JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, 2018
doi: 10.1111/1556-4029.13797
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Brandon Garrett,[1] J.D.; Gregory Mitchell,[1] J.D., Ph.D.; and Nicholas Scurich,[2] Ph.D.

Comparing Categorical and Probabilistic Fingerprint Evidence*

ABSTRACT: Fingerprint examiners traditionally express conclusions in categorical terms, opining that impressions do or do not originate from the same source. Recently, probabilistic conclusions have been proposed, with examiners estimating the probability of a match between recovered and known prints. This study presented a nationally representative sample of jury-eligible adults with a hypothetical robbery case in which an examiner opined on the likelihood that a defendant's fingerprints matched latent fingerprints in categorical or probabilistic terms. We studied model language developed by the U.S. Defense Forensic Science Center to summarize results of statistical analysis of the similarity between prints. Participant ratings of the likelihood the defendant left prints at the crime scene and committed the crime were similar when exposed to categorical and strong probabilistic match evidence. Participants reduced these likelihoods when exposed to the weaker probabilistic evidence, but did not otherwise discriminate among the prints assigned different match probabilities.

KEYWORDS: forensic science, fingerprints, fingerprint identification, expert testimony, jury decision-making, probabilistic evidence, statistical evidence, scientific evidence

For over a hundred years, latent fingerprint examiners have conducted detailed visual examinations of evidence and reached conclusions about the probative value of that evidence for a criminal investigation. Traditionally, examiners summarize their conclusions using categorical language: Latent fingerprints collected from a crime scene either were or were not suitable for comparison, and, for those latent prints deemed suitable for comparison, the latent prints and inked fingerprints obtained from a suspect either did or did not originate from the same source (1). When fingerprint examiners present their conclusions at trial using categorical language, jurors receive a clear message: The suspect's fingerprints either do or do not match the fingerprints collected from the crime scene.

This clear message is potentially misleading, however, because jurors may assign greater certainty to the match/no-match opinion than is warranted. While some forensic techniques, most notably DNA analysis, generate statistics-based estimates of a "random match probability" that will be presented to fact finders in probabilistic terms, other techniques, such as fingerprint examinations, rely on subjective assessments of a match that do not lend themselves to specific probabilities. Examination of the forensic analyst at trial may reveal the subjective and imprecise nature of a match opinion, but courts have long allowed opinions presented in categorical terms. Accordingly, the fact finder's understanding of the probability of a match may differ from the subjective probability assigned by the forensic analyst.

Concerns about the accuracy of subjective matching methodologies (e.g., [2–4]) have led forensic science laboratories and professional associations to consider alterations to their methodologies (5). One innovation in this respect is the adoption of statistical software that measures the similarity between markers identified in crime scene specimens and known specimens, to calculate the probability that the specimens come from the same versus different sources. Such software has been developed by the U.S. Defense Forensic Science Center (DFSC) for fingerprint examinations. This program, FRStat, generates an estimate of the likelihood that comparison prints would share observed similarities if those specimens come from the same versus different sources (6); additional information about this program is available at https://osf.io/pmkwf/. DFSC also promulgated model language to communicate the results of this statistical analysis to fact finders:

> RESULTS OF EXAMINATION
>
> The latent print on Exhibit 1 and the standards bearing the name DOE have corresponding ridge detail. The probability of observing this amount of correspondence is approximately [XXX] times greater when impressions are made by the same source rather than by different sources.

Defense Forensic Science Center fingerprint examiners now use this language to present the results of their fingerprint comparisons (6).

Innovations such as FRStat and the DFSC's model language for presenting the results of this program raise concerns that jurors may misuse or misunderstand statistical evidence (e.g., [7,8]). "Current research on jury interpretation of probabilistic evidence is sparse, but the research that has been done has tended to indicate that jurors are not particularly good at

[1] University of Virginia School of Law, 580 Massie Road, Chrlottesville, VA 22903.

[2] School of Social Ecology, University of California-Irvine, Irvine, CA.

*Supported by the Center for Statistics and Applications in Forensic Evidence (CSAFE) and the National Institute for Standards and Technology (NIST).

