Law and Human Behavior
2014, Vol. 38, No. 1, 84–93

0147-7307/14/$12.00 DOI: 10.1037/lhb0000051

Human Matching Performance of Genuine Crime Scene Latent Fingerprints

Matthew B. Thompson
The University of Queensland and National Information and Communications Technology Australia

Jason M. Tangen
The University of Queensland

Duncan J. McCarthy
Queensland Police Service

There has been very little research into the nature and development of fingerprint matching expertise. Here we present the results of an experiment testing the claimed matching expertise of fingerprint examiners. Expert (n = 37), intermediate trainee (n = 8), new trainee (n = 9), and novice (n = 37) participants performed a fingerprint discrimination task involving genuine crime scene latent fingerprints, their matches, and highly similar distractors, in a signal detection paradigm. Results show that qualified, court-practicing fingerprint experts were exceedingly accurate compared with novices. Experts showed a conservative response bias, tending to err on the side of caution by making more errors of the sort that could allow a guilty person to escape detection than errors of the sort that could falsely incriminate an innocent person. The superior performance of experts was not simply a function of their ability to match prints, *per se,* but a result of their ability to identify the highly similar, but nonmatching fingerprints as such. Comparing these results with previous experiments, experts were even more conservative in their decision making when dealing with these genuine crime scene prints than when dealing with simulated crime scene prints, and this conservatism made them relatively less accurate overall. Intermediate trainees—despite their lack of qualification and average 3.5 years experience—performed about as accurately as qualified experts who had an average 17.5 years experience. New trainees—despite their 5-week, full-time training course or their 6 months experience—were not any better than novices at discriminating matching and similar nonmatching prints, they were just more conservative. Further research is required to determine the precise nature of fingerprint matching expertise and the factors that influence performance. The findings of this representative, lab-based experiment may have implications for the way fingerprint examiners testify in court, but what the findings mean for reasoning about expert performance in the wild is an open, empirical, and epistemological question.

Keywords: fingerprints, decision making, expertise, testimony, law

Fingerprint examiners have been active in investigations and have presented identification evidence in criminal courts for more than a century (Cole, 2002). Remarkably, given that testimony about fingerprint matches is a product of human judgment and subjective decision making, there have been few scientific investigations of the human capacity to correctly match fingerprints. Examiners have claimed that fingerprint identification is infallible (Federal Bureau of Investigation, 1984) and that there is a zero error rate for fingerprint comparisons (Cole, 2005; Edwards, 2009). These claims of individualization and a zero error rate, however, are not supported by evidence and are scientifically implausible (Cole, 2010; National Research Council, 2009; Saks & Faigman, 2008). As a result, former President of the International Association for Identification suggested that members not assert 100% infallibility (zero error rate) of fingerprint comparisons (Garrett, 2009) and the Scientific Working Group on Friction Ridge Analysis, Study and Technology (2012) has drafted a standard for defining, calculating, and reporting error rates. Recently, there has been a shift in the way fingerprint identification is regarded (Tangen, 2013). The acknowledgment that humans cannot be detached from forensic decision making has been highlighted in a variety of recent inquiries by the U.S. National Research Council of the National Academy of Sciences (2009), the Scottish Fingerprint Inquiry (Campbell, 2011), and the National Institute for Standards and Training and the U.S. National Institute of Justice (2012).

The National Academy of Sciences (NAS, 2009) has highlighted the absence of solid scientific methods and practices in U.S. forensic science laboratories. Harry T. Edwards (a senior U.S. judge and cochair of the NAS Committee) noted that forensic science disciplines, including fingerprint comparison, are typically not grounded in scientific methodology, and forensic experts do not follow scientifically rigorous procedures for interpretation that ensure that the forensic evidence that is offered in court is valid

This article was published Online First July 22, 2013.

Matthew B. Thompson, School of Psychology, The University of Queensland, St. Lucia, Australia and National Information and Communications Technology Australia, Queensland Research Laboratory, Brisbane, Australia; Jason M. Tangen, School of Psychology, The University of Queensland; Duncan J. McCarthy, Forensic Services Branch, Queensland Police Service, Brisbane, Australia.

Correspondence concerning this article should be addressed to Matthew B. Thompson, School of Psychology, The University of Queensland, St. Lucia QLD 4072, Australia. E-mail: mbthompson@gmail.com

and reliable (Edwards, 2009; see also Risinger, Saks, Thompson, & Rosenthal, 2002; Saks & Koehler, 2005). The NAS report (2009) highlighted the absence of experiments on human expertise in forensic pattern matching: "The simple reality is that the interpretation of forensic evidence is not always based on scientific studies to determine its validity. This is a serious problem." They recommended that the U.S. Congress fund basic research to help the forensic community strengthen their field, rectify the lack of basic research, develop valid and reliable measures of performance, understand the effects of bias and human error, and establish evidence-based standards for analyzing and reporting forensic testimony. Subsequent reports in the United Kingdom and United States have focused directly on fingerprint evidence.

An inquiry into fingerprint evidence was conducted by Lord Campbell (2011) following the controversial McKie case in Scotland. The former police detective, Shirley McKie, was accused by fingerprint examiners of leaving her fingerprint on the bathroom door frame of a murder crime scene, a charge she denied. The report recommends that fingerprint evidence should be recognized as opinion evidence, not fact; examiners should discontinue reporting conclusions on identification or exclusion with a claim to 100% certainty or infallibility; and that examiners should receive training that emphasizes that their findings are based on their personal opinion and subjective interpretation.

Most recently, a large multidisciplinary collective—the Expert Working Group on Human Factors in Latent Print Analysis (2012)—was sponsored by the U.S. National Institute of Standards and Technology and the National Institute of Justice to investigate human factors in latent fingerprint identification. The authors recommended that examiners should be familiar with human factors issues such as fatigue, bias, cognitive and perceptual influences, and not state that errors are inherently impossible or that a method inherently has a zero error rate. They recommend that management foster a culture in which it is understood that some human error is inevitable and that a comprehensive testing program of competency and proficiency should be developed and implemented. Speaking generally, and taking the lead from medical and aviation research, the authors advocate that fingerprint identification would benefit from the human factors research and systems approaches to improve quality and productivity, and reduce the likelihood and consequences of human error.

As a result of these reports and of scholarly criticism, changes in policy and research programs have begun. There are two proposed bills currently before the United States Congress calling for more research into forensic identification and changes to the funding, organization, standards, and regulation of forensic science ("Leahy proposes," 2011; Maxmen, 2012), and research into fingerprint identification is well underway. Researchers have investigated the effect of contextual bias on fingerprint examiners (Dror & Cole, 2010; Dror & Rosenthal, 2008; Langenburg, Champod, & Wertheim, 2009), the special abilities and vulnerabilities of fingerprint examiners (Busey & Dror, 2010; Busey & Parada, 2010; Busey et al., 2011), the psychophysics of fingerprint identification (Vokey, Tangen, & Cole, 2009), the effect of technology (Dror & Mnookin, 2010; Dror, Wertheim, Fraser-Mackenzie, & Walajtys, 2012), and statistical models of fingerprint identification (Champod & Evett, 2001; Neumann, 2012; Neumann et al., 2007). Two recent experiments have been conducted to directly address the matching accuracy and expertise of examiners.

