Accuracy and reliability of forensic latent fingerprint decisions

Bradford T. Ulery[a], R. Austin Hicklin[a], JoAnn Buscaglia[b,1], and Maria Antonia Roberts[c]

[a]Noblis, 3150 Fairview Park Drive, Falls Church, VA 22042; [b]Counterterrorism and Forensic Science Research Unit, Federal Bureau of Investigation Laboratory Division, 2501 Investigation Parkway, Quantico, VA 22135; and [c]Latent Print Support Unit, Federal Bureau of Investigation Laboratory Division, 2501 Investigation Parkway, Quantico, VA 22135

Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved March 31, 2011 (received for review December 16, 2010)

The interpretation of forensic fingerprint evidence relies on the expertise of latent print examiners. The National Research Council of the National Academies and the legal and forensic sciences communities have called for research to measure the accuracy and reliability of latent print examiners' decisions, a challenging and complex problem in need of systematic analysis. Our research is focused on the development of empirical approaches to studying this problem. Here, we report on the first large-scale study of the accuracy and reliability of latent print examiners' decisions, in which 169 latent print examiners each compared approximately 100 pairs of latent and exemplar fingerprints from a pool of 744 pairs. The fingerprints were selected to include a range of attributes and quality encountered in forensic casework, and to be comparable to searches of an automated fingerprint identification system containing more than 58 million subjects. This study evaluated examiners on key decision points in the fingerprint examination process; procedures used operationally include additional safeguards designed to minimize errors. Five examiners made false positive errors for an overall false positive rate of 0.1%. Eighty-five percent of examiners made at least one false negative error for an overall false negative rate of 7.5%. Independent examination of the same comparisons by different participants (analogous to blind verification) was found to detect all false positive errors and the majority of false negative errors in this study. Examiners frequently differed on whether fingerprints were suitable for reaching a conclusion.

The interpretation of forensic fingerprint evidence relies on the expertise of latent print examiners. The accuracy of decisions made by latent print examiners has not been ascertained in a large-scale study, despite over one hundred years of the forensic use of fingerprints. Previous studies (1–4) are surveyed in ref. 5. Recently, there has been increased scrutiny of the discipline resulting from publicized errors (6) and a series of court admissibility challenges to the scientific basis of fingerprint evidence (e.g., 7–9). In response to the misidentification of a latent print in the 2004 Madrid bombing (10), a Federal Bureau of Investigation (FBI) Laboratory review committee evaluated the scientific basis of friction ridge examination. That committee recommended research, including the study described in this report: a test of the performance of latent print examiners (11). The need for evaluations of the accuracy of fingerprint examination decisions has also been underscored in critiques of the forensic sciences by the National Research Council (NRC, ref. 12) and others (e.g., refs. 13–16).

Background

Latent prints ("latents") are friction ridge impressions (fingerprints, palmprints, or footprints) left unintentionally on items such as those found at crime scenes (*SI Appendix*, *Glossary*). Exemplar prints ("exemplars"), generally of higher quality, are collected under controlled conditions from a known subject using ink on paper or digitally with a livescan device (17). Latent print examiners compare latents to exemplars, using their expertise rather than a quantitative standard to determine if the information content is sufficient to make a decision. Latent print examination can be complex because latents are often small, unclear, distorted, smudged, or contain few features; can overlap with other prints or appear on complex backgrounds; and can contain artifacts from the collection process. Because of this complexity, experts must be trained in working with the various difficult attributes of latents.

During examination, a latent is compared against one or more exemplars. These are generally collected from persons of interest in a particular case, persons with legitimate access to a crime scene, or obtained by searching the latent against an Automated Fingerprint Identification System (AFIS), which is designed to select from a large database those exemplars that are most similar to the latent being searched. For latent searches, an AFIS only provides a list of candidate exemplars; comparison decisions must be made by a latent print examiner. Exemplars selected by an AFIS are far more likely to be similar to the latent than exemplars selected by other means, potentially increasing the risk of examiner error (18).