Received 4 Dec. 2017; and in revised form 2 Feb. 2018, 24 Mar. 2018; accepted 26 Mar. 2018.

assigning it appropriate weight within the context of a case" ([5], p. 80).

Much of the research into jury comprehension of statistical evidence has involved DNA evidence, which presents both methodological and numerical complexities. Studies have shown that jurors are sensitive to the probabilities associated with DNA evidence, but they sometimes misunderstand random match probabilities (9–11). Small changes in how DNA statistics are presented (e.g., using percentages versus frequencies to communicate random match probabilities) can alter the weight given to the evidence (12). But clear explanations of the proper interpretation of DNA evidence can improve juror comprehension (13).

This study sought to examine how jury-eligible adults comprehend probabilistic fingerprint evidence, in particular fingerprint evidence presented using the DFSC's new model language relative to fingerprint evidence presented in traditional categorical (match/no-match) terms. By comparing the weight given to probabilistic versus categorical match evidence, we can obtain a better understanding of the match probabilities that jurors assign to categorical fingerprint evidence and examine whether jurors share common interpretations of different levels of probabilistic evidence (e.g., does fingerprint evidence presented as one million times more likely to be from the same than different persons have a greater impact on jurors than fingerprint evidence presented as only ten times more likely to come from the same person?).

In prior research, we found that the particular categorical language used by fingerprint examiners mattered little: Categorical match opinions carried considerable weight regardless of whether the examiner stated that opinion starkly, bolstered it with unwarranted statements of certainty, or qualified it by admitting that someone other than the defendant might be the source of the latent prints (1). This prior research did not examine, however, whether the subjective probabilities of a match, as determined by an examiner, corresponded to the jurors' assigned probabilities. Although the present research does not gather data on fingerprint examiners' subjective probabilities, it does provide insight into the weight that jurors assign to a fingerprint identification described in categorical terms.

Method

Participants

We commissioned Qualtrics to recruit a nationally representative sample of adult participants with respect to gender, race/ethnicity, age, income, and geographic region in the United States. The 1050 adults who participated in the study each received approximately $3 for their participation, which took less than 10 min. We report the results based on responses of 858 participants who passed three quality control checks while completing the survey (14): (i) After giving informed consent, these participants committed to reading the questions and giving their best answers; (ii) these participants responded correctly to a question approximately midway through the survey directing them to enter a particular answer to ensure that they were attending to the questions; (iii) these participants responded correctly to a question near the end of the survey asking them to give the sum of one plus three. Inclusion of the full sample in the analyses reported below does not alter the results in any substantial way, but we focus on this subsample because we have greater confidence that they carefully considered the questions that we posed to them. The full data set and the experimental materials used in the study are publicly available at https://osf.io/pmkwf/.

The high-attention subsample consisted of 419 (49%) males and 439 (51%) females, with a mean age of 46.54 (SD = 16.82), a median age of 46, and an age range of 18–90. Seventy-five percent of the sample had at least some post-high-school education, 23.8% had a 4-year college degree, and 11.8% had a professional or doctorate degree. Sixty-eight percent of the subsample identified as White, 12% as Black, 4% as Asian, 13.4% as Hispanic, 0.5% as American Indian/Alaska Native, and 1.3% as other. Thirty percent self-identified as Republican, 39.5% as Democrat, and the remainder stated that they had no consistent political preference. Twenty-two percent of participants resided in the northeastern region of the United States, while 23% resided in the west, 17.4% in the midwest, and 37.5% in the south. The median household annual income fell in the range of $40,000–$49,999. Approximately one-third of the subsample reported previously serving on a jury, and 31.7% stated that they or a family member had been arrested by the police.

Procedure and Materials

The experimental materials were created and provided to the participants using Qualtrics' online survey platform. After participants completed initial demographic questions, they were presented with a vignette describing a robbery that occurred at a convenience store. In brief, an assailant wearing a mask robbed a convenience store with a gun, but the assailant dropped the gun when exiting the store. An individual was arrested shortly after the crime, but no proceeds of the crime were found on the person and the store clerk could not positively identify the arrestee as the robber because the robber had worn a mask.