Ulery, Hicklin, Buscaglia, and Roberts (2011) set out to measure the matching performance of latent print examiners. They had 169 latent print examiners each compare around 100 pairs of latent and exemplar fingerprints from a pool of 744 pairs. They focused on examiners' accuracy in the comparison process (i.e., the extent to which examiners can accurately match a latent print to its source). The researchers manufactured their own latent fingerprints so the ground truth is known, and they included similar, but nonmatching, distractors from a search of a national computer database containing approximately 580 million individual fingerprints. They reported an overall false alarm rate of 0.1% (i.e., incorrectly judging nonmatching prints to be a "match"). And 85% of examiners made at least one miss (i.e., incorrectly judging matching prints to be a "nonmatch") for an overall miss rate of 7.5%. Refer to Figure 1 for a description of the two ways of being right and two ways of being wrong in a basic fingerprint comparison task. Note, however, that Ulery et al. (2011) allowed examiners to give "inconclusive" and "no value" responses, and when the no value responses are discounted and the inconclusive responses are translated into misses, the overall miss rate is closer to 60% (an extremely conservative response bias). Although the experiment did not include a comparison group of participants (e.g., laypersons), it is clear that fingerprint examiners demonstrate impressive pattern matching abilities that may rival those of medical diagnosticians (Thompson, Tangen, & McCarthy, 2013). The rigorous experimental design, coupled with the large number of participants and stimuli, makes it one of the most important contributions to our understanding of expert matching performance.

Tangen, Thompson, and McCarthy (2011) set out to determine whether fingerprint experts are any more accurate at matching prints than the average person, and to get an idea of how often experts make errors of the sort that could allow a guilty person to escape detection compared with how often they make errors of the sort that could falsely incriminate an innocent person. In a two-alternative forced choice design, 37 qualified fingerprint experts

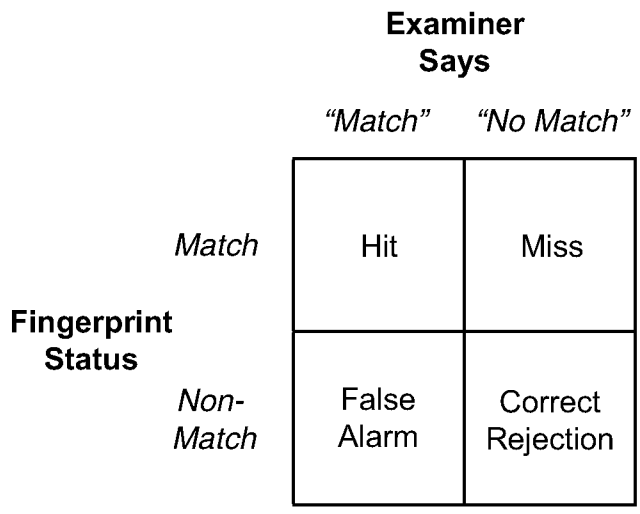

Figure 1. Fingerprint discrimination contingency table. A 2 × 2 contingency table depicting the four possible outcomes of a forced choice fingerprint discrimination task where two prints match or not and an examiner declares them as a "match" or "no match."

and 37 undergraduate students were presented with pairs of fingerprints and asked to indicate whether a simulated crime scene print matched a potential "suspect" or not. Some of the print pairs matched, and others were highly similar but did not match. Thirty-six simulated crime scene prints were paired with fully rolled exemplar prints. Across participants, each simulated print was paired with a fully rolled print from the same individual (match), with a nonmatching but similar exemplar (similar distractor), and with a random nonmatching exemplar (nonsimilar distractor). The simulated prints and their corresponding fully rolled print were from the Forensic Informatics Biometric Repository (see FIB-R.com for details), so, unlike genuine crime scene prints, they had a known true origin (Cole, Welling, Dioso-Villa, & Carpenter, 2008; Koehler, 2008). Similar distractors were obtained by searching the Australian National Automated Fingerprint Identification System. For each simulated print, the most highly ranked nonmatching exemplar from the search was used if it was available in the Queensland Police 10-print hardcopy archives, which contains approximately 1 million 10-print cards (10 million individual prints) from approximately 300,000 to 400,000 people (one person may have more than one 10-print card on record). Of the prints that actually matched, the experts correctly declared 92.12% of them as matching (hits). Of the prints that did not actually match, the experts incorrectly declared 0.68% of them as matching (false alarms). The rate of expert false alarms is impressive considering the corresponding false alarm rate for novices was 55.18%. Tangen et al. (2011) concluded that the qualified court-practicing fingerprint experts were exceedingly accurate compared with novices, and that the experts tended to err on the side of caution by making errors of the kind that would fail to identify a criminal rather than provide incorrect evidence to the court.

The experiment by Tangen et al. (2011) did not focus on the absolute performance of experts but on the comparison between experts and novices and between matching and nonmatching prints. So even though a false alarm rate for experts of 0.68% is impressive in its own right, this experiment cannot determine whether this rate reflects the false alarm rate of the field more generally. But it can be concluded that a false alarm rate of 55.18% for novices pales in comparison with experts (Thompson et al., 2013).

Despite the above contributions to forensic decision making, still very little is known about human fingerprint matching performance, the nature of expertise in fingerprint identification, the factors that affect matching accuracy, and the basis on which examiners can reasonably testify in court. Considering the shift toward viewing the human as an integral part of the forensic identification process, systematic programs of research are needed to understand the skills, abilities, and limits of fingerprint examiners, and to understand the nature of their expertise. Research programs that are under way or are to be developed include understanding the nature of forensic expertise, the influence of cognitive and perceptual biases, the impact of technology, how best to present pattern evidence to judges and juries, the best ways to turn novices into experts, and the most effective and efficient work practices, environments, and tools. Before these research programs can advance, however, a foundation for understanding expertise and accuracy in human fingerprint identification is needed. Here, we present a first step in our research program into the nature of forensic expertise in fingerprint identification.

Overview of the Present Research

In the experiment reported here we investigated the matching performance and expertise of human fingerprint examiners by replicating and extending on the work of Tangen et al. (2011). We increased the fidelity of the discrimination task (i.e., the resemblance of the discrimination task to actual casework) by using genuine crime-scene latents (and their matched exemplars) from police training materials, compiled from casework. The increased fidelity, however, reduces experimental control because the ground truth of the matched fingerprint pairs cannot be certain (Thompson et al., 2013). We also made the addition of two trainee groups and asked four groups of people to perform a fingerprint discrimination task—novices, new trainees, intermediate trainees, and qualified experts—in order to compare their relative performance.

