JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, November 2013, Vol. 58, No. 6
doi: 10.1111/1556-4029.12203
Available online at: onlinelibrary.wiley.com

PAPER

GENERAL

Matthew B. Thompson,[1,2] *B.Inf.Tech., B.Sc. (Hons); Jason M. Tangen,*[1] *Ph.D.; and Duncan J. McCarthy,*[3] *B.App.Sc.*

Expertise in Fingerprint Identification*

ABSTRACT: Although fingerprint experts have presented evidence in criminal courts for more than a century, there have been few scientific investigations of the *human* capacity to discriminate these patterns. A recent latent print matching experiment shows that qualified, court-practicing fingerprint experts are exceedingly accurate (and more conservative) compared with novices, but they do make errors. Here, a rationale for the design of this experiment is provided. We argue that fidelity, generalizability, and control must be balanced to answer important research questions; that the proficiency and competence of fingerprint examiners are best determined when experiments include highly similar print pairs, in a signal detection paradigm, where the ground truth is known; and that inferring from this experiment the statement "The error rate of fingerprint identification is 0.68%" would be unjustified. In closing, the ramifications of these findings for the future psychological study of forensic expertise and the implications for expert testimony and public policy are considered.

KEYWORDS: forensic science, fingerprints, decision making, expertise, testimony, judgment, law, policy

Maintaining high standards of forensic evidence is vital for an effective justice system and for ensuring that innocent people are not wrongfully accused. Although fingerprint experts have presented evidence in criminal courts for more than a century, there have been few scientific investigations of the human capacity to discriminate these patterns and impressions. Contrary to popular belief (and television shows like CSI), computers are not relied upon to match crime scene fingerprints. Instead, human fingerprint experts decide whether a print belongs to a suspect or not. These experts make thousands of fingerprint identifications, per day, to be used as evidence in courts of law. Until recently, it was unclear what role expertise plays or whether expertise is even necessary to conduct accurate fingerprint comparisons.

In 2011, we (Tangen, Thompson, and McCarthy) published the results of an experiment testing the accuracy and claimed expertise of fingerprint examiners (1). These results showed that qualified, court-practicing fingerprint experts are exceedingly accurate (and more conservative) compared with novices, but they do make errors. Here, the current state of fingerprint testimony, measures of accuracy, and the research culture in forensic science are discussed. A rationale for the "Identifying Fingerprint Expertise" (1) experimental design is provided, and the steps taken to balance fidelity, generalizability, and control; ensure validity and ground truth; create a signal detection framework with highly similar prints; establish expertise with a novice control group; and establish meaningful error rates are described. Given the brevity of the original research article, this rationale will provide context for interpreting the results for the benefit of researchers, forensic examiners, forensic managers, lawyers, and judges. In closing, the ramifications of the findings for the future study of forensic expertise and the implications for expert testimony and public policy are considered.

[1]School of Psychology, The University of Queensland, St Lucia, QLD 4072, Australia.
[2]Queensland Research Laboratory, National Information and Communications Technology Australia, St Lucia, QLD 4072, Australia.
[3]Forensic Services Branch, Queensland Police Service, 200 Roma St, Brisbane, QLD 4000, Australia.

*Supported by a Fulbright Scholarship to Thompson, and an Australian Research Council Linkage Grant to Tangen and McCarthy (see The Forensic Reasoning Project at ForensicReasoning.com).

Received 2 Dec. 2011; and in revised form 31 July 2012; accepted 11 Aug. 2012.

Fingerprint Expert Testimony

In 2009, the National Academy of Sciences (NAS) delivered a landmark report highlighting the absence of solid scientific methods and practices in the forensic science domain (2). Harry T. Edwards (a senior U.S. judge and co-chair 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 offered in court is valid and reliable [(3); see also (4–7)]. The NAS report (2) 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" (p. 8). They recommended that the U.S. Congress fund basic research to help the forensic community strengthen their field, develop valid and reliable measures of performance, and establish evidence-based standards for analyzing and reporting testimony.

Courts rely heavily on forensic evidence to convict the guilty and protect the innocent. The presentation of flawed forensic evidence has obvious implications for individual cases, but it also calls into question the integrity of the entire criminal justice system—innocent people may be wrongfully convicted and people may lose trust in the justice system (3). It is important, there-

fore, that the claims made by fingerprint examiners testifying in court are accurate, substantiated, and reasonable. Fingerprint examiners have claimed that fingerprint identification is infallible (8) and that there is a zero error rate for fingerprint comparisons (3,9). Several commentators, however, have suggested that the claims of individualization and a zero error rate are not supported by evidence and, moreover, are scientifically implausible (e.g., [2,10]). Past President of the International Association for Identification suggested that members not assert 100% infallibility (zero error rate) of fingerprint comparisons (11) and the Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST) has drafted a standard for defining, calculating, and reporting error rates (12).

These issues are made more complicated by the nature of the legal system. The adversarial approach to the submission of evidence in court is not well suited to establishing "scientific truth." According to Edwards (3), judges and lawyers generally lack the expertise necessary to evaluate forensic evidence scientifically; defense attorneys often lack the resources to challenge the evidence; judges make admissibility decisions without the benefit of judicial colleagues and time to research; and cases are seldom appealed on the basis of disputed forensic evidence. It may be unwise, therefore, to rely on the judicial system to address the challenges facing fingerprint expert testimony.

Proficiency Tests and Accuracy

Fingerprint proficiency tests are available—such as those provided by Collaborative Testing Services, Inc. (CTS)—where the goal is to measure the accuracy of participating laboratories as a unit. The results are sometimes reported as a function of the kind of source print, sometimes as a function of the kind of correct response, sometimes as a function of the kind of error, and sometimes as the proportion of examiners/laboratories producing various responses. Proficiency tests of fingerprint examiners and previous studies of examiners' performance have been heavily criticized for (among other things) failing to include large, counterbalanced samples of targets and distractors for which the ground truth is known (see [13–15]). A weakness of proficiency tests and previous experiments is that a particular crime scene (or "latent") print either forms part of a match comparison or part of a distractor comparison for every individual who takes the test—a particular latent never serves as part of a target trial for one examiner/laboratory and a distractor trial for another examiner/laboratory. The result is that even a single highly distinctive latent on a distractor trial can artificially improve discrimination by reducing false positives. Or a single highly distinctive latent in a match trial can artificially improve discrimination by increasing hits (15).

There is nothing inherently wrong with the proficiency tests, like those provided by CTS, if the goal is to measure examiners' performance on exactly the same set of items. It may be possible to narrow in on particular features that cause difficulty (e.g., a peculiar pattern type) or what prints a particular department has trouble with. Indeed, if CTS made their materials and all their results widely available, they may provide a useful tool to measure performance on specific items and for assessing reliability. But, the tests are insufficient for measuring accuracy. To make general claims—beyond those of specific prints at a specific level (e.g., accuracy with whorl patterns)—different (and randomized) sets of prints for each examiner are needed. Otherwise, information in the specific prints used for the test will influence performance, making it difficult to generalize the results (15). Proficiency tests have not adequately addressed the general issue of expert matching accuracy and are not designed to disentangle the factors that affect matching accuracy.

There is, however, a growing body of research on fingerprint matching. Researchers have investigated the effect of contextual bias on fingerprint examiners (e.g., [16–21]); some of the special abilities and vulnerabilities of fingerprint examiners (e.g., [22–25]); the effect of technology (e.g., [26,27]); statistical models of fingerprint identification (e.g., [28–31]); and the accuracy of fingerprint examiners' decisions (e.g., (32–36) but see [15,37,38]). But, despite its 100 year history, there have still been few peer-reviewed studies directly examining the extent to which experts can correctly match fingerprints to one another, how competent and proficient fingerprint experts are, how and on what basis examiners make their decisions, or the factors that affect matching accuracy and what is the effect of expertise. In this study, we focus our efforts on the claimed matching expertise of fingerprint examiners.