Participants were randomly assigned to receive one of nine different versions of the hypothetical case. Participants assigned to the control condition received no further information about the case (i.e., these participants received no information about fingerprint evidence). Participants assigned to one of the eight fingerprint evidence conditions were told that a fingerprint examiner compared fingerprints recovered from the handle of the gun dropped at the crime scene to the defendant's fingerprints and determined that the prints matched.

In two of the fingerprint evidence conditions, the fingerprint examiner used categorical language to describe the fingerprint match: In the "simple match" condition, the examiner concluded that "the fingerprint was individualized as the right thumb of the defendant"; in the "strong match" condition, the examiner added to the simple match language that it was a "practical impossibility that the prints came from a different source." The remaining six scenarios presented the fingerprint examiner's testimony in probabilistic terms based on the DFSC's model language:

> The probability of observing this amount of correspondence is approximately [1,000,000; 100,000; 10,000; 1,000, 100; 10] times greater when impressions are made by the same source rather than by different sources. This conclusion was reached using software that measures the degree of similarity between fingerprint impressions.

It is worth noting that how the FRstat software is described to jurors may affect juror acceptance of the results of that software; however, we did not in this study test the impact of alternative ways of presenting the software. Future research should examine this possibility.

After reading about the case, participants responded to our main dependent variables by (i) stating whether they would convict the defendant (yes or no); (ii) estimating the likelihood that the defendant left his prints on the gun on a scale ranging from 0 (certainly did not leave his prints on the gun) to 100 (certainly did leave his prints on the gun), with 50 corresponding to complete uncertainty; and (iii) estimating the likelihood that the defendant committed the robbery using the same scale. Participants were given the following information on how to use the 0–100 scales: "Numbers below 50 indicate that you think it is more likely that the defendant did *not* commit the robbery (the smaller the number, the less likely). Numbers above 50 indicate that you think it is more likely the defendant *did* commit the robbery (the bigger the number, the more likely). 0 represents certainty that the defendant did *not* commit the robbery. One hundred represents certainty that the defendant *did* commit the robbery. Fifty indicates uncertainty as to whether the defendant did or did not commit the robbery" (participants were given similar guidance on how to use the scale to indicate the likelihood the defendant left prints on the gun). After completing these main dependent measures, participants completed a few general questions about forensic evidence and their aversions to false convictions and false acquittals.

Results

Across all conditions, 50% percent of participants indicated that they would vote to convict the defendant. In the control condition, where no fingerprint evidence was presented, only 22.5% of participants indicated that they would vote for conviction, whereas 53.7% would vote for conviction when fingerprint evidence was presented. Thus, when fingerprint evidence of any kind was presented, participants were significantly more likely to vote in favor of conviction, $\chi^2(1) = 34.89$, $p < 0.001$. A logistic regression comparing the odds of conviction in each fingerprint condition to the odds of conviction in the control condition reported significant results for each comparison, with all Exp (β)'s > 11.00 and all p's < 0.01.

We next compared conviction rates in the simple match condition to conviction rates in the strongest and weakest probabilistic match conditions (i.e., the conditions in which the examiner opined that it was 1 million versus 10 times more likely that the latent print came from the defendant than someone else), but we found no significant differences in conviction rates. In other words, when participants had to make the dichotomous choice of convict or not, participants were just as likely to convict when the fingerprint evidence was presented in categorical terms as probabilistic terms, regardless of the specific language used to describe the match or probability of a match (Fig. 1 presents the conviction rates for each condition).