Method

Participants

Four distinct groups participated in the experiment: novices, new trainees, intermediate trainees, and qualified, court-practicing experts. Novices were 37 undergraduates from The University of Queensland who participated for course credit and who had no experience with fingerprints. New trainees, intermediate trainees, and qualified experts were from five police organizations: The Australian Federal, New South Wales, Victoria, South Australia, and Queensland Police. New trainees included nine people who were training to be fingerprint experts. Five of these trainees had completed a 5-week training program on the day of testing, and four had been working in a fingerprint department for 5 or 6 months.

Intermediate trainees included eight people who were training to be fingerprint experts. Of these, one had 1 year of experience, one had 2 years of experience, two had 3 years of experience, one had 4 years of experience, and three had 5 years of experience (M = 3.5, SD = 1.51). The distinction between the two types of trainees is arbitrary and was decided before the data were analyzed. Experts were 37 qualified court-practicing fingerprint experts with experience ranging from 5 to 32 years (M = 17.45, SD = 7.53).

Procedure

Participants were presented with pairs of prints displayed side-by-side on a computer screen, as illustrated in Figure 2. They were asked to judge whether the prints in each pair matched, using a confidence rating scale ranging from 1 (*sure different*) to 12 (*sure same*). Judgments were reported by moving a scroll bar to the left ("different") or right ("same"). The scale forced a "match" or "no match" decision, where ratings of 1 through 6 indicated *no match*, and ratings of 7 through 12 indicated a *match*. Judgments of "inconclusive," which are often made in practice, were not permitted in this two-alternative forced-choice design, so it was possible to distinguish between accuracy and response bias (Green & Swets, 1966). A thorough explanation of the advantages of this approach can be found in Thompson et al. (2013). The methodology of this experiment emulates one aspect of the identification

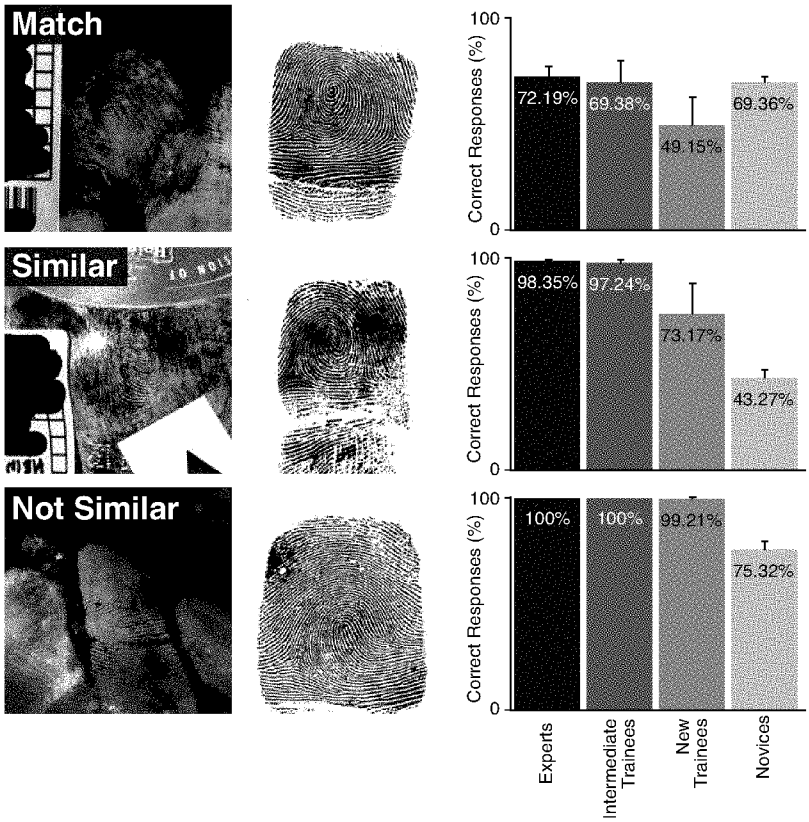

Figure 2. Stimuli and mean percentage of correct responses. On each trial, participants were presented with a genuine crime scene latent print on the left and a fully rolled candidate print on the right, and they were asked to judge whether the prints in each pair matched using a confidence rating scale. On some trials, the two prints came from the same individual (top row); on others, the prints were similar but came from two different individuals (middle row); and on others, the prints came from two different individuals and were paired randomly (bottom row). The three graphs on the right depict the mean percentage of correct responses in these three conditions for experts, intermediate trainees, new trainees, and novices. Error bars represent 95% within-cell confidence intervals.

process, namely, the extent to which a print can be accurately matched to its source.

Stimuli

The stimuli consisted of 45 latent prints from a larger police training examination set and were paired with fully rolled prints. The latent prints were taken from actual crime scene casework and were used for training purposes. An examiner (the third author) developed the training set to provide comparison materials that would expose trainee experts with a larger volume of latent comparisons, and chose the experimental stimuli from the larger set such that they provided clear ridges for the NAFIS system to search on. The corresponding fully rolled matches were declared previously as identifications, and were verified by at least three expert examiners. Information about whether these identifications had any associated, and potentially corroborating, information such as a guilty plea, conviction, or independent DNA match was not available.

Given that the prints were matched during casework, a qualified expert must have decided that each matching fingerprint pair contained sufficient information to make an identification. Across participants, each latent print was paired with a fully rolled print from the "same" individual (match), with a nonmatching but similar exemplar (similar distractor), and with a random nonmatching exemplar (nonsimilar distractor). For each participant, each latent print was randomly allocated to one of the three trial types, with the constraint that there were 15 prints in each condition. Unlike the simulated latent prints taken from the Forensic Informatics Biometric Repository (as used by Tangen et al., 2011), the ground-truth of the matches cannot be certain.

Similar distractors were obtained by searching each crime scene latent print on the Australian National Automated Fingerprint Identification System. For each latent print, the most highly ranked, nonmatching exemplar from the search was used if it was available in the Queensland Police 10-print hardcopy archives, which contains approximately 1 million 10-print cards (10 million individual prints) from approximately 300,000 to 400,000 people (one person may have more than one 10-print card on record). The corresponding 10-print card was retrieved from the archives, scanned, and extracted. In practice, highly similar nonmatches

retrieved from large national databases are likely to increase the chance of incorrect identifications (Dror & Mnookin, 2010). Distinguishing such highly similar, but nonmatching, print pairs from actual matching print pairs is potentially the most difficult task that fingerprint examiners face (Dror & Mnookin, 2010; Thompson et al., 2013).

Latent prints were from printed photographs and were scanned in color as a 600-dpi lossless Tagged Information File Format (TIFF) file, converted to grayscale, cropped to 600 × 600 pixels, and isolated in the frame. The matching and nonmatching exemplars were fully rolled fingerprint impressions made using a standard elimination pad and a 10-print card or were digitally scanned via LiveScan™. Each card was photocopied at 600-dpi and scanned in color as a 600-dpi lossless Tagged Information File Format (TIFF) file. Each print was then converted to grayscale, cropped to 600 × 600 pixels, and isolated in the frame.