The prevailing method for latent print examination is known as analysis, comparison, evaluation, and verification (ACE-V) (19, 20). The ACE portion of the process results in one of four decisions: the analysis decision of no value (unsuitable for comparison); or the comparison/evaluation decisions of individualization (from the same source), exclusion (from different sources), or inconclusive. The Scientific Working Group on Friction Ridge Analysis, Study and Technology guidelines for operational procedures (21) require verification for individualization decisions, but verification is optional for exclusion or inconclusive decisions. Verification may be blind to the initial examiner's decision, in which case all types of decisions would need to be verified. ACE-V has come under criticism by some as being a general approach that is underspecified (e.g., refs. 14 and 15).

Latent-exemplar image pairs collected under controlled conditions for research are known to be mated (from the same source) or nonmated (from different sources). An individualization decision based on mated prints is a true positive, but if based on nonmated prints, it is a false positive (error); an exclusion decision based on mated prints is a false negative (error), but is a true negative if based on nonmated prints. The term "error" is used in this paper only in reference to false positive and false negative conclusions when they contradict known ground truth. No such absolute criteria exist for judging whether the evidence is sufficient to reach a conclusion as opposed to making an inconclusive or no-value decision. The best information we have to

Author contributions: B.T.U., R.A.H., J.B., and M.A.R. designed research; B.T.U., R.A.H., J.B., and M.A.R. performed research; B.T.U. and R.A.H. contributed new analytic tools; B.T.U., R.A.H., J.B., and M.A.R. analyzed data; and B.T.U., R.A.H., J.B., and M.A.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: joann.buscaglia@ic.fbi.gov.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1018707108/-/DCSupplemental.

evaluate the appropriateness of reaching a conclusion is the collective judgments of the experts. Various approaches have been proposed to define sufficiency in terms of objective minimum criteria (e.g., ref. 22), and research is ongoing in this area (e.g., ref. 23). Our study is based on a black box approach, evaluating the examiners' accuracy and consensus in making decisions rather than attempting to determine or dictate how those decisions are made (11, 24).

Study Description

This study is part of a larger research effort to understand the accuracy of examiner conclusions, the level of consensus among examiners on decisions, and how the quantity and quality of image features relate to these outcomes. Key objectives of this study were to determine the frequency of false positive and false negative errors, the extent of consensus among examiners, and factors contributing to variability in results. We designed the study to enable additional exploratory analyses and gain insight in support of the larger research effort.

There is substantial variability in the attributes of latent prints, in the capabilities of latent print examiners, in the types of casework received by agencies, and the procedures used among agencies. Average measures of performance across this heterogeneous population are of limited value (25)—but do provide insight necessary to understand the problem and scope future work. Furthermore, there are currently no means by which all latent print examiners in the United States could be enumerated or used as the basis for sampling: A representative sample of latent print examiners or casework is impracticable.

To reduce the problem of heterogeneity, we limited our scope to a study of performance under a single, operationally common scenario that would yield relevant results. This study evaluated examiners at the key decision points during analysis and evaluation. Operational latent print examination processes may include additional steps, such as examination of original evidence or paper fingerprint cards, review of multiple exemplars from a subject, consultation with other examiners, revisiting difficult comparisons, verification by another examiner, and quality assurance review. These steps are implemented to reduce the possibility of error.

Ideally, a study would be conducted in which participants were not aware that they were being tested. The practicality of such an approach even within a single organization would depend on the type of casework. Fully electronic casework could allow insertion of test data into actual casework, but this may be complex to the point of infeasibility for agencies in which most examinations involve physical evidence, especially when chain-of-custody issues are considered. Combining results among multiple agencies with heterogeneous procedures and types of casework would be problematic.

In order to get a broad cross-section of the latent print examiner community, participation was open to practicing latent print examiners from across the fingerprint community. A total of 169 latent print examiners participated; most were volunteers, while the others were encouraged or required to participate by their employers. Participants were diverse with respect to organization, training history, and other factors. The latent print examiners were generally highly experienced: Median experience was 10 y, and 83% were certified as latent print examiners. More detailed descriptions of participants, fingerprint data, and study procedures are included in *SI Appendix*, *Materials and Methods*.