The Research Culture

There is little doubt, among critics and proponents alike, that fingerprint identification is a valuable tool for law enforcement. Fingerprint identification errors are unlikely to be made because of malicious actions—fingerprint experts do their best to provide accurate fingerprint evidence to the courts and uphold civil liberties. Indeed, in the wake of the Mayfield case of false identification, the FBI stated their intention to make certain they are employing the most effective means to ensure the integrity of their expert fingerprint examinations (39). But—unlike other areas of expertise where decisions are safety-critical, such as aviation and medicine—there is currently no culture of research in fingerprint identification (40). Steady advances in the fingerprint development process have been made, but the critical human decision-making element has been neglected. Examiners are eager to demonstrate their abilities and advance their field, but rarely receive the support and resources to do so. The Director of the FBI's Investigation Lab described the gap between basic research and its application in solving crimes as the "valley of death" because "nobody wants to pay for it, nobody really wants to do it" (41).

It appears that fingerprint examiners are expected to strengthen the scientific basis of their field, while they relentlessly make identifications, search databases, and testify in court. Examiners, however, do not have the time, infrastructure, training, expertise, or research culture necessary to mount studies of human performance to ensure their field meets scientific and legal standards of evidence. By analogy, it would be like expecting the local doctor to find a cure for cancer. Clearly, examiners are not well positioned to address the challenges leveled at their field alone and, traditionally, there has not been a good working relationship between examiners and researchers. Research on expertise and complex systems is the domain of Cognitive Science and of Human Factors. These researchers have the reward structures already in place for conducting and publishing high-quality research and are well positioned to work with examiners to strengthen the field.

Much of the existing research on the cognitive factors involved in fingerprint judgment has investigated the influence of contextual information on examiners' performance. Dror et al. (17) used a highly publicized case of exposed fingerprint error to determine whether biasing information could lead an examiner to change their prior judgment. They covertly evaluated five

examiners, with an average of 17 years of experience, who consented to being tested at an unknown time over 12 months. The five examiners were each given a print to identify by a colleague, who advised them that the fingerprints were from a famous case of misidentification by the FBI for the 2004 Madrid train bombings. One examiner reported that the prints matched, three reported that the prints did not match, and one reported inconclusive. Unbeknownst to the examiners, however, the prints that they were asked to identify were taken from their own previous case history where they had previously declared them a match. With four of the five examiners subsequently changing their previous judgment of the prints as matching, it seems that top-down, contextual influences can affect expert judgments (see also [16,19]). Dror and Rosenthal (20) also conducted a meta-analysis to determine the degree to which examiners would make the same or conflicting decisions if extraneous information about the case was added. Although good data were sparse, the authors concluded that examiners are susceptible to bias.

It is clear that fingerprint experts have special abilities, but their decisions can be influenced by extraneous contextual information (22,42), and researchers have suggested ways contextual bias can be mitigated (43). Even with this contribution, relatively little research on human fingerprint identification has been conducted by academics and professionals alike. The U.S. NAS (2) and others have called for the development of a research culture within forensic science. Mnookin et al. (40) argue that there is a legitimate role for experience-based claims of knowledge, but also that pattern identification disciplines must develop a scientific foundation, through research, that is grounded in the values of empiricism and skepticism. The experiment described below is a step toward addressing the call from the NAS for the urgent development of objective measures of accuracy and expertise in fingerprint identification.

The "Identifying Fingerprint Expertise" Experiment

The "Identifying Fingerprint Expertise" experiment (1) was designed to find out whether fingerprint experts were any more accurate at matching prints than the average person and to get an idea of how often they make errors of the sort that could lead to a failure to identify a criminal compared with how often they make errors of the sort that could lead to inaccurate evidence being presented in court. Thirty-seven qualified fingerprint experts and 37 undergraduate students were given pairs of fingerprints to examine and decide whether a simulated crime scene print matched a potential "suspect" or not. Some of the print pairs matched, while others were highly similar but did not match.

Thirty-six simulated latent 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 matches were from our Forensic Informatics Biometric Repository, so, unlike genuine crime scene prints, they had a known true origin (13,44). All fingerprints were authentic (i.e., not simulated), but the latents were "simulated" in the sense of representing those found at crime scenes during casework. (For human matching performance with genuine crime scene prints, see [45].) Similar distractors were obtained by searching the Australian National Automated Fingerprint Identification System (NAFIS). 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 hard-copy archives, which contains approximately one million 10-print cards (10 million individual prints) from approximately 300,000–400,000 people (one person may have more than one 10-print card on record).

The results were striking. 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)—impressive expert performance, considering the corresponding false alarm rate for novices was 55.18%. We concluded that qualified court-practicing fingerprint experts are exceedingly accurate compared with novices, but are not infallible. 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. Even so, they made the kind of error that could result in incorrect evidence being presented to the court in a criminal trial.

Balancing Fidelity, Generalizability, and Control

Readers will react differently to the methodology employed in the "Identifying Fingerprint Expertise" experiment. The intuition that the best experiment should resemble "real-life" as much as possible is understandable. However, this intuition is incorrect (see [46] for a defense of external *in*validity), as has been demonstrated in other complex, safety-critical domains, such as aviation and medicine. When designing studies of human performance, the challenge is to find the appropriate balance between fidelity, generalizability, and control, to produce data that are best suited to answering the research question (47,48).

Fidelity is the similarity of an experimental task to its reference domain (i.e., fingerprint identification; [49,50]). How well does the task represent the particular work domain? Is the expertise of the participants high or low? Are the experimental situations full-featured or simplified? Are the available tools restricted or complete?

Generalizability is the theoretical depth or breadth of applicability of the results to situations beyond those examined in the study. Can the results and conclusions of the experiment be extended to situations that are different from the experiment, and is this the goal?

Control is the latitude available to the experimenters to isolate and manipulate variables. Is control high or low, and will the data collected be sensitive enough to detect differences between the experimental manipulations?

The perfect, but unattainable, experiment, will have high fidelity, high control, and high generalizability. But, these variables must be balanced to answer the research question appropriately (51,52). To understand accuracy rates in fingerprint identification, it is tempting to think that the best option is to insert test prints—unbeknownst to the examiners—into their regular workflow and measure the number of errors that come out the other end. This arrangement of high fidelity comes at the cost of reduced generalizability (we cannot apply the results from one experiment in a particular laboratory to fingerprint identification more broadly) and reduced control (when errors occur, we have no way of knowing how they arose and, therefore, what we might do to prevent them).

Our goal in measuring fingerprint expertise was to compare expert and novice performance at matching fingerprints and to get an idea of how often they fail to declare matching prints as

such (misses) in a matching task compared with how often they declare that prints match when they actually do not (false alarms). In designing the experiment, fidelity, generalizability, and control were balanced to answer these questions. The goal was not to generalize the results from these laboratory-based experiments to the "real world" (46). The fingerprint examiners who participated in this experiment did not have their usual tools available to allow them to zoom, rotate, or apply filters to the images; the latent prints that were used were collected as part of the Forensic Informatics Biometric Repository; and examiners conducted the experiment during their break on laptops that were provided in bureau conference rooms. This situation is not—purposefully—analogous to casework. The majority of expert participants, nonetheless, reported that the task represented their day-to-day work.

Unlike opinion polls and surveys, scientific experiments are not about seeing how well a sample approximates the general population from which it was selected. Laboratory-based experiments are intentionally artificial, because they allow for the control of all factors that are not of interest (e.g., the benefit of software tools, the role of verification, the type of crime, inconclusive responses, lifting agents, etc.) and for the systematic manipulation of only the factors of interest (e.g., the difference between novices and qualified experts, comparing performance on match trials and nonmatch trials, ensuring the ground truth of the prints, using highly similar distractors from a national database search, etc.). Generalizability in this context refers to the extent to which the difference between expert and novice performance is "real," not the extent to which the laboratory setting resembles the everyday operations of a fingerprint bureau.