Results

While analyzing the data, the pattern of results at the level of the trial type suggested that one of the latent prints in the set might not truly match the target exemplar. We sent the target pair to a qualified fingerprint examiner who declared that the prints, in fact, did not match. The source of the error arose from the police training materials spreadsheet that incorrectly labeled the finger type of a 10-print card, and so the incorrect print was extracted from the 10-print card. This transcription error has no relation to casework. As a result, all trials containing the latent (even the unaffected similar and nonsimilar distractor trials) were removed from the analysis. Because the latents were randomly allocated to either a target, similar or nonsimilar distractor pair, the proportion of trials removed was randomly distributed across trial types. With the offending latent removed there were 44 fingerprint comparison data points from each participant, rather than 45. For the 37 experts, for example, there were 537 matching trials, 547 similar nonmatching trials, and 544 nonsimilar nonmatching trials, rather than 555 per condition.

For each participant, we calculated the percentage of trials that were responded to correctly in each condition. The three graphs on the right side of Figure 2 depict the average percentage of correct responses for the 37 experts, 10 intermediate trainees, nine new trainees, and 37 novices. Participants were anonymous, so it is not possible to link a participant's performance to a particular person or police agency.

Matching trials included pairs of prints that originated from the same source. We use the term match here as shorthand, but as indicated above, the ground truth of the print pairs is uncertain. As depicted in Figure 2, experts correctly labeled 72.19% (SD = 18.10%) of the matching pairs on average "match" (hits), but incorrectly labeled 27.81% of the matching pairs "no match" (misses). Intermediate trainees correctly labeled 69.38% (SD = 17.34%) of the matching pairs "match" (hits), but incorrectly labeled 30.62% of them "no match" (misses). New trainees correctly labeled 49.15% (SD = 21.66%) of these matching pairs "match" (hits), but incorrectly labeled 50.85% "no match" (misses). Novices correctly labeled 69.36% (SD = 13.02%) of these matching pairs "match" (hits), but incorrectly labeled 30.64% "no match" (misses).

Highly similar nonmatching trials included pairs of prints that did not originate from the same source but are, according to the national database search algorithm, highly similar. We use the term *similar* nonmatch here as shorthand. Experts correctly labeled 98.35% (SD = 4.01%) of the highly similar nonmatching pairs "no match" (correct rejections), but incorrectly reported 1.65% of them "match" (false alarms) on average. Specifically, seven experts incorrectly labeled nine pairs out of the 547 highly similar nonmatching pairs a "match"—six experts made one false alarm each and one expert made three false alarms. Confidence ratings for the nine false alarms were 8, 9, 9, 10, 12, and 12 for the six experts with one false alarm each and 8, 9, 9, for the one expert with three false alarms. The nine false alarm errors occurred on eight different print pairs (i.e., each false alarm was made on a different latent and similar exemplar pair except for one). The fact that the nine false alarms were spread across prints and across people suggests that human factors are likely to be good predictors of these errors, rather than factors in the prints themselves. We caution readers, however, to avoid over interpreting individual confidence reports because the appropriate level of analysis in this experiment is expertise, not the individual confidence ratings by individual examiners on individual trials. Intermediate trainees correctly labeled 97.24% (SD = 4.91%) of these pairs "no match" (correct rejections), but incorrectly labeled 2.76% "match" (false alarms). Specifically, four intermediate trainees made four false alarms on four different print pairs. New trainees correctly labeled 73.17% (SD = 23.22%) of these pairs "no match" (correct rejections), but incorrectly labeled 26.83% "match" (false alarms). Novices correctly labeled 43.27% (SD = 14.79%) of these pairs "no match" (correct rejections), but incorrectly labeled 56.73% "match" (false alarms). Most striking is the difference in the rate of false alarms between experts and novices: 1.65% for experts compared with 56.73% for novices.

Nonsimilar nonmatching trials included pairs of prints that did not originate from the same source and were sampled randomly from the set. We use the term nonmatch here as shorthand. Of the trials in which the prints did not match, and were not similar, both experts and intermediate trainees correctly labeled 100% of these pairs "no match" (correct rejections), and so they did not incorrectly label any pairs (false alarms). New trainees correctly labeled 99.21% (SD = 2.39%) of these pairs on average "nonmatch," but incorrectly labeled 0.79% "match." Novices correctly labeled 75.32% (SD = 15.48%) of these pairs on average "nonmatch," but incorrectly labeled 24.68% "match."

Experts and intermediate trainees responded much more toward the extreme ends of the confidence scale compared with new trainees and novices: 83% of expert and 75% of intermediate responses were either one or 12 compared with 53% for new trainees and 20% for novices. We subjected the percentages of correct responses to a 4 (expertise: experts, intermediate trainees, new trainees, novices) × 3 (trial type: match, similar nonmatch, nonsimilar nonmatch) mixed analysis of variance (ANOVA). The analysis revealed significant main effects of expertise, $F(3, 89) = 109.450$, $MSE = 0.014$, $p < .001$, $\eta^2 = .79$, 95% CI [.71, .82], and trial type, $F(2, 178) = 66.038$, $MSE = .019$, $p < .001$, $\eta^2 = .43$, 95% CI [.33, .50], and a significant interaction between the two, $F(6, 178) = 29.385$, $MSE = .019$, $p < .001$, $\eta^2 = .50$, 95% CI [.40, .55].

Simple effects analyses revealed a significant benefit of expertise on all trial types: match, $F(3, 89) = 4.759$, $MSE = .027$, $p = .004$, $\eta^2 = .14$, 95% CI [.03, .23], similar nonmatch, $F(3, 89) = 142.391$, $MSE = .015$, $p < .001$, $\eta^2 = .83$, 95% CI [.77, .86], and nonsimilar nonmatch, $F(3, 89) = 45.999$, $MSE = .010$, $p < .001$, $\eta^2 = .61$, 95% CI [.49, .67].

Follow-up pairwise comparisons revealed that, for matches, only new trainees were different from all other levels of expertise: new trainees versus novices, $p = .001$, $d = 1.13$, 95% CI [32.0, 8.0]; new trainees versus intermediate trainees, $p = .009$, $d = 1.03$, 95% CI [35.0, 5.1]; new trainees versus experts, $p < .001$, $d = 1.15$, 95% CI [35.0, 11.0]. For similar nonmatches, both novices and new trainees were different from all other levels of expertise: novices versus new trainees, $p < .001$, $d = 1.54$, 95% CI [38.8, 21.0]; novices versus intermediate trainees, $p < .001$, $d = 4.89$, 95% CI [62.5, 45.4]; novices versus experts, $p < .001$, $d = 5.08$, 95% CI [60.7, 49.5]; new trainees versus intermediate trainees, $p = .001$, $d = 1.43$, 95% CI [35.1, 13.0]; new trainees versus experts, $p < .001$, $d = 1.51$, 95% CI [34.1, 16.3]. For nonsimilar nonmatches, only novices were different from all other levels of expertise: novices versus new trainees, $p < .001$, $d = 2.16$, 95% CI [32.2, 16.6]; novices versus intermediate trainees, $p < .001$, $d = 2.25$, 95% CI [31.7, 17.7]; novices versus experts, $p = .001$, $d = 2.25$, 95% CI [29.2, 20.1].