The fingerprint data included 356 latents, from 165 distinct fingers from 21 people, and 484 exemplars. These were combined to form 744 distinct latent-exemplar image pairs. There were 520 mated and 224 nonmated pairs. The number of fingerprint pairs used in the study, and the number of examiners assigned to each pair, were selected as a balance between competing research priorities: Measuring consensus and variability among examiners required multiple examiners for each image pair, while incorporating a broad range of fingerprints for measuring image-specific effects required a large number of images.

We sought diversity in fingerprint data, within a range typical of casework. Subject matter experts selected the latents and mated exemplars from a much larger pool of images to include a broad range of attributes and quality. Latents of low quality were included in the study to evaluate the consensus among examiners in making value decisions about difficult latents. The exemplar data included a larger proportion of poor-quality exemplars than would be representative of exemplars from the FBI's Integrated AFIS (IAFIS) (*SI Appendix*, Table S4). Image pairs were selected to be challenging: Mated pairs were randomly selected from the multiple latents and exemplars available for each finger position; nonmated pairs were based on difficult comparisons resulting from searches of IAFIS, which includes exemplars from over 58 million persons with criminal records, or 580 million distinct fingers (*SI Appendix*, section 1.3). Participants were surveyed, and a large majority of the respondents agreed that the data were representative of casework (*SI Appendix*, Table S3).

Noblis developed custom software for this study in consultation with latent print examiners, who also assessed the software and test procedures in a pilot study. The software presented latent and exemplar images to the participants, allowed a limited amount of image processing, and recorded their decisions, as indicated in Fig. 1 (*SI Appendix*, section 1.2). Each of the examiners was randomly assigned approximately 100 image pairs out of the total pool of 744 image pairs (*SI Appendix*, section 1.3). The image pairs were presented in a preassigned order; examiners could not revisit previous comparisons. They were given several weeks to complete the test. Examiners were instructed to use the same diligence that they would use in performing casework. Participants were assured that their results would remain anonymous; a coding system was used to ensure anonymity during analysis and in reporting.

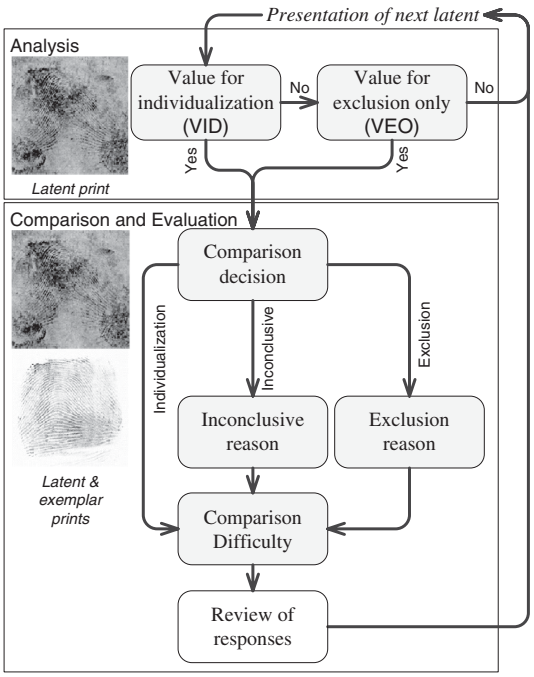

Fig. 1. Software workflow. Each examiner was assigned a distinct, randomized sequence of image pairs. For each pair, the latent was presented first for a value decision; if it was determined to be no value, the test proceeded directly to the latent from the next image pair; otherwise, an exemplar was presented for comparison and evaluation (*SI Appendix*, section 1.5).