The unit of analysis in this particular experiment is the comparison between matching and nonmatching prints or between experts and novices, not their absolute performance. So, even though a false alarm rate for experts of 0.68% is impressive in its own right, we cannot determine from this experiment whether this rate reflects the false alarm rate of the field more generally. We can conclude, however, that a false alarm rate of 55.18% for novices pales in comparison with experts under the same conditions. These results provide sufficient evidence for examiners to legitimately claim specialized knowledge, which may satisfy legal admissibility criteria. These results do not allow one to conclude that 0.68% is the misidentification rate for the field. A full scale, end-to-end "black box" experiment would allow us to pinpoint the precise rate of accuracy in a particular manifestation of practice, but it is inappropriate and inefficient to conduct a black box experiment to make simple claims like, "Are experts better than novices?" and "Do experts make errors?" Much more will be said about black box experiments in the *Determining Error Rates* section below. Put simply, the design of an experiment needs to be targeted specifically at the question that one sets out to address.

Validity and Ground Truth

Validity is a cornerstone of the scientific method. It is a measure of whether a method, instrument, questionnaire, construct, etc., measures what it is supposed to measure. Validity can be demonstrated by comparing the outcomes of a method with the ground truth. So, to assess validity in fingerprint identification decision making, the conclusion of the identification process (i.e., match or nonmatch) can be compared with that which is known (i.e., the ground truth). For example, if the ground truth of a pair of prints is that they were left by two different individuals, but the examiner incorrectly declares that the prints match, then the examiner has made a "false alarm" type of error; if the ground truth of a pair of prints is that they were left by the same finger from the same individual, but the examiner concludes that the prints do not match, then the examiner has made a "miss" type of error. The same goes for the two ways the examiner can reach the correct conclusion: that the prints actually match and the examiner correctly declares them as such (a "hit") or that the prints do not actually match and the examiner correctly declares them a nonmatch (a "correct rejection").

Most tests of proficiency and studies of accuracy (with the exception of [32]) used print pairs from casework where the ground truth was uncertain (see [13–15,53]). In assessing validity—the degree to which a test measures what it is supposed to measure—in fingerprint identification decision making, experiments ought to use print pairs for which the ground truth is known.

To make use of ground truth prints in our expertise study, fingerprint pairs were sourced from the Forensic Informatics Biometric Repository—an open biometric repository that we created to increase the availability of high-quality forensic materials where the ground truth is known. Details on the Forensic Informatics Biometric Repository are available at www.FIB-R.com. Forensic Informatics Biometric Repository contains a range of crime related materials: fingerprints, palm prints, shoe prints, face photographs, handwriting samples, voice samples, and iris photographs. The materials are collected from participants using a standardized methodology and vary systematically in quality. The repository contains multiple types of materials converging on a single source, and the ground truth of the source is built into the system. Materials are also collected from participants over two sessions to approximate the natural variation that is commonly found in forensic evidence (e.g., changes in facial hair, clothes, and shoe decay). Participants are first-year undergraduates who participate in 1 h of data collection for course credit and who provide informed consent.

The fingerprint materials contained in Forensic Informatics Biometric Repository are 10-prints, palm prints, and latent prints. Ink is used to capture each fingerprint onto standard 10-print cards, rolled fully from nail-edge to nail-edge, as well as "slap impressions" (pressing, not rolling, the fingers on the card) and fully rolled palms. Latent prints are taken from common crime scene surfaces (determined in consultation with fingerprint examiners) including: gloss-painted timber, smooth metal, glass, and smooth plastic. Participants are instructed to interact with the surfaces by "pushing on the gloss-painted timber to open the door" or "safely grabbing the knife by the blade." By interacting with objects in this way, the aim is to approximate the variation in materials that are commonly found at actual crime scenes.

In our experiment, the latent prints used were mated with their matching fully rolled exemplar so that the ground truth of match trials was known. The use of ground truth print pairs means that we can compare participants' responses to reality.

Signal Detection

A signal detection framework was used to measure the matching performance of fingerprint examiners (see also [15,54]). Signal detection is a method of quantitating a person's ability to distinguish signal from noise. In fingerprint identification, for example, *signal* refers to print pairs that truly match and *noise* refers to print pairs that do not truly match. Signal detection was initially applied to radar operators who were trying to discriminate

friendly and enemy aircraft and has since been used to measure all areas of human performance. Several factors may affect a person's ability to distinguish signal from noise, such as experience, expectations, context, physiological, and psychological states. To conduct a signal detection analysis of novice and expert fingerprint matching performance, the two ways of being right and the two ways of being wrong were separated; performance on matching and nonmatching prints was compared; and accuracy and response bias were separated.

Separate the Two Ways of Being Right and the Two Ways of Being Wrong

When an examiner compares two fingerprints, there are two ways for her to be right and two ways to be wrong, as shown in Fig. 1. To get a comparison right, she can correctly say the prints match when they actually do (a hit) or she can correctly say they do not match when they do not (a correct rejection). These decisions could result in providing correct evidence to the court or help eliminate potential suspects. To get a comparison wrong, she can incorrectly say that the prints match when they do not (a false alarm), or she can incorrectly say that the prints do not match when in fact they do (a miss). These decisions could lead to providing incorrect evidence to the court or a failure to identify a criminal.

Compare Performance on Matching and Nonmatching Prints

To properly measure performance, examiners must compare both matching and nonmatching prints. As discussed in the section on *Proficiency Tests and Accuracy* above, most previous studies have included no or few distractors, making it impossible to measure the two ways of being right and the two ways of being wrong, leading to artificially inflated accuracy rates.

In this experiment, to avoid the problem of distractors, each latent print in the set formed part of a match, similar distractor, and nonsimilar distractor trial. (The reasoning for providing similar distractors is described in the *Similarity* section below.) This way, match trials can be directly compared with the same number of nonmatch trials in the other two conditions. For each participant, each latent print was randomly allocated to one of the three trial types, with the constraint that there were 12 prints in each condition. This way, each latent print has an equal chance to act in either a match, similar distractor or nonsimilar distractor trial, and so eliminating the possibility that a particularly easy/difficult/distinctive/high-quality latent print could artificially influence examiners' performance.

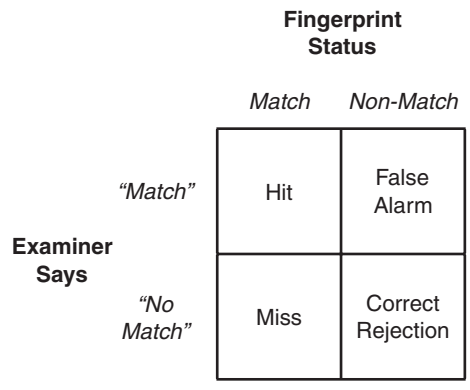

FIG. 1—*A 2 × 2 contingency table depicting the four possible outcomes of a forced choice fingerprint matching task where two prints match or not and an examiner declares them as a "match" or "no match."*

Separate Accuracy and Response Bias

There are two distinct measures of performance in a fingerprint comparison task. The obvious one is accuracy—an examiner's ability to discriminate matches from nonmatches. The less obvious measure is response bias—the decision rule employed by an examiner when they are uncertain about a comparison. That is, their tendency to say "match" or "no match" regardless of whether the prints match or not. If an examiner is unsure whether two prints match, and declares that they do, then they have made a "liberal" decision. If an examiner is unsure whether two prints match, and declares that they do not, then they have made a "conservative" decision. Averaged across several comparisons, the criteria used to make these decisions add up to so-called liberal and conservative response biases. Put simply, a response bias is a measure of a person's willingness to say, "yes": if they say "yes" a lot, then they have a liberal response bias; if they say "no" a lot, then they have a conservative response bias.

