	→	99.8976 %
60 %	→	99.9317 %
70 %	→	99.9561 %
80 %	→	99.9744 %
90 %	→	99.9886 %
100 %	→	100%

Appendix C: Descriptive Statistics by Experimental Condition

Type of Random Match Evidence	Probability of Guilt					Guilty Verdicts		
Testimony about Lab Error	Mean	Std. Error	SD	Median	N	%	Std. Error	N
No Forensics (1)	0.31	0.02	0.23	0.30	135	0.04	0.02	134
Mere Match (2)	0.52	0.02	0.28	0.50	134	0.25	0.04	133
Frequency								
Unquantified (3)	0.62	0.02	0.29	0.60	135	0.38	0.04	133
Quantified (4)	0.63	0.02	0.28	0.62	136	0.40	0.04	139
Aggregated (5)	0.61	0.02	0.28	0.60	131	0.35	0.04	135
Total	0.62	0.01	0.28	0.60	402	0.37	0.02	407
Likelihood Ratio								
Unquantified (6)	0.68	0.03	0.30	0.75	137	0.48	0.04	134
Quantified (7)	0.65	0.02	0.28	0.75	131	0.49	0.04	134
Aggregated (8)	0.65	0.02	0.27	0.65	137	0.41	0.04	138
Total	0.66	0.01	0.29	0.75	405	0.46	0.02	406
Chart								
Unquantified (9)	0.67	0.02	0.29	0.75	134	0.47	0.04	135
Quantified (10)	0.69	0.03	0.30	0.80	131	0.50	0.04	133
Aggregated (11)	0.75	0.02	0.26	0.83	133	0.64	0.04	135
Total	0.70	0.01	0.28	0.80	398	0.54	0.02	403

Table 1. ANOVA of the Effect of Type of Random Match and Lab Error Testimony on Ratings of the Probability of Guilt. (N = 1205.)

Source	Sum of Squares	df	Mean Square	F	p	Eta Squared
Random Match	1.29	2	0.64	7.97	<0.001	0.013
Lab Error	0.05	2	0.02	0.28	0.76	0.000
Random Match x Lab Error	0.53	4	0.13	1.63	0.17	0.005
Error	96.74	1196	0.08			

Table 2. Loglinear Analysis of the Effect of Type of Random Match and Lab Error Testimony on Number of Guilty Verdicts. (N = 1216.)

Effect	df	Likelihood Ratio Chi-Square	p
Random Match	2	23.07	<0.001
Lab Error	2	0.64	0.73
Random Match x Lab Error	4	11.09	0.03

Table 3. False Positive Expectancies.

Source of Match Report	Untutored Expectancies		Tutored Expectancies	
	mean	median	mean	median
Random Match	0.019 (n=134; c.i.= ±0.010)	0 .0001	0.013 (n=1195; c.i.= ±0.004)	0.0001
Lab Error	0.019 (n=532; c.i.= ±0.004)	0.001	0.016 (n = 810; c.i.= ±0.004)	0.001
Miscellaneous	0.024 (n=1336; c.i.= ±0.004)	0.001	not applicable	

Figure 1. Mean rated probability of guilt and 95% confidence interval for each condition.

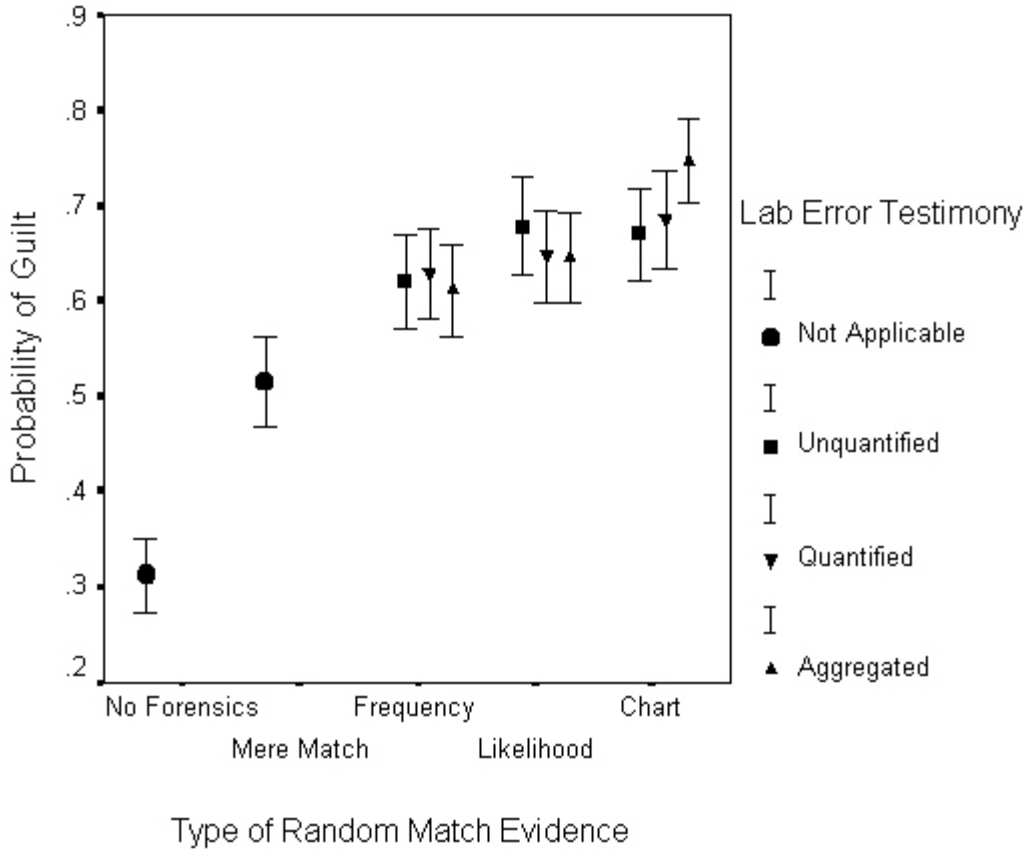

Figure 2. Proportion preferring guilty verdict and 95% confidence interval for each condition.