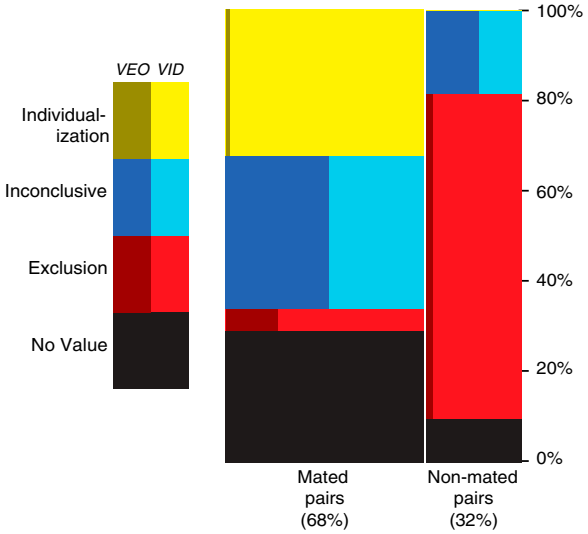

Fig. 2. Distribution of 17,121 decisions. 23% of all decisions resulted in no-value decisions (no comparison was performed); comparison decisions were based on latents of VID and of VEO; 7.5% of comparisons of mated pairs resulted in exclusion decisions (false negatives); 0.1% of comparisons of nonmated pairs resulted in individualization decisions (false positives—too few to be visible) (*SI Appendix*, Table S5).

Results

A summary of examiner decisions is shown in Fig. 2. We emphasize that individual examiner decisions are only a part of an overall operational process, which may include verification, quality assurance, and reporting. Our results do not necessarily reflect the performance of this overall operational process.

The true negative rate was greater than the true positive rate. Much of this difference may be explained by three factors: The amount of information necessary for an exclusion decision is typically less than for an individualization decision, examiners operate within a culture where false positives are seen as more serious errors than false negatives (5), and the mated pairs included a greater proportion of poor-quality prints than the nonmated pairs (*SI Appendix*, section 1.3). Whereas poor-quality latents result in the no-value decisions in Fig. 2, the poor-quality exemplars contribute to an increase in the proportion of inconclusive decisions.

Rates of comparison decisions can be calculated as a percentage of all presentations (PRES), including latents of no value; of comparisons where the latent was of value for individualization (VID); or of all comparisons (CMP), which includes comparisons where the latent was of value for exclusion only (VEO) as well as VID. Because standard operating procedures typically include only VID comparisons, this is our default basis for reporting these rates.

False Positives

Six false positives occurred among 4,083 VID comparisons of nonmated pairs (false positive rate, $FPR_{VID} = 0.1\%$) (*SI Appendix*, Tables S5 and S8; confidence intervals are discussed in *SI Appendix*, section 2.1). The image pairs that resulted in two of the false positives are shown in Fig. 3. Two of the false positive errors involved a single latent, but with exemplars from different subjects. Four of the five distinct latents on which false positives occurred (vs. 18% of nonmated latents) were deposited on a galvanized metal substrate, which was processed with cyanoacrylate and light gray powder. These images were often partially or fully tonally reversed (light ridges instead of dark), on a complex background (Fig. 3, image pair C). It is not known if other complex backgrounds or processing artifacts would have a similar increased potential for error.

The six errors were committed by five examiners, three of whom were certified (including one examiner who made two errors); one was not certified; one did not respond to our background survey. These correspond to the overall proportions of certifications among participants (*SI Appendix*, section 1.4). In no case did two examiners make the same false positive error: Five errors occurred on image pairs where a large majority of examiners correctly excluded; one occurred on a pair where the majority of examiners made inconclusive decisions. This suggests that these erroneous individualizations would have been detected if blind verification were routinely performed. For verification to be truly blind, examiners must not know that they are verifying individualizations; this can be ensured by performing verifications on a mix of conclusion types, not merely individualizations. The general consensus among examiners did not indicate that these were difficult comparisons, and only for two of the six false positives did the examiner making the error indicate that these were difficult (*SI Appendix*, Table S8).