Discussion

We set out to determine whether fingerprint experts are any more accurate at matching prints than trainees and lay people. We also wanted to get an idea of how often these groups make "misses" (i.e., errors comparable with allowing a guilty person to escape detection) compared with how often they make "false alarms" (i.e., errors comparable with falsely incriminating an innocent person). In this experiment, we made use of genuine crime scene prints that are highly representative of casework, but where the ground truth is uncertain.

We found that experts and novices were equally accurate at identifying print pairs that actually matched; both groups were around 70% accurate. Experts, however, were much more accurate than novices at identifying prints that did not actually match, but were highly similar; experts were 98.35% accurate compared with 43.27% for novices. It seems that superior expert performance lies, not in the ability to match prints *per se,* but in the ability to identify highly similar, but nonmatching, prints as such. The comparison with novices is important for demonstrating expertise, and shows that the discrimination task was difficult enough for experts to perform accurately, but for novices to perform relatively poorly.

The results of this experiment are similar to those reported by Tangen et al. (2011). It is possible to compare performance across these two experiments because it is only the stimulus sets that differ—simulated crime scene latents were used by Tangen et al. and genuine crime scene latents were used in this experiment. The performance difference between experts and novices for similar nonmatches was about the same in both experiments; experts were around 55% more accurate than novices in both experiments. The performance difference between experts and novices for matches, however, was different in the two experiments. In Tangen et al., experts were around 18% more accurate than novices for prints that matched. In the present experiment, experts and novices were equally accurate for prints that matched. It appears that experts are even more conservative in their decision making when dealing with genuine crime scene prints than when dealing with the simulated crime scene prints: 72% hits on matching genuine latents compared with 92% hits on matching simulated latents.

The performance of trainees, in the present experiment, was more nuanced. For print pairs that matched, intermediates were just as accurate as experts, although new trainees were less accurate than any other group. For the similar print pairs, experts and intermediates were equally accurate, although new trainees were less accurate than both experts and intermediates, but they were more accurate than novices.

There was very little difference between the overall performance of experts with an average of 17.5 years of experience and intermediate trainees with an average of 3.5 years of experience. Although these results provide some insight into the development of expertise (i.e., how long it takes to turn a novice into an expert), much more research needs to be conducted. For example, one could track the development of novice examiners over time to determine precisely what aspects of their performance change as novices become experts and how quickly these capabilities develop. Future experiments could also pinpoint the nature of this expertise. That is, the relative contribution of formal rules compared with the accumulation of experience (Norman & Brooks, 1997; Norman, Young, & Brooks, 2007), as well as the role of corrective feedback (Eva & Regehr, 2005). In addition to the development of expertise, we need to better understand how age affects identification separately from years of experience. In medicine, for example, older/more experienced doctors generally have greater diagnostic accuracy (Eva, 2002), but are less likely to be influenced by the presentation of clinical features that are inconsistent with their initial hypothesis (Eva, Link, Lutfey, & McKinlay, 2010). Similar analyses need to be conducted in forensic reasoning to establish the relationship between age, experience, and fingerprint matching performance.

This experiment was not designed to determine the likelihood of errors in practice, nor the performance of individuals or departments, and examiners were not provided with their usual tools or independent verification. It was designed to determine performance differences based on expertise using genuine crime scene latents. Inferring from these results that experts are 98.35% accurate in practice or that the overall error rate of fingerprint identification is 1.65%, would be unjustified. It may be necessary, or the courts may demand, that particular rates of error are established for particular situations. At the extreme, an examiner could report how accurate they are at matching an arch type print, lifted from glass, using white powder, in a particular department, with particular training, on a Tuesday, and so on. But unless it has been demonstrated that accuracy (or proficiency, or reliability, or competence) varies systematically in any one of those situations, then it may be best to report measures of accuracy at a broader level (Koehler, 2008; Thompson et al., 2013).

Discrimination and Response Bias

In describing how well someone performs a given task, people usually count the number of correct items relative to the total number of items in the task. But in our experiment, when an

examiner compares two fingerprints, there are two ways to be right and two ways to be wrong. To get a comparison right, as shown in Figure 2, one can correctly say the prints are from the same source when they actually are (a hit), or correctly say the prints are not from the same source when they actually are not (a correct rejection). To get a comparison wrong, one can incorrectly say that the prints are from the same source when they actually are not (a false alarm), or incorrectly say the prints are not from the same source when they actually are (a miss). Therefore, simply counting up, say, just the number of similar, but nonmatching prints that experts correctly judged to be "no match" (i.e., 98.35%) is only half of the story. These examiners could have scored 100% correct on these nonmatching prints by simply saying "no match" to every pair of prints. By adopting such a conservative response bias, however, they would have incorrectly deemed every pair of matching prints a "no match" as well. On the other hand, they could adopt an extremely liberal response bias and say "match" to every pair of prints. These examiners would score 100% correct for all the prints that actually match, but they would incorrectly declare every nonmatching pair a "match" as well. The only way to perform perfectly in this experiment is to adopt a neutral response bias and correctly label all of the matching pairs "match" and label all of the nonmatching pairs "no match."

By adopting a signal detection methodology, we can distinguish people's tendency to say "match" or "no match" from their ability to distinguish prints that actually match from those that actually do not match—we can separate an examiner's response bias from their ability to discriminate matching and nonmatching fingerprints (Green & Swets, 1966; Phillips, Saks, & Peterson, 2001). The results from our experiment indicate that experts and intermediate trainees both have a tendency to say "no match" regardless of whether the prints actually match or not. Adopting such a strong conservative response bias certainly reduces the rate of false alarm errors (i.e., errors that could lead to falsely incriminating an innocent person), but it will also necessarily increase the rate of miss errors (i.e., errors that could lead to a guilty person escaping detection). A false alarm rate of 1%–3% is indeed impressive, particularly compared with a 57% false alarm rate for novices. But there is a direct tradeoff between preventing a false alarm and allowing a miss. The cost of such a low false alarm rate for experts and intermediate trainees in the current experiment equates to a substantial miss rate of roughly 30%.