We also examined whether the different ways of presenting the fingerprint evidence affected responses on the two more sensitive-dependent variables, which were scored on 0–100 scales: (i) the likelihood the defendant left prints on the gun and (ii) the likelihood the defendant committed the crime. On both of these variables, there were significant differences in ratings across the fingerprint evidence conditions ($F(7,748) = 2.529$, $p = 0.014$ and $F(7,748) = 3.159$, $p = 0.003$, respectively). These findings held when the data were standardized (z-scored) as well. We then compared ratings on these two dependent variables within the simple match condition to the ratings in the strongest and weakest probabilistic match conditions (Fig. 2 presents mean ratings on the likelihood left prints and likelihood committed the crime questions across the experimental conditions). Ratings on these two continuous variables did not differ significantly between the simple match and strongest probabilistic match condition, but the ratings on both variables differed significantly between the simple match and weakest probabilistic match condition (both F's > 10.00 and both p's = 0.001). Finally, we compared ratings on these two variables between the strongest and weakest probabilistic match conditions: The difference on the likelihood of committing the crime variable was statistically significant ($F(1,193) = 4.56$, $p = 0.034$), but the difference on the likelihood of leaving prints on the gun was not ($F(1,193) = 2.84$, $p = 0.094$). When we compared the second strongest probabilistic evidence (100,000 times more likely that defendant rather than another was the source) to the weakest probabilistic evidence (only 10 times more likely defendant was the source), we observed no significant difference in ratings on either of the continuous dependent variables.

We next examined the contribution of individual difference variables on participants' ratings of the likelihood the defendant left his prints on the gun dropped at the crime scene. We entered into the regression the following possible explanatory variables: experimental condition (i.e., an indicator variable for the fingerprint evidence condition to which the participant was assigned), participant sex, age, race/ethnicity, political preference, regional location, prior jury service (yes/no), prior arrest of self or family member (yes/no), numeracy as measured by an objective test of persons' mathematical knowledge, and error aversion (whether the person viewed a false conviction or false acquittal as more serious or saw the errors as equally bad). The model that most efficiently explained variance in ratings on the "left prints" variable contained only numeracy, error aversion, and age ($R^2 = 0.06$, $p < 0.001$) (adding the political preference variable to numeracy, error aversion, and age significantly improved prediction but only slightly increased the variance explained [change in $R^2 = 0.005$, $p < 0.05$]). Regardless of experimental condition, participants higher in numeracy tended to give higher ratings to the likelihood the defendant's prints were left on the gun ($r_{\mathrm{partial}} = 0.202$). Participants who believed false convictions are worse than false acquittals tended to give lower ratings ($r$partial = 0.129). Older participants tended to give higher ratings ($r_{\mathrm{partial}} = 0.120$). Persons rating false convictions and false acquittals as equally bad on average rated the likelihood that the defendant left his prints on the gun highest ($M = 65.96$), followed by those rating false acquittals as worse ($M = 62.66$) and those rating false convictions worse ($M = 59.85$). Only the difference in ratings between the "equally bad" and "false convictions worse" group was statistically significant ($t(711) = 2.813$, $p = 0.005$).

The results for the error aversion variable, which to date has not received much attention as an influence on evidentiary assessments, are particularly interesting for two reasons. First, the majority of respondents indicated that they considered false convictions and false acquittals equally bad ($n = 431$; 50.2%), while approximately one-third of respondents rated false convictions as worse than false acquittals ($n = 282$; 32.9%) and a smaller but still sizeable percentage rated false acquittals as worse ($n = 145$; 16.9%). Recall that we collected a nationally representative sample, which suggests that many Americans do not share the common assumption within the law that false convictions are more serious than false acquittals. Second, we observed sex differences on this variable: About equal