There has been discussion (24, 26, 27) regarding the appropriateness of using qualified conclusions in investigation or testimony. The effects of qualified conclusions could be assessed in this study, as "inconclusive with corresponding features" (*SI Appendix*, section 1.5). Qualified conclusions potentially yield many additional "leads": 36.5% of VID comparisons resulted in individualization decisions, and an additional 6.2% resulted in qualified conclusions. However, 99.8% of individualization decisions were mated, as opposed to only 80.6% of qualified conclusions (*SI Appendix*, section 2). Only one of the six image pairs

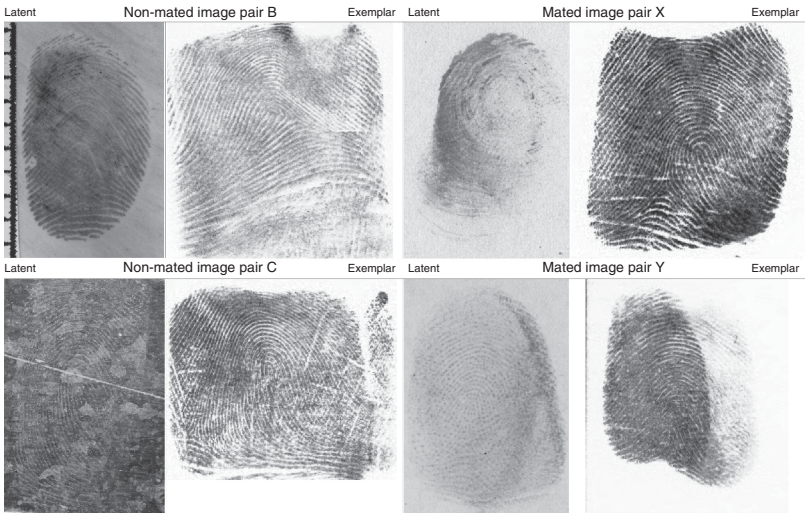

Fig. 3. Examples of fingerprint pairs used in the study that resulted in examiner errors. Pairs B and C resulted in false positive errors: 1 of 30 examiners made an individualization decision on B (24 exclusions); 1 of 26 examiners made an individualization decision on C (22 exclusions). The processing of the latent in C (cyanoacrylate with light gray powder) tonally reversed the image so that portions of ridges were light rather than dark. Pairs X and Y resulted in false negative errors, with no true positives made by any examiner: X was excluded by 13 of 29 examiners, presumably because the latent was deposited with a twisting motion that resulted in misleading ridge flow; Y was excluded by 15 of 18 examiners; the exemplar was particularly distorted. For use in this figure, these images were cropped to reduce background area.

that resulted in false positives had a plurality of inconclusive decisions, and none had a plurality "with corresponding features."

False Negatives

False negatives were much more prevalent than false positives (false negative rate: $FNR_{VID} = 7.5\%$) (*SI Appendix*, Table S5). Including VEO comparisons had no substantial effect: $FNR_{CMP} = 7.5\%$. Eighty-five percent of examiners made at least one false negative error, despite the fact that 65% of participants said that they were unaware of ever having made an erroneous exclusion after training (*SI Appendix*, section 1.4, no. 25); awareness of previous errors was not correlated with false negative errors on this test. False negatives were distributed across half of the image pairs that were compared. The likelihood of false negatives varied significantly by examiner (discussed further under *Examiner Skill*, below), and by image pair (*SI Appendix*, Figs. S3 and S5 *C* and *D*). Of the image pairs that were most frequently associated with false negatives, most had distorted latents and/or exemplars that gave an appearance of a different ridge flow pattern.