Two examiners can be equally accurate in their ability to discriminate or "see" matching prints, but—if they have a different response bias—they may come to opposite conclusions. It follows that there is no universal best response bias. There is an optimal criterion that minimizes false alarms and misses, but the appropriate decision criterion will depend on the costs and benefits of committing both types of error and both types of success. Only when the number of response alternatives is limited can an examiner's response bias be separated from their ability to discriminate prints.

Forcing a choice is a widely used paradigm for measuring human performance. In our experiment, participants were asked to judge whether print pairs matched, using a confidence rating scale ranging from 1 ("sure different") to 12 ("sure same") anchored at the center (i.e., 6.5). The response scale forced a "no match" or "match" decision because ratings of 1 through 6 indicated no match, whereas ratings of 7 through 12 indicated a match. (Note that these ratings were described incorrectly in the *Procedure* section of our original paper.) That is, subjects were required to move the scrollbar either left (to six or less, "different") or right (to seven or more, "same"); they could not make a rating of 6.5. This 12-point confidence scale was not designed to reflect the decisions made, and terms used, by examiners during casework. Judgments that the information was "inconclusive," which are often made in practice, were not permitted in this match/no match forced-choice design, making it possible to distinguish between accuracy and response bias (55). Interestingly, experts responded much more toward the extreme ends of the scale compared with novices: 92% of expert responses were either a 1 or a 12 compared with 32% for novices.

Aside from the capacity of the forced-choice procedure to differentiate the roles of accuracy and response bias, there are difficulties with measuring "inconclusive" judgments. There is no ground truth for sufficiency, that is, there is no way of knowing whether a print contains sufficient information for a human to discriminate it. The best that can be done is to ask several experts about the sufficiency of the information in several prints to see whether they agree with each other and themselves on repeated examinations (i.e., between and within participant reliability). What one means by "insufficient" is also tricky. Does it mean, "There is not sufficient information in the latent print to make an identification," or does it mean, "I am unwilling to make a

judgment (match/identification, no match/exclusion) either way." In fact, if a sufficient amount of information or signal was present (whatever that means), and an examiner declared it "inconclusive," then this ought to be regarded as a "miss" type of error.

Sufficiency of information in this experiment was partially controlled by only using prints that an expert declared as having sufficient information to make an identification. Participants' uncertainty in their judgments was also controlled using a 12-point confidence scale where a rating of 6, for example, would be counted as a "nonmatch" decision.

Similarity

A pair of fingerprints will appear similar or dissimilar to each other (or somewhere in-between), depending on the amount of information in each and depending on the experience of the examiner. There is no agreed upon definition or measure of similarity for the comparison of prints, but there have been attempts to create an objective measure of similarity. For example, Vokey et al. (15) converted a set of fingerprint images into their raw pixel values (i.e., the brightness values in each fingerprint image) and projected each print into the multidimensional space of all the prints in a set to return a vector, where the similarity of one print to another is given by the cosine of the angle between their vectors. A cosine value close to 1 indicates that the prints are virtually identical; whereas cosines close to zero indicate that the prints are highly dissimilar. This technique, therefore, provides an objective measure of similarity because it uses only the raw pixel values in the images and so requires no human input. We did not make use of this objective measure of accuracy for this experiment but, instead, used a national fingerprint database search.

For over 20 years, examiners have had the ability to search large databases, with the aid of computer algorithms, for potential matches to latent crime scene fingerprints. Although no formal data exist, it is likely that the majority of fingerprint comparisons made today use database queries (suspect-absent cases) rather than with a closed set of known prints from a suspect (suspect-present cases). A database query on a latent print will return a list of candidates that are most similar (according to the algorithm) to that latent. As Vokey et al. (15) and Dror and Mnookin (26) discuss, a database query makes an examiner's task much more difficult by returning a set of highly similar distractor prints—prints that look very much alike (according to the algorithm) but come from different people. These searches, by their very nature, are maximizing the conditions conducive to false positives. What's more, Vokey et al. (15) found that novices made more false alarms on comparisons that were similar—as measured by the distance between vectors of pixel maps—than those that were not similar. Given that distinguishing highly similar, but nonmatching, prints from genuine prints is likely to be the most difficult and common task that examiners face, similarity was included as a factor in our experiment.

Similar distractor prints were obtained by searching our simulated crime scene latents on the NAFIS. The latents were first auto-coded and then hand-coded by a qualified expert. For each simulated crime scene print, the most highly ranked nonmatching exemplar from the search was used if it was available in the Queensland Police 10-print hard-copy archives, which contains approximately one million 10-print cards (10 million individual prints) from approximately 300,000–400,000 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 the individual print of interest was extracted. Due to the proprietary nature of NAFIS, the information it uses in its search algorithms is unknown; but, it almost certainly relies on the minutiae, features, direction, relative and spatial relationships of the fingerprints as identified by human examiners, rather than low-level pixel values. These print comparisons were labeled "similar distractors," but it is important to note that the NAFIS algorithm is a search aid and was not designed to model human performance; so, what the NAFIS algorithm regards as similar may or may not correspond to what a human examiner considers similar.

Establishing Expertise: Novices as a Control Group

Research on the nature of human expertise has developed over decades and ranges from the seemingly disparate domains of chess to medical diagnosis. One goal of expertise research in cognitive science is to understand the mechanisms that account for the superior performance of experts, across domains (56). Another is understanding the domain-specific nature of expertise: Why does superior performance in one domain not transfer to others? Can we hasten the transition from novice to expert? Broadly, what makes an expert, an expert?

To study the nature of expertise in fingerprint identification, it first needs to be demonstrated that expertise actually exists; that is, are there people who possess exceptional abilities for matching latent fingerprints to their source? Despite over 100 years of fingerprint testimony—and although implicit in the terminology referring to fingerprint examiners who are qualified to testify in court (i.e., a "fingerprint expert")—there had been no study demonstrating that qualified examiners have specialized discrimination skills or abilities superior to those of the person on the street (but see [21] for a bias experiment with novices). Superior expert discrimination performance in fingerprint identification had been assumed, and the nature of that expertise had not been proposed or demonstrated. But should people who have no training or experience be expected to accurately match prints to their source?

It is clear from Vokey et al. (15) that novices generally have substantial abilities to match fingerprints. In a fingerprint matching task, naïve undergraduates were able to discriminate fingerprint matches from nonmatches quite well, or well above chance at least. With these findings in mind—and without any requisite experiments of expert performance in the forensic use of fingerprint identification—it was not obvious (to us at least) that experts would outperform novices in our experiment as much as they did. Furthermore, and as Vokey et al. (15) note, one pioneer of fingerprinting, Sir Francis Galton (57), believed that experts would quickly become unnecessary and that lay juries would eventually evaluate fingerprint evidence.

In 2002, Louis H. Pollak (a senior federal judge in Philadelphia; [58]) ruled, in *United States v. Llera Plaza*, that fingerprint evidence does not meet the standards set for scientific testimony and that experts in the field cannot testify that a suspect's prints definitely match those found at a crime scene. Pollak ruled that fingerprint experts could still point out the similarities between prints from a crime scene and those of a defendant, but the ultimate decision should be left to the jury. This decision was eventually overturned (59), but it is clear that one option for expert testimony under consideration is for experts to present the physical evidence, with commentary attached, and allow lay juries to decide whether a latent crime scene print matches the suspect.