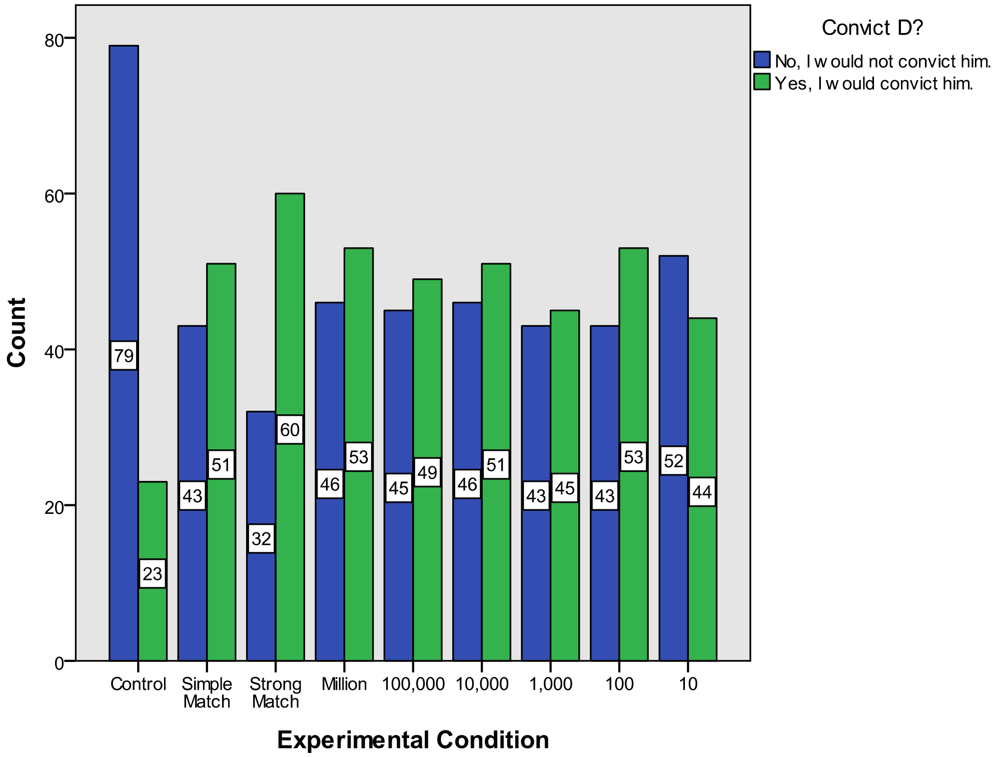

FIG. 1—*Conviction rates across experimental condition.*

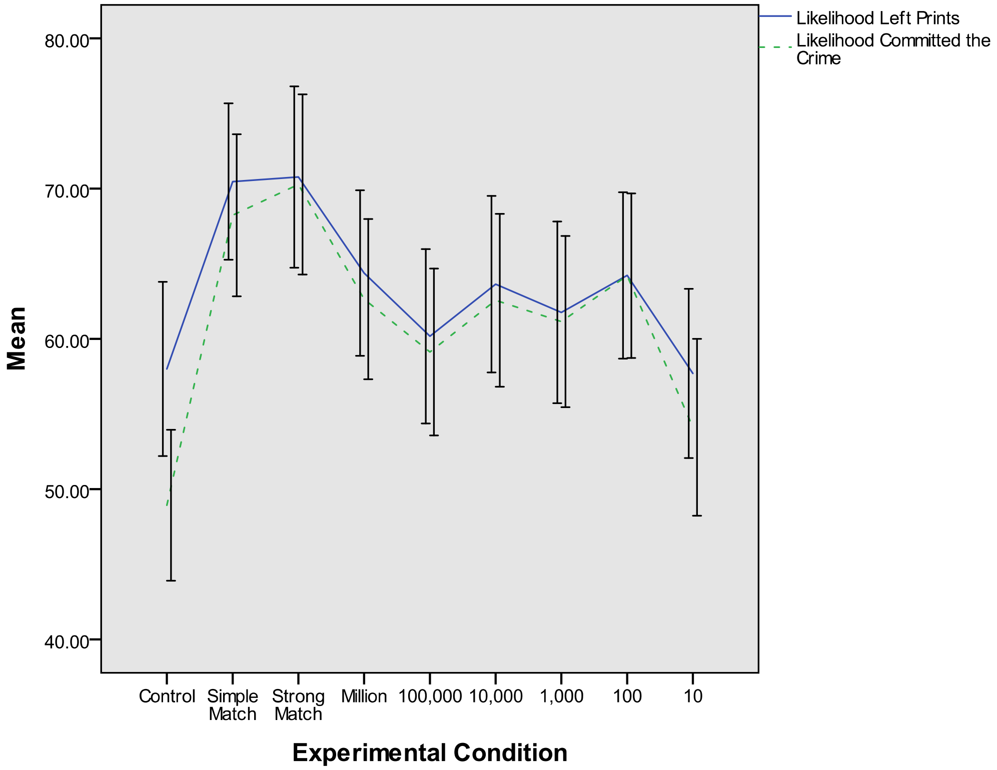

FIG. 2—*Mean ratings of likelihood defendant left prints and committed the crime across experimental conditions. Note: error bars = 95% confidence interval.*