Verification of exclusions (especially blind verification) is not standard practice in many organizations, in part due to the large number encountered in casework. To investigate the potential benefits of blind verification, we posed the following question: Given a mated image pair, what is the probability, p_v, that two examiners would both reach exclusion decisions? If exclusions were equally likely for all image pairs (independence assumption), we would estimate that exclusions by two examiners would occur at the rate $p_v = FNR^2_{PRES} = 5.3\% \times 5.3\% = 0.3\%$ (*SI Appendix*, Table S5). However, the data show that the independence assumption is not valid: Some mated pairs are more likely to be excluded than others. Because the outcomes of blind verifications are not statistically independent but depend on the image pairs, we estimate $p_v = 0.85\%$ (*SI Appendix*, section 11). This suggests that blind verification of exclusions could greatly reduce false negative errors; agency policy would have to balance this benefit with the impact on limited resources.

For exclusions where the latent was VID, examiner assessment of comparison difficulty was a good predictor of accuracy, but even "Very Easy/Obvious" exclusions were sometimes incorrect: Among 450 false negatives where the latent was VID, 13 were rated "Very Easy/Obvious" by 11 distinct examiners (*SI Appendix*, Fig. S8). Latent value (VEO vs. VID) had no predictive value for false negative errors; however, exclusions were more likely to be true negatives when the latent was VID than when it was VEO. This counterintuitive result is due to the fact that VEO determinations were more often inconclusive, hence most exclusion decisions were associated with VID latents (*SI Appendix*, Fig. S7).

Posterior Probabilities

False positive and false negative rates are important accuracy measures, but assume a priori knowledge of true mating relationships, which of course are not known in forensic casework. In practice, knowledge of mating relationships is based solely on examiners' decisions: It is important to know the likelihood that these decisions are correct. Positive predictive value (PPV) is the percentage of individualization decisions that are true positives; negative predictive value (NPV) is the percentage of exclusion decisions that are true negatives. Fig. 4 depicts PPV and NPV as functions of the prior prevalence of mated pairs among the examinations performed: As the proportion of mated pairs increases, PPV increases and NPV decreases (*SI Appendix*, section 9). The prior prevalence of mated pair comparisons varies substantially among organizations, by case type, and by how candidates are selected. Mated comparisons are far more prevalent in cases where the candidates are suspects determined by nonfingerprint means than in cases where candidates were selected by an AFIS.

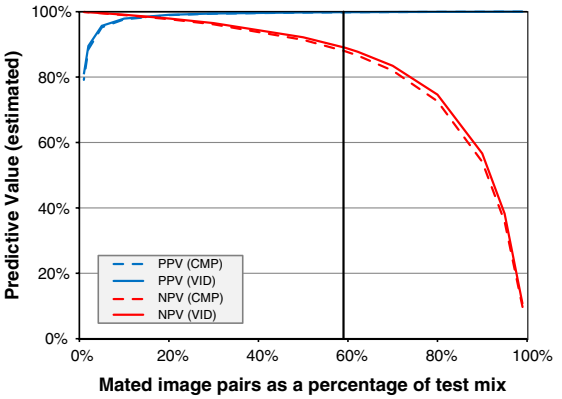

Fig. 4. PPV and NPV as a function of mate prevalence in workload. The observed predictive values ($PPV_{VID,59\%} = 99.8\%$ and $NPV_{VID,59\%} = 88.9\%$ for VID comparisons) correspond to the actual test mix (indicated) where 59% of VID comparisons were mated pairs; other predictive values are calculated as a function of mate prevalence. Sixty-two percent of all comparisons (VEO and VID) were performed on mated pairs, and $PPV_{CMP,62\%} = 99.8\%$ and $NPV_{CMP,62\%} = 86.6\%$.

Consensus

Each image pair was examined by an average of 23 participants. Their decisions can be regarded as votes in a decision space (Fig. 5). Consensus was limited on both mated and nonmated pairs: VID decisions were unanimous on 48% of mated pairs and 33% of nonmated pairs. Votes by latent print examiners also provide a basis for assessing sufficiency for value decisions, as shown in Fig. 6; consensus on individualization and exclusion decisions is shown in *SI Appendix*, Fig. S6.