The relationship between discrimination and response bias for each of the four groups in the above experiment, and the novices and experts in Tangen et al. (2011), is depicted in Figure 3. The figure represents the space of all possible results from experiments like ours. Each of the tables that comprise Figure 3 is a version of the contingency table in Figure 2 with different combinations of average "match" and "no match" responses from participants when we ask them to compare 50 print pairs that actually match and 50 prints pairs that don't actually match. Moving along the *y*-axis from the bottom to the top of the figure, participants become more capable of discriminating matching and nonmatching prints. That is, they correctly say "match" to matching prints and "no match" to nonmatching prints, thereby increasing the values in the top left cell (hits) and bottom right cell (correct rejections) in each of the tables that comprise Figure 3. The table at the apex depicts perfect discrimination—50 hits and 50 correct rejections. Participants here can distinguish between matching and nonmatching prints perfectly. The tables along the bottom depict chance discrimination. Participants here cannot distinguish between matching and nonmatching prints (their performance is like a coin flip), but there are several ways to reach the same level of overall performance. Overall performance—the number of comparisons they got correct—is depicted by the large number in bold at the center of each table and is the sum of the two diagonal cells (hits and correct rejections). Overall performance ranges from 50 (chance discrimination) at the bottom of the figure to 100 (perfect discrimination) at the top. Results from novices and new trainees lie toward the bottom of this figure—they are reasonably poor discriminators—compared with intermediate trainees and experts, who are closer to the top.

Moving along the *x*-axis from the left to the right, participants become more conservative in their responses; on the left side of Figure 3, they say "match" much more often than they say "no match," regardless of whether the prints actually match or not. The opposite is true on the right side of the figure; they say "no match" much more often than they say "match." A liberal response bias (on the left of the figure) is represented by a higher column total for the two cells on the left side of each table (i.e., a tendency to say "match"), compared with the two cells on the right. A conservative response bias (on the right of the figure) is represented by a higher column total for the two cells on the right side of each table (i.e., a tendency to say "no match"), compared with the two cells on the left.

An extremely liberal response bias, coupled with low accuracy, means that participants say "match" to every comparison. They will get half of the comparisons correct in this case, but they will also get half incorrect; they make many hits and many false alarms, while not making any misses or correct rejections. An extremely conservative response bias, on the other hand, coupled with low accuracy, means that participants say "no match" to every comparison. Again, they will get half of the comparisons correct in this case, but they will also get half incorrect; they make many misses and many correct rejections, while not making any hits or false alarms. Results from novices lie closer to the left of the figure—they have a reasonably liberal response bias—compared with trainees and experts, who lie closer to the right and have a very conservative response bias.

Theoretically, there is an optimal decision criterion, that minimizes errors, where the participant shows no response bias and is equally likely to say "match" or "no match" across all comparisons (i.e., straight up and down the middle of Figure 3 where the row totals are equal). This is true only when the base rates—the signal and noise distributions—are equal (i.e., the column totals are equal), as in Figure 3. The picture changes dramatically, however, when the base rates are unequal, which would add a third dimension to Figure 3. For example, if there are many more matches than there are nonmatches—as may be the case in practice when examiners compare crime scene latent prints to suspects already in custody—a liberal response bias would result in a high number of hits and a low number of false alarms. If, on the other hand, there are many more nonmatches than there are matches—as may be the case in practice when examiners search large databases in the absence of a suspect—a liberal response bias would result in a high number of false alarms and a low number of hits. The decision criterion (i.e., the propensity to say "match" or "no match") that a search algorithm, examiner, or department adopts will depend on

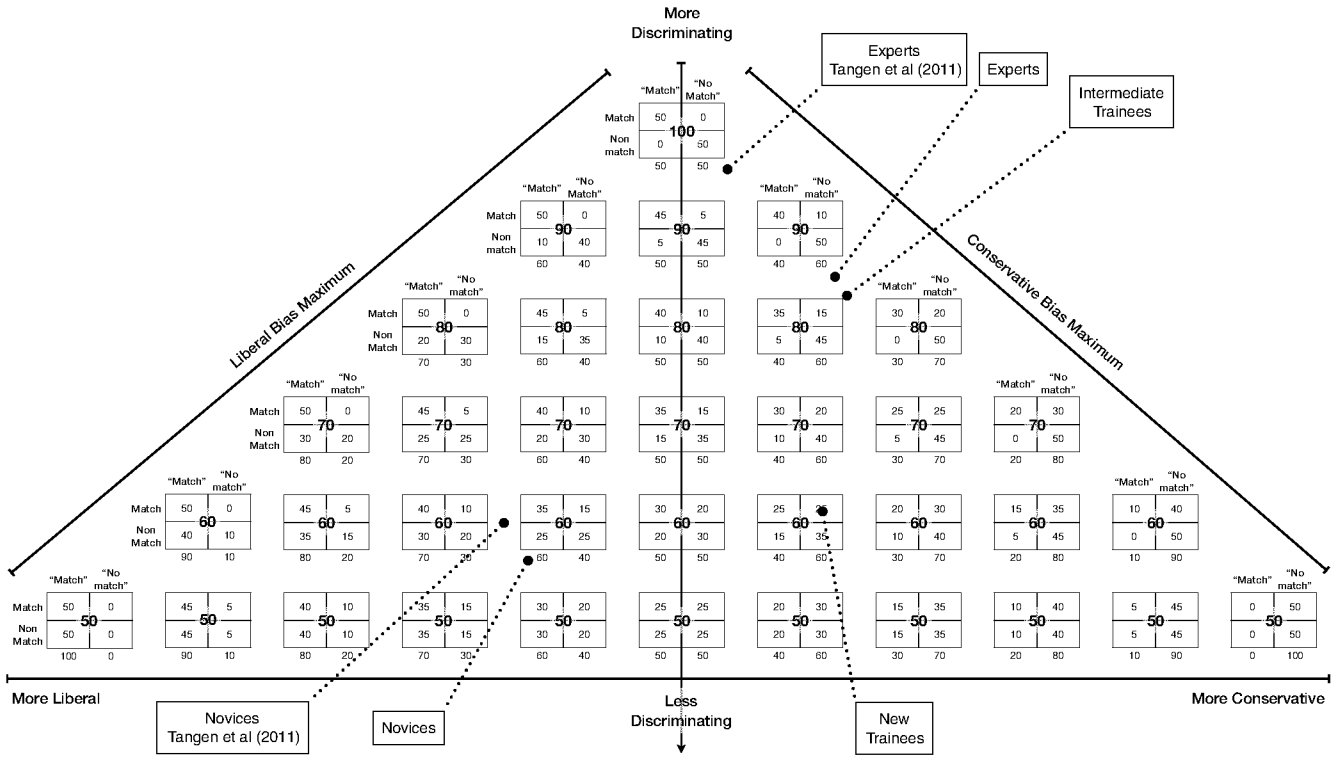

Figure 3. Signal detection space of all possible results. The space represents all possible performance results from a fingerprint discrimination task and the relationship between discrimination and response bias. Each of the tables that comprise the figure is a 2 × 2 contingency table depicting the four possible outcomes of a forced choice fingerprint discrimination task where two prints match or not and an examiner labels them a "match" or "no match." The numbers align with hits, false alarms, misses, and correct rejections in Figure 1. The large number in bold at the center of each table depicts the sum of the two diagonal cells ranging from 50 (chance discrimination) at the bottom of the figure to 100 (perfect discrimination) at the top. The column totals at the bottom of each table depict response bias with liberalism represented by a higher column total for the two cells on the left side of each table and conservatism represented by a higher column total for the two cells on the right side of each table. Pinpointed in the space are the locations of the actual results for each of the four groups in the current experiment and the novices and experts in Tangen et al. (2011), with nonsimilar distractors omitted and the number of trials scaled to give a total of 100.

the real world costs and benefits. Policy decisions about the ideal decision criterion and subsequent response bias will be ideological, not empirical, in nature. Interventions in training, technology, management, safety culture, and public policy will influence the signal and noise distributions, and the ratio of errors (false alarms vs. misses) that examiners will make in practice (Clark, 2012; Wixted & Mickes, 2012).