percentages of women and men rated false acquittals as worse (15.9% vs. 17.9%), but fewer women than men rated false convictions as worse (32.9% vs. 37.9%), and more women than men rated the two errors as equally bad (50.2% vs. 44.2%).

Finally, we collected data on general views about the reliability of fingerprint and DNA evidence and their value as unique identifiers of individuals. The modal response for both fingerprint and DNA evidence in terms of reliability was "very reliable": 70.9% rated fingerprint evidence as very reliable or reliable, whereas 87.1% rated DNA evidence as very reliable or reliable; 23.9% rated fingerprint evidence as somewhat reliable compared to 9.7% for DNA evidence; fewer than 5% rated fingerprint evidence somewhat unreliable or unreliable compared to fewer than 3% for DNA evidence. Generally, both fingerprint and DNA evidence were seen as reliable forms of identification evidence, with DNA faring somewhat better. In terms of their value as unique identifiers of individuals, both types of evidence were seen as unique identifiers of particular individuals. We asked participants to estimate how many other people of the approximately 7 billion people in the world have fingerprints and DNA identical to their own: The modal response on both questions was zero, which was the response given by 76% of the sample on the fingerprint question and 77% on the DNA question. A few respondents estimated that a handful of other people shared their fingerprints or DNA, but 90% of the sample believed that 12 or fewer other people shared their fingerprints or DNA. These results suggest that most Americans assume that their fingerprints and DNA are unique to themselves (i.e., do not match the fingerprints or DNA of anyone else).

Discussion

Our results suggest that the traditional categorical approach to fingerprint evidence, in which the fingerprint examiner declares a match between latent and inked prints, carries great weight with laypersons, but our results also show that a strong probabilistic statement about the likelihood of a match carries similar weight. Potential jurors, on reading about weaker probabilistic fingerprint evidence, adjusted downward their estimates of the likelihood a defendant left prints at the crime scene, as they should have. However, we also found that jurors did not meaningfully distinguish between a wide range of probabilities that objectively differed greatly. While the probability of 1,000,000 times greater produced significantly higher estimates that the defendant left his prints on the gun and committed the crime, the lower probabilities ranging from 100,000 to just ten times more probable did not produce significant differences in these estimates (we did, however, observe a downward trend in these estimates with the lowest probability evidence, as shown in Fig. 2). This finding suggests the need for further research on how to better promote juror discrimination among degrees of probability, including quite disparate degrees of probability. Our findings suggest that it may be a challenge to do so by relying simply on conclusion language, given our finding that even the objectively weaker probabilistic fingerprint evidence was sufficient to move many jurors from a vote of acquittal to a vote of conviction. One reason why even relatively weak fingerprint evidence is likely to be convincing is that our survey showed that most jurors enter trials with the prior belief that fingerprint evidence is reliable and that fingerprints are unique to particular individuals.

We did not observe the confusion that studies have found when studying perceptions of DNA evidence. However, the model language utilized by the DFSC keeps a discussion of the technical details to a minimum and presents the probabilities derived from the FRStat software in straightforward terms. A searching cross-examination that raises questions about the validity of the estimates generated and about the meaning of the probabilistic language might introduce confusion or concerns that would reduce the understanding and weight of DFSC fingerprint examiners. Future studies should examine this possibility. Furthermore, we intentionally kept the evidence offered in our hypothetical criminal case minimal, to focus participants on the fingerprint evidence testimony. Future studies should examine how jurors react to different formulations of fingerprint examiner testimony when there is a greater mix of pro-prosecution and pro-defense evidence. It is likely that juror interpretations of relatively weaker probabilistic match evidence will depend on the strength of the corroborating evidence.