Considering both the evidence for the reasonable performance of novices and the notion that juries should make the ultimate decision, it seems that the most appropriate comparison group to demonstrate expertise should be novices who have no training

with fingerprints whatsoever. In the "Identifying Fingerprint Expertise" experiment, the matching performance of qualified fingerprint examiners was compared with the performance of novice undergraduates who had no experience or training with prints to establish the supposition of expertise in fingerprint identification. Novices were 37 psychology undergraduates from The University of Queensland who participated for course credit. Experts were 37 qualified practicing fingerprint experts from five police organizations (the Australian Federal, New South Wales, Queensland, South Australia, and Victoria Police) who volunteered during our visit to their department. Their experience with prints ranged from 5 to 32 years and was 17.45 (SD = 7.53) years on average. We found that qualified, court-practicing fingerprint experts are exceedingly accurate compared with novices. Even though the novices could reliably distinguish matching and nonmatching prints, they made a large number of errors.

The performance difference between experts and novices on trials in which the prints matched was relatively small (92.12% correct for experts vs. 74.55% for novices). Comparably, the performance difference between experts and novices on trials in which the prints did not match, and were not similar, was also relatively small (100% correct for experts vs. 77.03% for novices). The performance difference between experts and novices on trials in which the prints were highly similar but did not match, however, was substantial; novice participants mistakenly identified 55.18% of the similar, nonmatching distractor prints as matches, whereas the corresponding rate for experts was 0.68%. The largest performance difference between novices and experts seems to lie in identifying highly similar, but nonmatching prints, as such. A comparison with novices was important for demonstrating expertise and shows that the matching task was difficult enough for experts to perform accurately, but for novices to perform relatively poorly.

Error Rates

Much has been made about "error rates" in fingerprint identification, and more so in light of the NAS report (2). In this section, we attempt to characterize error rates in our experiment and for fingerprint identification more generally.

Expert Matching Accuracy

In our experiment, 37 experts each compared 36 print pairs for a total of 1332 comparisons. Of the 444 comparisons in which the prints matched (targets), 22 of the 37 examiners incorrectly declared at least one of these matching prints as nonmatches, for an absolute total of 35 misses (hits = 92.12%; misses = 7.88%). Misses are the kind of error that could allow a guilty person to escape detection. Of the 444 comparisons in which the prints did not match, and were not similar (nonsimilar distractors), all of the examiners correctly declared the prints as nonmatches (correct rejections = 100%; false alarms = 0%). Of the 444 comparisons in which the prints did not match, but were highly similar (similar distractors), three examiners incorrectly declared three of these print pairs as matches (correct rejections = 99.32%; false alarms = 0.68%). These three print pairs were made up of three different latents. False alarms are the kind of error that could falsely incriminate an innocent person. (As an aside, it is not possible to link a participant's performance to the identity of a particular individual because experts and novices participated anonymously.) What, then, can be concluded about error rates from this experiment?

Determining Error Rates

Our study was not designed to determine the likelihood of errors in practice, nor the performance of individual practitioners or departments. It was designed to demonstrate expertise in fingerprint identification. Inferring, from our results, that "Fingerprint examiners are 99.32% accurate," or "The error rate of fingerprint identification is 0.68%," would be unjustified. Any claim of accuracy would have to be followed by a list of qualifiers. For example, "Some qualified fingerprint examiners are 99.32% accurate at correctly declaring nonmatching prints as such when the prints were obtained from the most similar nonmatch according to the NAFIS, and when examiners are not provided with their usual tools, or independent verification," and so on. The qualifiers are limitations only when trying to make an overgeneralization like, "fingerprint examiners are 99.32% accurate," which is close to impossible to make in any area of expertise (let alone on the basis of our results). It is, however, legitimate to conclude that experts are more accurate (and conservative) than novices, for example. If this is what we are concluding (and we are), then all of the qualifying remarks above are completely irrelevant.

Readers might be now considering some particular qualifiers to explain these results. For example, one might imagine that the ability to zoom and rotate prints will improve expert performance or that verification will reduce experts' error rate to zero. But, it is unlikely that one particular qualifier will be enough to entirely explain the results. Regardless, these qualifiers (and many others) are all testable hypotheses, should the answers be seen as necessary and important. If our goal—in addressing the critics and advancing the field—is to determine the precise error rate for each fingerprint examiner in each department, or even the field as a whole, then the necessary experiments become unwieldy.

We would need to unpack the different types of error that are possible (e.g., clerical, identification, sufficiency, misses, false alarms, disagreement, inadmissible rulings, etc.). We would need to establish who is going to be tested (e.g., trainees, intermediates, qualified experts, supervisors), under what conditions (e.g., distractions, interruptions, sleep deprivation, time and resource constraints, etc.), and on which part of the process (e.g., latent development, analysis, comparison, testifying, etc.). Will we include verification (and will it be blind)? How will we ensure ground truth? What sort of materials will be examined (e.g., lifts, surface types, lifting agents, powders, whorls, arches, low quality, distorted, highly similar, etc.)? What tools will be available (e.g., image enhancements, digital markers, zoom, rotation, computer search algorithms, statistical models, etc.)? How much time will they have? Can they collaborate? How many items will they be tested on, and so on. In the end, we would be left with some numbers (e.g., 85% hits and 6% false alarms for Jones under x, y, and z conditions). What, then, do we do with this information? It is difficult to see how the incredible amount of time, money, and resources required to get such an answer, would pay off. As discussed earlier, this approach may be ineffective and inefficient. And, we will be unable to locate the source of errors, to say nothing of taking steps to avoid them.

Levels of Analysis

Is a measure of individual error necessary for the science or for the court? The fingerprint profession, of course, will be concerned with their performance to make sure that they are on track and to

ensure continuous improvement. And, of course, the courts will be concerned with data that will help fact finders make optimal decisions. But, for example, demands are not made for individual error rates for a medical doctor or a field-wide error rate in medical diagnosis; only performance measures of the instrument or test on average are sought. To ask for error rates associated with a particular individual on a particular test seems, rightly, inappropriate in medicine. Similarly, focusing on the individual is the wrong level of analysis when attempting to characterize the accuracy of forensic fingerprint identification processes and systems.

A broader question concerns the level of analysis that is appropriate for presenting evidence and associated rates of error in court. At the extreme, an examiner could report how accurate they are at matching a whorl type print, lifted from a crime scene, on a wooden surface, using magnetic black powder, in a particular department, in a particular country, on a Tuesday, and so on. It may be necessary, or the courts may demand, that particular rates of error are established for particular situations. But unless it has been demonstrated that the level of accuracy (or proficiency, or reliability, or competence) varies systematically at any one of these levels, then the default should be to opt for reporting accuracy at the broader level.

Error Rates in Other Domains

Comparing the matching performance of fingerprint experts with experts in similar domains may give us an appreciation for their relative performance. Unlike with fingerprints, all people have expertise with faces. Psychologically, faces are similar to prints in that they are both complex visual patterns (24). People can easily recognize familiar faces despite changes in facial expression, context, and viewpoint (60). Unfamiliar faces, however, are extremely difficult to identify across these changes. Even in ideal conditions, where an unfamiliar target face is presented next to a set of candidate faces—with similar lighting, poses and no time constraints—people only match 68% of the faces correctly (61). Even people whose job requires them to identify unfamiliar faces from identity cards perform poorly at this simultaneous matching task (62). Based on the results of Tangen et al. (1), it is clear that fingerprint experts have impressive pattern matching abilities.

The accuracy of fingerprint experts becomes more impressive when compared with medical experts. Just as in fingerprint identification, however, it is difficult or impossible to determine general rates of field-wide error. But it is known that roughly 5% of autopsies reveal lethal diagnostic errors for which a correct diagnosis coupled with treatment could have averted death, and an estimated 40,000–80,000 U.S. hospital deaths result from misdiagnosis annually (63,64). These figures suggest that more Americans are killed in U.S. hospitals every 6 months than died in the entire Vietnam War and is equivalent to three fully loaded jumbo jets crashing every other day (but see [65]).