Conclusions

We found that qualified, court-practicing fingerprint experts were exceedingly accurate at discriminating prints compared with novices. Our experts showed a conservative response bias, tending to err on the side of caution by making more errors of the sort that could allow a guilty person to escape detection than errors of the sort that could falsely incriminate an innocent person. The performance difference between experts and novices provides further evidence for expertise in fingerprint identification. How novices would perform under different levels of motivation and incentive, after brief training, after the costs of a false alarm versus a miss are conveyed, and so forth, is still unknown. The superior performance of experts in this experiment was not simply a function of their ability to match prints, *per se*, but a result of their ability to identify highly similar, but nonmatching fingerprints as such. Novices, nonetheless, correctly identified almost the same number of matching prints as experts. This experiment was designed to be difficult. The fact that experts made so few errors is evidence for impressive human pattern matching performance possibly exceeding that of experts in other comparable domains of expertise (Thompson et al., 2013). When these results are compared with those of Tangen et al., (2011) we see that experts were even more conservative in their decision making when dealing with genuine crime scene prints than when dealing with simulated crime scene prints, and this conservatism made them relatively less accurate overall. How experts would perform with their usual tools, peer verification, statistical models, different lifting agents and surface types, different response types, time and resource constraints, different types of training, experience, and qualifications, and so forth, is unknown.

The performance of the trainee groups was surprising. First, intermediate trainees—despite their lack of qualification and average 3.5 years experience—performed about as accurately as qualified experts who had an average 17.5 years experience. This finding provides some insight into the development of expertise, that is, how long it takes to turn a novice into an expert. It appears that people can learn to distinguish matching from similar nonmatching prints to roughly the same level of accuracy as experts after a few years of experience and training. Much more research needs to be conducted, however, to make precise and definitive conclusions about the factors that lead to fingerprint matching expertise. Second, new trainees—despite their 5-week, full-time training course or their 6 months of experience—were not any better than novices at discriminating matching and similar nonmatching prints, they were just more conservative (see Figure 3). It appears that early training and/or experience may not necessarily result in more accurate judgments, but may simply result in a more conservative response bias (i.e., a tendency to say "no match" more often).

This experiment was limited by the small number of trainee participants, so one needs to be cautious when interpreting the relative performance of these groups. Small sample sizes in trainee comparison groups will be difficult to overcome considering that we recruited the majority that existed across Australia. Also, more trials (i.e., the number of fingerprint comparisons) per participant across conditions would help bear out false alarm errors in order to understand their nature. Given that appropriate casework stimuli are so rare and manufacturing stimuli is difficult and expensive, future experiments could attempt to pull expert performance off ceiling by adding artificial noise or constraining the task environment.

Measuring the relative performance of trainees is a useful first step, but programmatic or longitudinal experiments are needed to answer questions such as: What sets an expert apart from a novice? How does fingerprint expertise develop over time? Does training help and can training time be reduced without compromising performance? What is the best way to provide feedback to examiners about their performance? More research is needed to determine the nature of forensic reasoning, the influence of deadline pressure (Brewer, Weber, Wootton, & Lindsay, 2012), the role of feedback and self-assessment (Eva & Regehr, 2005), and, more generally, the respective contribution of training (the formal rules) versus daily exposure to a multitude of prints (the accumulation of instances; Norman et al., 2007).

The findings of this representative, lab-based experiment may have implications for the way fingerprint examiners testify in court (Edmond, Thompson, & Tangen, in press). What the findings mean for understanding expert performance in the wild is an open, empirical, and epistemological question that is part of an ongoing conversation (e.g., Koehler, 2008, 2012; Mnookin, 2008; Thompson et al., 2013). Whether performance data come from lab-based experiments, statistical models, proficiency tests, or full-scale black box interrogations of a system, one still needs to make an inference to performance in a particular manifestation of practice or to reason about the value of the evidence in a particular case. Edmond, Thompson, & Tangen (in press) have proposed a guide for the reporting of emerging empirical data about the performance of fingerprint examiners in order to help nonexperts understand the value of fingerprint evidence. The question, "What is the error rate in practice?" may not be the right one. Better questions might be: What is known about expert performance in situations similar to practice or to the particular case? What can reasonably be inferred from the general (data from experiments like this one) to the particular (the evidence in the case)? What information, and in what form, will help a trier of fact make optimal decisions? Taking the lead from research in health care and aviation, having empirical evidence from several experiments that address various research questions, at multiple levels of analysis, is sure to be the best way to help researchers reason about performance in the wild and to help triers of fact reason about forensic evidence.

References

Brewer, N., Weber, N., Wootton, D., & Lindsay, D. (2012). Identifying the bad guy in a lineup using confidence judgments under deadline pressure. *Psychological Science, 23,* 1208–1214. doi:10.1177/0956797612441217

Busey, T., & Dror, I. (2010). Special abilities and vulnerabilities in forensic expertise. In A. McRoberts (Ed.), *The fingerprint sourcebook* (pp. 1–23). Washington, DC: National Institute of Justice Press.

Busey, T., & Parada, F. (2010). The nature of expertise in fingerprint examiners. *Psychonomic Bulletin & Review, 17,* 155–160. doi:10.3758/PBR.17.2.155

Busey, T., Yu, C., Wyatte, D., Vanderkolk, J., Parada, F., & Akavipat, R. (2011). Consistency and variability among latent print examiners as revealed by eye tracking methodologies. *Journal of Forensic Identification, 60,* 61–91.

Campbell, A. (2011). *The fingerprint inquiry report.* Edinburgh, UK: APS Group Scotland.

Champod, C., & Evett, I. (2001). A probabilistic approach to fingerprint evidence, *Journal of Forensic Identification, 51,* 101–122.

Clark, S. E. (2012). Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science, 7,* 238–259. doi:10.1177/1745691612439584

Cole, S. A. (2002). *Suspect identities: A history of fingerprinting and criminal identification.* Cambridge, MA: Harvard University Press.

Cole, S. A. (2005). More than zero: Accounting for error in latent fingerprint identification. *Journal of Criminal Law & Criminology, 95,* 985–1078.