Finally, our study suggests an important line of inquiry into the role that jurors' error aversions play in their assessments of evidence. Jurors predisposed to concerns about false acquittals may have lower thresholds for acceptance of fingerprint evidence or prosecution evidence more generally. We did not instruct our mock jurors on the presumption of innocence or reasonable doubt; whether such instructions can overcome the concerns many jurors have about false acquittals poses an important question for further study (15).

Conclusion

This study demonstrates that jury-eligible adults can make distinctions among fingerprint evidence presented in probabilistic terms and will place as much weight on high-probability match evidence as categorical match evidence. The results provide some support for the move toward assigning probabilities to forensic identifications and for a move away from categorical conclusions that may mask the degree of uncertainty attached to an identification opinion. However, jurors did not discriminate among any but the highest and lowest probability evidence, and failed to distinguish between match probabilities ranging from ten to one hundred thousand. Thus, these results also suggest that more research is needed on how to present statistical evidence in a way that will cause jurors to distinguish more carefully between degrees of probability.

References

1. Garrett B, Mitchell G. How jurors evaluate fingerprint evidence: the relative importance of match language, method information and error acknowledgement. J Empir Leg Stud 2013;10(3):484–511.
2. National Research Council Committee on Identifying the Needs of the Forensic Science Community. Strengthening forensic science in the United States: a path forward. Washington, DC: The National Academies Press, 2009;136–45.
3. President's Council of Advisors on Science and Technology (PCAST). Scientific criteria for validity and reliability of forensic feature-comparison methods. In: Report to the President: forensic science in criminal courts: ensuring scientific validity of feature-comparison methods. North Charleston, SC: CreateSpace Publishing, 2016;44–65.
4. American Association for the Advancement of Sciences (AAAS). Forensic science assessments: a quality and gap analysis, latent fingerprint examination, 2017;8–12; https://www.aaas.org/page/forensic-science-assessments-quality-and-gap-analysis (accessed March 26, 2018).
5. Eldridge H. The shifting landscape of latent print testimony: an American perspective. J Forensic Sci Med 2017;3(2):72–81.
6. Defense Forensic Science Center (DFSC). Information paper: modification of latent print technical reports to include statistical calculations, 2017; https://osf.io/pmkwf/(accessed March 26, 2018).
7. Goodman J. Jurors' comprehension and assessment of probabilistic evidence. Am J Trial Advoc 1992;16:361–90.
8. Tribe LH. Trial by mathematics: precision and ritual in the legal process. Harvard Law Rev 1971;84(6):1329–93.
9. Koehler JJ, Chia A, Lindsey S. The random match probability (RMP) in DNA evidence: irrelevant and prejudicial? Jurimetrics 1995;35(2):201–19.
10. Thompson WC, Kaasa SO, Peterson T. Do jurors give appropriate weight to forensic identification evidence? J Empir Legal Stud 2013;10(2):359–97.

11. Thompson WC, Schumann EL. Interpretation of statistical evidence in criminal trials: the prosecutor's fallacy and the defense attorney's fallacy. Law Hum Behav 1987;11(3):167–87.
12. Koehler J, Macchi L. Thinking about low-probability events. An exemplar-cuing theory. Psychol Sci 2004;15(8):540–6.
13. Kaye DH, Hans VP, Dann BM, Farley E, Albertson S. Statistics in the jury box: how jurors respond to mitochondrial DNA match probabilities. J Empir Legal Stud 2007;4(4):797–834.
14. Oppenheimer DM, Meyvis T, Davidenko N. Instructional manipulation checks: detecting satisficing to increase statistical power. J Exp Soc Psychol 2009;45(4):867–72.
15. Scurich N, John RS. Jurors' presumption of innocence. J Legal Stud 2017;46(1):187–206.

Additional information and reprint requests:
Brandon Garrett, J.D.
School of Law
University of Virginia
580 Massie Road
Charlottesville
VA 22903
E-mail: bgarrett@virginia.edu