The prevalence of false positive diagnostic errors in perceptual specialties, such as radiology and pathology, is typically <5%, and increases to the range of 10–15% in emergency room type settings (66,67). Researchers are now focused on ways to reduce (not eliminate) diagnostic errors and on creating policy that defines acceptable rates of error (64). Of course it is difficult to define "error rates," let alone compare them across domains. But, it is clear, from the available data, that fingerprint experts demonstrate impressive pattern matching abilities that may rival those of medical diagnosticians; even despite the distinction that, arguably, identification (as in fingerprints) is a more difficult task than categorization (as in medical diagnosis).

Summary

Thus far, we have expanded on the results of the Identifying Fingerprint Expertise experiment (1) and explained that previous experiments and tests of proficiency were problematic and that the expertise of human fingerprint examiners had been assumed but not demonstrated. We have described the decisions and compromises that we made to design an experiment that tests the claimed expertise of human fingerprint examiners. In summary:

- Fidelity, generalizability, and control must be balanced to answer research questions. Our experiment was "artificial" for good reason. The goal was to understand the extent to which the difference between expert and novice performance is real, not the extent to which the experimental setting resembles the everyday operations of a fingerprint bureau.
- The validity, proficiency, and competence of fingerprint examiners are best determined when experiments include highly similar print pairs where the ground truth is known. Prints from the Forensic Informatics Biometric Repository were used to ensure ground truth.
- To best quantitate matching performance, a signal detection paradigm can be employed to separate the two ways of being right and the two ways of being wrong, to compare performance on matching and nonmatching prints, and to separate accuracy and response bias.
- Distinguishing highly similar, but nonmatching, prints from genuine prints is likely to be the most difficult and common task that examiners face. Similar distractor prints were obtained by searching simulated crime scene latents on the NAFIS to emulate this task.
- Considering both the evidence for the reasonable performance of novices and the notion that juries should make the ultimate decision, the most appropriate comparison group to demonstrate expertise should be novices who have no training with fingerprints whatsoever.
- Our study was not designed to determine the likelihood of errors in practice, nor the performance of individual practitioners or departments. As such, inferring from our results that, "Fingerprint examiners are 99.32% accurate," or "The error rate of fingerprint identification is 0.68%," would be unjustified.
- Determining error rates with black box studies may be unnecessary at best and ineffective and inefficient at worst, and unless one can demonstrate that a particular qualifier will systematically affect accuracy, the default should be to report accuracy at the broader level.
- Fingerprint experts posses impressive pattern matching abilities that may rival those of medical diagnosticians.

It appears that expertise in fingerprint identification does exist. That is, there are people who have demonstrable and specialized abilities for matching latent fingerprints to their source, and those abilities are superior to the person on the street. An examiner's expertise seems to be situated, not in their ability to match prints per se, but in their superior ability to identify highly similar, but nonmatching fingerprints as such. These results, and their comparison with novices, show that the accuracy of qualified examiners is substantially higher than inexperienced novices. Moreover, the 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.

It seems that some combination of training and the daily comparison of untold numbers of fingerprints leads to an uncanny ability to match fingerprints to their source. Experts are drawing on an entire career of experiences in making their decisions, as well as their training in fundamentals of fingerprint impressions to understand the "behavior" of minutiae. Experts, likely implicitly, understand the structure, regularities, and acceptable variation of fingerprint impressions. Future experiments could pinpoint the nature of this expertise. That is, whether expertise arises mainly from formal rules or the accumulation of instances (68).

Implications for Expert Testimony

The results from Tangen et al. (1) demonstrate that qualified fingerprint experts perform much better than novices at matching fingerprints, and their rates of error may be lower than those in diagnostic medicine, for example. The implications of these results for current models of expert admissibility, testimony, and policy are discussed below.

The Current Model

Current models of expert testimony vary from country to country and from state to state. The two largest bodies that provide consensus guidelines and standards for fingerprint identification—the Scientific Working Group on Friction Ridge Analysis, Science and Technology (69) and the International Association for Identification (70)—both stipulate that examiners are only permitted to testify to three conclusions: exclusion, inconclusive, and individualization. An individualization is defined as, "The determination by an examiner that there is sufficient quality and quantity of detail in agreement to conclude that two friction ridge impressions originated from the same source" (71, p. 6). When testifying, examiners often do not provide evidence of their claimed expertise or attempt to characterize their level of proficiency. Examiners may, when pushed by the courts, report that all fingerprints are unique or point to prenatal development and persistence. Despite considerable acrimony (2), examiners continue to make claims of individualization or similar ([72], but see [11]).

The Implications of the Identifying Fingerprint Expertise Experiment

Admissibility

Information about accuracy and performance—along with the relative performance of laypersons—is required for courts to make informed decisions about the admissibility of expert testimony. Experts outperform novices, but they do make errors (1). These results make it less likely that examiners themselves will suffer unfounded attacks on their expertise. If an examiner's expertise is challenged, then our methodology and design ought to be the target of criticism rather than the examiners themselves; assuming, of course, that their testimony does not extend beyond what our experiments can support.

The distinction, between the performance of experts and novices, is fundamental to the question of expert testimony, because it demonstrates specialized knowledge. Our experiment could be used as evidence for this distinction to satisfy legal admissibility criteria. And the results suggest that relying on juries to evaluate fingerprint evidence—when presented with the physical evidence alone, without expert commentary (56)—could result in a substantial number of false identification errors.

The NAS Report (2,3) noted the frequent absence of solid scientific research demonstrating the validity of forensic methods in general; of quantifiable measures of the reliability and accuracy of forensic analyses; and of quantifiable measures of uncertainty in the conclusions of forensic analyses. Fingerprint examiners have taken a first step in demonstrating their claimed expertise in controlled, representative situations in which ground truth is known. Examiners are now working with researchers toward understanding the source of identification errors, the factors that influence performance, and the nature of expertise in identification. In light of the NAS report, and the model for demonstrating expertise provided here by fingerprint examiners and researchers, it behooves other forensic pattern identification disciplines—such as shoeprints, bloodstains, DNA, ballistics, toolmarks, bitemarks, CCTV face identification, etc.—to conduct similar experiments to demonstrate expertise and performance in their own disciplines.

Testimony

Documented cases of false identification (9), issues of plausibility reported by the NAS (2), and recent experiments (1,32–34) highlight the need for a contemporary model of forensic testimony. Following developments in the United States and Canada, Edmond (73) has suggested that Australia adopt a reliability standard, and the U.K. Law Commission (4) has announced similar recommendations for admissibility practice in England and Wales. Indeed, science and legal commentators are beginning to call for empirical demonstrations of accuracy and performance, along with details about the relative performance of laypersons, across forensic science. A failure to respond to criticism means that judges are in danger of acting irrationally and being left behind by practical and ongoing reforms in the forensic sciences. While it is likely that courts will start to develop an admissibility and testimony jurisprudence more directly concerned with reliability in the near future, there is an independent need for forensic scientists and technicians to pay much closer attention to the evidence for ability and reliability (74). Edmond (73) suggests that reliability standards will help to make criminal trials fairer and ensure outcomes reflect the known value of expert evidence.

It is clear that an alternative to the current model of fingerprint testimony is required. But, what should an acceptable alternative and contemporary model look like? Several factors must be considered; these include the role of scientific experiments on the accuracy and reliability of forensic identification; whether it is necessary to report department or individual scores on properly controlled proficiency tests; the state of the science in other areas of pattern and impression identification; the impact that the testimony has on jury decision making; finding the right balance between accurate scientific reporting and the ability of judges and juries to understand expert testimony; and decisions about whether to report on the degree to which the specimen matches the source (e.g., "lends limited support"), the degree of confidence in a match (e.g., "highly confident that *x* matches *y*"), opinions about the evidence (e.g., "it is my opinion that…"), or statements about the particular hypotheses in question (e.g., the evidence is more consistent with *x* than *y*).