Cole, S. A. (2010). Who speaks for science? A response to the National Academy of Sciences report on forensic science. *Law, Probability & Risk, 9,* 25–46. doi:10.1093/lpr/mgp032

Cole, S. A., Welling, M., Dioso-Villa, R., & Carpenter, R. (2008). Beyond the individuality of fingerprints: A measure of simulated computer latent print source attribution accuracy. *Law, Probability & Risk, 7,* 165–189. doi:10.1093/lpr/mgn004

Dror, I. E., & Cole, S. A. (2010). The vision in "blind" justice: Expert perception, judgment, and visual cognition in forensic pattern recognition. *Psychonomic Bulletin & Review, 17,* 161–167. doi:10.3758/PBR.17.2.161

Dror, I. E., & Mnookin, J. L. (2010). The use of technology in human expert domains: Challenges and risks arising from the use of automated fingerprint identification systems in forensic science. *Law, Probability & Risk, 9,* 47–67. doi:10.1093/lpr/mgp031

Dror, I. E., & Rosenthal, R. (2008). Meta-analytically quantifying the reliability and biasability of forensic experts. *Journal of Forensic Sciences, 53,* 900–903. doi:10.1111/j.1556-4029.2008.00762.x

Dror, I. E., Wertheim, K., Fraser-Mackenzie, P., & Walajtys, J. (2012). The impact of human-technology cooperation and distributed cognition in forensic science: Biasing effects of AFIS contextual information on human experts. *Journal of Forensic Sciences, 57,* 343–352. doi:10.1111/j.1556-4029.2011.02013.x

Edmond, G., Thompson, M. B., & Tangen, J. M. (in press). A guide to interpreting forensic testimony: Scientific approaches to fingerprint evidence. Law, Probability & Risk.

Edwards, H. T. (2009). *Statement of the Honorable Harry T. Edwards: Strengthening forensic science in the United States: A path forward.* Washington, DC: United States Senate Committee on the Judiciary.

Eva, K. W. (2002). The aging physician: Changes in cognitive processing and their impact on medical practice. *Academic Medicine, 77,* S1–S6. doi:10.1097/00001888-200210001-00002

Eva, K. W., Link, C. L., Lutfey, K. E., & McKinlay, J. B. (2010). Swapping horses midstream: Factors related to physicians' changing their minds about a diagnosis. *Academic Medicine, 85,* 1112–1117. doi:10.1097/ACM.0b013e3181e16103

Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: A reformulation and research agenda. *Academic Medicine, 80,* S46–S54. doi:10.1097/00001888-200510001-00015

Expert Working Group on Human Factors in Latent Print Analysis. (2012). *Latent print examination and human factors: Improving the practice through a systems approach.* Washington, DC: U.S. Government Printing Office.

Federal Bureau of Investigation. (1984). *The science of fingerprints: Classification and uses.* Washington, DC: U.S. Government Printing Office.

Garrett, R. (2009). *Memorandum from the President of the International Association for Identification, 02/19/09, International Association for Identification.*

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York, NY: Wiley.

Koehler, J. J. (2008). Fingerprint error rates and proficiency tests: What they are and why they matter. *Hastings Law Journal, 59,* 1077–1100.

Koehler, J. J. (2012). Proficiency tests to estimate error rates in the forensic sciences. *Law, Probability & Risk, 12,* 89–98. doi:10.1093/lpr/mgs013

Langenburg, G., Champod, C., & Wertheim, P. (2009). Testing for potential contextual bias effects during the verification stage of the ACE-V methodology when conducting fingerprint comparisons. *Journal of Forensic Sciences, 54,* 571–582. doi:10.1111/j.1556-4029.2009.01025.x

Leahy proposes landmark forensics reform legislation bill would improve reliability of forensic evidence. (2011, January 25) Retrieved from http://www.leahy.senate.gov/press/leahy-proposes-landmark-forensics-reform-legislation/

Maxmen, A. (2012, July 13). Proposed bill calls for better forensic science. *Nature News Blog.* Retrieved from http://blogs.nature.com/news/2012/07/proposed-bill-calls-for-better-forensic-science.html

Mnookin, J. L. (2008). Of black boxes, instruments, and experts: Testing the validity of forensic science. *Episteme, 5,* 343–358. doi:10.3366/E1742360008000440

National Research Council, Committee on Identifying the Needs of the Forensic Science Community. (2009). *Strengthening forensic science in the United States: A path forward.* Washington, DC: The National Academies Press.

Neumann, C. (2012). Fingerprints at the crime-scene: Statistically certain, or probable? *Significance, 9,* 21–25. doi:10.1111/j.1740-9713.2012.00539.x

Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A., & Bromage-Griffiths, A. (2007). Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *Journal of Forensic Sciences, 52,* 54–64. doi:10.1111/j.1556-4029.2006.00327.x

Norman, G. R., & Brooks, L. R. (1997). The non-analytical basis of clinical reasoning. *Advances in Health Sciences Education, 2,* 173–184. doi:10.1023/A:1009784330364

Norman, G., Young, M., & Brooks, L. (2007). Non-analytical models of clinical reasoning: The role of experience. *Medical Education, 41,* 1140–1145. doi:10.1111/j.1365-2923.2007.02914.x

Phillips, V. L., Saks, M. J., & Peterson, J. L. (2001). The application of signal detection theory to decision-making in forensic science. *Journal of Forensic Sciences, 46,* 294–308. doi:10.1520/JFS14962J

Risinger, D., Saks, M., Thompson, W., & Rosenthal, R. (2002). The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. *California Law Review, 90,* 1–56. doi:10.2307/3481305

Saks, M. J., & Faigman, D. L. (2008). Failed forensics: How forensic science lost its way and how it might yet find it. *Annual Review of Law and Social Science, 4,* 149–171. doi:10.1146/annurev.lawsocsci.4.110707.172303

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science, 309,* 892–895. doi:10.1126/science.1111565

Scientific Working Group On Friction Ridge Analysis Study And Technology. (2012, March 5). *Standard for the definition and measurement of rates of errors and non-consensus decisions in friction ridge examination (latent/tenprint), Ver. 1.2.* Retrieved from http://www.swgfast.org/documents/error/120305_Rates-of-Error_1.2.pdf

Tangen, J. M. (2013). Identification personified. *Australian Journal of Forensic Sciences.* doi:10.1080/00450618.2013.782339

Tangen, J. M., Thompson, M. B., & McCarthy, D. J. (2011). Identifying fingerprint expertise. *Psychological Science, 22,* 995–997. doi:10.1177/0956797611414729

Thompson, M. B., Tangen, J. M., & McCarthy, D. J. (2013). Expertise in fingerprint identification. *Journal of Forensic Sciences.* doi:10.1111/1556-4029.12203

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences of the United States of America, 108,* 7733–7738. doi:10.1073/pnas.1018707108

Vokey, J. R., Tangen, J. M., & Cole, S. A. (2009). On the preliminary psychophysics of fingerprint identification. *The Quarterly Journal of Experimental Psychology, 62,* 1023–1040. doi:10.1080/17470210802372987

Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science, 7,* 275–278. doi:10.1177/1745691612442906

Received November 29, 2012
Revision received May 27, 2013
Accepted May 29, 2013 ■