There is much research and consideration needed to develop an acceptable alternative model of fingerprint testimony. Several debates on this topic are raging internationally between academics,

statisticians, lawyers, forensic examiners, and managers. We are working toward proposing recommendations that do not extend beyond the capabilities of examiners or experimental findings while substantially engaging with critics to develop robust empirical guides to practice.

To Develop a Research Culture in Forensic Science

Researchers and professionals (e.g., [40]) have highlighted the need for a research culture in forensic science. Currently, however, it appears that professionals are expected to strengthen the scientific basis of their field but are not provided with the financial or intellectual support to do so. It is clear that examiners are doing their best to capture criminals and uphold civil liberties. But the lack of funding and resources in already overworked forensic departments makes basic research exceedingly unlikely. In addition, few have the methodological skills and expertise in the psychology of perception, cognition, bias, memory, accuracy, and decision making to ensure that their practice meets legal admissibility standards emerging internationally.

It is essential that we move beyond the adversarial system currently impeding advancement of the field and develop a culture of cooperation between researchers and examiners. The emergence of such a culture would fundamentally change examiners' relationship with empirical data and affect how evidence is understood and reported. Indeed, forensic examiners have expressed a desire to address the shortfalls of their discipline and engage in research.

Considering that forensic identification is based on human judgment, the field would benefit from further research on expert decision making. Clinical reasoning in medicine, for example, has developed over the last 40 years after it became increasingly apparent that physicians' decisions resulted in adverse consequences for patients (63). Much has been learned about the nature of medical expertise, the influence of perceptual and cognitive biases, and how to best incorporate such knowledge into practice. Researchers need to provide a scientific basis for demonstrating the validity of forensic methods and measures of uncertainty in the judgments of forensic analyses.

From here, more sophisticated questions can be asked than those about error rates and expertise. For example, what is the most effective way to train novices? What information is the most important for matching or excluding prints? What elements of the matching task best distinguish experts and novices? How do experts and novices differ in their use of this information? How does expertise with fingerprints develop over time? What is the relationship between the Analysis and Comparison phase of the identification process? How does time pressure influence performance? What is best practice in providing feedback and self-assessment? What is the most effective way to present fingerprint evidence to juries? The practical outcomes from answering questions such as these include a better understanding of the source of potential identification errors and factors that influence performance, a reduction in training time from novice to expert, more effective recruitment and training methods, and greater validity in presenting forensic evidence in court.

Maintaining high standards of evidence is vital for an effective justice system and ensuring that innocent people are not wrongfully accused. The reliability of forensic evidence and the value of expert testimony in the criminal justice system can be maximized by examining forensic reasoning and decision making. Given the inevitability of human error, the move should be toward fostering resilient systems capable of minimizing and acknowledging errors (75). We—collectively, forensic professionals, researchers, legal scholars, and the courts—need to define acceptable rates of error, foster a work environment conducive to learning from error, and promote a blame-free safety culture, as medicine is working toward (64,76).

This approach will allow police, intelligence systems, and investigators to interpret evidence more effectively and efficiently, assist forensic examiners in the development of evidence-based training programs, discourage exaggerated interpretations of forensic evidence, and help in the development of a model of expert testimony that does not extend beyond the capabilities of examiners or beyond the scope of experimental findings. Further psychological research into forensic decision making will help to ensure the integrity of forensics as an investigative tool available to police, so the rule of law is justly applied.

Acknowledgments

We thank Professor Gary Edmond for his valuable input toward this publication and Morgan Tear and Hayley Thomason for assistance with data collection.

References

1. Tangen JM, Thompson MB, McCarthy DJ. Identifying fingerprint expertise. Psychol Sci 2011;22(8):995–7.
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.
3. Edwards HT. 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, 2009.
4. Law Commission of England and Wales. Expert evidence in criminal proceedings in England and Wales. London: Stationery Office, 2011.
5. Loftus EF, Cole SA. Contaminated evidence. Science 2004;304 (5673):959.
6. Saks MJ, Koehler JJ. The coming paradigm shift in forensic identification science. Science 2005;309(5736):892–5.
7. Spinney L. The fine print. Nature 2010;464:344–6.
8. Federal Bureau of Investigation. The science of fingerprints: classification and uses. Washington, DC: U.S. Government Printing Office, 1984.
9. Cole SA. More than zero: accounting for error in latent fingerprint identification. J Crim Law Criminol 2005;95(3):985–1078.
10. Cole SA. Who speaks for science? A response to the National Academy of Sciences report on forensic science. Law Prob Ris 2010;9:25–46.
11. Garrett R. Memorandum from the President of the International Association for Identification, February 19, 2009, http://www.theiai.org/current_affairs/nas_memo_20090219.pdf. (accessed December 1, 2011).
12. Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST). Standard for the definition and measurement of rates of errors and non-consensus decisions in friction ridge examination (latent/tenprint). Ver. 1.1, 09/16/11, SWGFAST, 2011.
13. Cole SA, Welling M, Dioso-Villa R, Carpenter R. Beyond the individuality of fingerprints: a measure of simulated computer latent print source attribution accuracy. Law Prob Ris 2008;7(3):165–89.
14. Haber L, Haber RN. Scientific validation of fingerprint evidence under Daubert. Law Prob Ris 2008;7(2):119–26.
15. Vokey JR, Tangen JM, Cole SA. On the preliminary psychophysics of fingerprint identification. Q J Exp Psychol (Hove) 2009;62(5):1023–40.
16. Dror I, Charlton D. Why experts make errors. J Forensic Identif 2006;56 (4):600–56.
17. Dror I, Charlton D, Peron A. Contextual information renders experts vulnerable to making erroneous identifications. Forensic Sci Int 2006;156 (1):74–8.
18. Dror I, Cole S. The vision in "blind" justice: expert perception, judgment, and visual cognition in forensic pattern recognition. Psychon B Rev 2010;17(2):161–7.
19. Dror I, Péron A, Hind S, Charlton D. When emotions get the better of us: the effect of contextual top-down processing on matching fingerprints. Appl Cognitive Psychol 2005;19(6):799–809.

20. Dror I, Rosenthal R. Meta-analytically quantifying the reliability and biasability of forensic experts. J Forensic Sci 2008;53(4):900–3.
21. Langenburg G, Champod C, Wertheim P. Testing for potential contextual bias effects during the verification stage of the ACE-V methodology when conducting fingerprint comparisons. J Forensic Sci 2009;54(3):571–82.
22. Busey T, Dror I. Special abilities and vulnerabilities in forensic expertise. In: Holder EH Jr, Robinson LO, Laub JH, editors. Fingerprint sourcebook. Washington, DC: National Institute of Justice Press, 2010;15-1–15-23.
23. Busey T, Parada F. The nature of expertise in fingerprint examiners. Psychon B Rev 2010;17(2):155–60.
24. Busey T, Vanderkolk J. Behavioral and electrophysiological evidence for configural processing in fingerprint experts. Vision Res 2005;45(4):431–48.
25. Busey T, Yu C, Wyatte D, Vanderkolk J, Parada F, Akavipat R. Consistency and variability among latent print examiners as revealed by eye tracking methodologies. J Forensic Identif 2011;60(1):61–91.
26. Dror IE, Mnookin JL. The use of technology in human expert domains: challenges and risks arising from the use of automated fingerprint identification systems in forensic science. Law Prob Ris 2010;9(1):47–67.
27. Dror IE, Wertheim K, Fraser-Mackenzie P, Walajtys J. The impact of human-technology cooperation and distributed cognition in forensic science: biasing effects of AFIS contextual information on human experts. J Forensic Sci 2012;57(2):343–52.
28. Champod C, Evett I. A probabilistic approach to fingerprint evidence. J Forensic Identif 2001;51(2):101–22.
29. Neumann C, Champod C, Puch-Solis R, Egli N, Anthonioz A, Bromage-Griffiths A. Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. J Forensic Sci 2007;52:54–64.
30. Neumann C, Champod C, Puch-Solis R, Egli N, Anthonioz A, Meuwly D, et al. Computation of likelihood ratios in fingerprint identification for configurations of three minutiae. J Forensic Sci 2006;51(6):1255–66.
31. Neumann C. Fingerprints at the crime-scene: statistically certain, or probable? Significance 2012;9(1):21–5.
32. Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. Accuracy and reliability of forensic latent fingerprint decisions. Proc Natl Acad Sci USA 2011;108(19):7733–8.
33. Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. Repeatability and reproducibility of decisions by latent fingerprint examiners. PLoS ONE 2012;7(3):e32800.
34. Dror IE, Champod C, Langenburg G, Charlton D, Hunt H, Rosenthal R. Cognitive issues in fingerprint analysis: inter- and intra-expert consistency and the effect of a "target" comparison. Forensic Sci Int 2011;208(1):10–17.
35. Langenberg G. Performance study of the ACE-V process: a pilot study to measure the accuracy, precision, reproducibility, repeatability, and biasability of conclusions resulting from the ACE-V process. J Forensic Identif 2009;59(2):219–57.
36. Wertheim K, Langenburg G, Moenssens A. A report of latent print examiner accuracy during comparison training exercises. J Forensic Identif 2006;56(4):55–93.
37. Haber L, Haber RN. Letter to the editor. Re: a report of latent print examiner accuracy during comparison training exercises. J Forensic Identif 2006;56(4):493–9.
38. Wertheim K, Langenburg G, Moenssens A. Authors' response to letter: letter to the editor. Re: a report of latent print examiner accuracy during training exercises. J Forensic Identif 2006;56(4):500–10.
39. FBI responds to the office of inspector general's report on the fingerprint misidentification of Brandon Mayfield. 2006, http://www.fbi.gov/news/pressrel/press-releases/fbi-respondsto-the-office-of-inspector-general2019s-reporton-the-fingerprint-misidentification-of-brandon-mayfield (accessed November 3, 2011).
40. Mnookin JL, Cole SA, Dror IE, Fisher BAJ, Houck M, Inman K, et al. The need for a research culture in the forensic sciences. UCLA Law Rev 2011;58(3):725–79.
41. Spinney L. Forensic science braces for change. Nature 2010, doi: 10.1038/news.2010.369, http://www.nature.com/news/2010/100722/full/news.2010.369.html (accessed December 1, 2011).
42. Dror IE. The paradox of human expertise: why experts get it wrong. In: Kapur N, editor. The paradoxical brain. Cambridge: Cambridge University Press, 2011;177–88.
43. Dror IE. Combating bias: the next step in fighting cognitive and psychological contamination. J Forensic Sci 2012;57:276–7.
44. Koehler JJ. Fingerprint error rates and proficiency tests: what they are and why they matter. Hastings Law J 2007;59:1077–1110.
45. Thompson MB, Tangen JM, McCarthy DJ. Human matching performance of genuine crime scene latent fingerprints. Law and Hum Behav 2013 In press.
46. Mook DG. In defense of external invalidity. Am Psychol 1983;38(4):379–87.
47. Brinberg D, McGrath JE. Validity and the research process. Beverly Hills, CA: SAGE Publishing Co, 1985.
48. Woods DD. The observation problem in psychology (Westinghouse Technical Report). Pittsburgh, PA:Westinghouse Corporation, 1985.
49. Brunswik E. Perception and the representative design of psychological experiments, 2nd rev. ed. & enl. Berkeley, CA: University of California Press, 1956.
50. Rasmussen J, Pejtersen AM, Goodstein LP. Cognitive systems engineering. New York, NY: John Wiley & Sons, Inc, 1994.
51. Sanderson P. Capturing the essentials: simulator-based research in aviation and healthcare. Proceedings of the Eighth International Symposium of the Australian Aviation Psychology Association; 2008 Apr 8–11; Sydney, Australia. Sydney: Australian Aviation Psychology Association, 2008.
52. Sanderson P, Liu D, Jenkins S, Watson M, Russell WJ. Summative display evaluation with advanced patient simulation: fidelity, control, and generalizability. Brisbane, QLD: Cognitive Engineering Research Group, The University of Queensland, 2010. Report No.: CERG-2010-01.
53. Haber L, Haber RN. Error rates for human fingerprint examiners. In: Rath NK, Bolle RM, editors. Automatic fingerprint recognition systems. New York, NY: Springer-Verlag, 2004;339–60.
54. Phillips VL, Saks MJ, Peterson JL. The application of signal detection theory to decision-making in forensic science. J Forensic Sci 2001;46:294–308.
55. Green DM, Swets JA. Signal detection theory and psychophysics. New York, NY: Wiley, 1966.
56. Ericsson KA, Smith J. Toward a general theory of expertise: prospects and limits. Cambridge: Cambridge University Press, 1991.
57. Galton F. Decipherment of blurred fingerprints: supplementary chapter to "finger prints." London: Macmillan, 1893.
58. Cho A. Fingerprinting doesn't hold up as a science in court. Science 2002;295(5554):418.
59. Cho A. Judge reverses decision on fingerprint evidence. Science 2002;295(5563):2195–7.
60. Johnston R, Edmonds A. Familiar and unfamiliar face recognition: a review. Memory 2009;17(5):577–96.
61. Megreya AM, Burton AM. Matching faces to photographs: poor performance in eyewitness memory (without the memory). J Exp Psychol-Appl 2008;14(4):364–72.
62. Kemp R, Towell N, Pike G. When seeing should not be believing: photographs, credit cards and fraud. Appl Cognitive Psychol 1997;11(3):211–22.
63. Institute of Medicine. To err is human: building a safer health system. Washington, DC: National Academy Press, 2000.
64. Newman-Toker DE, Pronovost PJ. Diagnostic errors—the next frontier for patient safety. JAMA 2009;301(10):1060–2.
65. Hayward RA, Hofer TP. Estimating hospital deaths due to medical errors. JAMA 2001;286(4):415–20.
66. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. Am J Med 2008;121(5):S2–3.
67. Norman GR, Eva KW. Diagnostic error and clinical reasoning. Med Educ 2010;44(1):94–100.
68. Norman GR, Young M, Brooks L. Non-analytical models of clinical reasoning: the role of experience. Med Educ 2007;41(12):1140–5.
69. Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST). Standards for Conclusions. Ver. 1.0, 9/11/03, SWGFAST, 2003.
70. International Association For Identification (IAI). IAI position concerning latent fingerprint identification, 2007, www.onin.com/fp/IAI_Position_Statement_11-29-07.pdf. (accessed December 1, 2011).
71. Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST). Standard terminology of friction ridge examination. Ver. 3, 3/23/11, SWGFAST, 2011.
72. Cole SA. Forensics without uniqueness, conclusions without individualization: the new epistemology of forensic identification. Law Prob Ris 2009;8(3):233–55.
73. Edmond G. Specialised knowledge, the exclusionary discretions and reliability: reassessing incriminating expert opinion evidence. UNSW L J 2008;31(1):1–55.

74. Edmond G. Actual innocents? Legal limitations and their implications for forensic science and medicine. Aust J Forensic Sci 2011;43(2):177–212.
75. Hollnagel E, Woods DD, Leveson N. Resilience engineering: concepts and precepts. Alershot: Ashgate Publishing Company, 2006.
76. Woods DD, Johannesen L, Dekker S, Cook R, Sarter N. Behind human error, 2nd edn. Burlington, VT: Ashgate Publishing, Ltd., 2010.

Additional information and reprint requests:
Matthew B. Thompson, B.Inf.Tech., B.Sc.
School of Psychology
The University of Queensland
St Lucia QLD 4072
Australia
E-mail: mbthompson@gmail.com