Lack of consensus among examiners can be attributed to several factors. For unanimous decisions, the images were clearly the driving factor: Unusable or pristine prints resulted in unanimous decisions, and therefore different data selection would have affected the extent of consensus. When there was a lack of consensus, much of the variation could be explained by examiner differences: Examiners showed varying tendencies toward no-

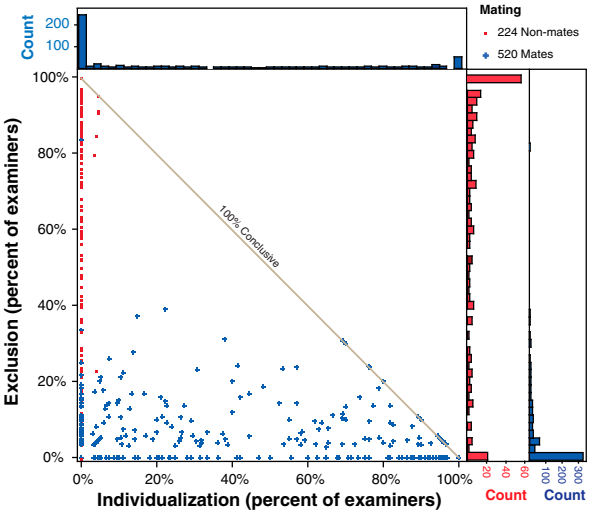

Fig. 5. Decision rates on each image pair. Percentage of examiners making an individualization decision (*x* axis) vs. exclusion decision (*y* axis) on each image pair; mean 23 presentations per pair. VEO and no-value decisions are treated as inconclusive. Marginal distributions are shown as histograms. Of mated pair decisions, 10% were unanimous true positives, 38% unanimous inconclusives. Of nonmated pair decisions, 25% were unanimous true negatives, 9% were unanimous inconclusives. Points along diagonal represent pairs on which all examiners reached conclusions. The prevalence of false negatives is evident in the vertical spread of mated pairs; the few false positives are evident in the limited horizontal spread of the nonmated pairs.

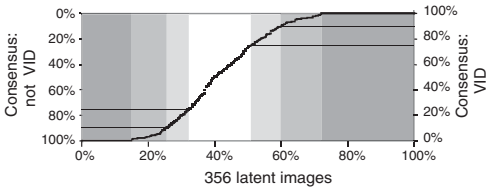

Fig. 6. Examiner consensus on VID decisions, showing the percentage of examiners reaching consensus (*y* axis) on each latent (*x* axis). Areas of unanimous (100%), decile (10%, 90%), and quartile (25%, 75%) consensus are marked. For example, at a 90% level of consensus (*y* axes), examiners agreed that 40% of the latents were VID (interval from 60% to 100% indicated by a horizontal line in upper right) (*SI Appendix*, Table S11). Such measures of consensus may be useful in developing quantity and quality metrics.

value or inconclusive decisions, or toward conclusions (*SI Appendix*, Fig. S4). Examiners differed significantly in conclusion rates, and we see this effect as secondary to image characteristics in explaining lack of consensus. Other factors accounting for lack of consensus include intraexaminer inconsistency and (presumably) test environment (*SI Appendix*, Fig. S3).

It was not unusual for one examiner to render an inconclusive decision while another made an individualization decision on the same comparison. This result is consistent with previous observations (1, 5, 28). Among all decisions based on mated pairs, 23.0% resulted in decisions other than individualization even though at least one other examiner made a true positive on the same image pair; 4.8% were not individualization decisions even though the majority of other examiners made true positives. This has operational implications in that some potential individualizations are not being made, and contradictory decisions are to be expected.

When examiners reached contradictory conclusions (exclusion and individualization) on a single comparison, the exclusion decision was more frequently in error: 7.7% of independent examinations of conclusions on mates were contradictory, vs. 0.23% on nonmates. Which of the contradictory decisions is more likely to be erroneous depends on the prior prevalence of mated vs. nonmated pairs: Exclusion decisions are more likely to be erroneous except in situations where the prior prevalence of nonmated pairs is very high.

Examiner Skill

The criminal justice system relies on the skill of latent print examiners as expert witnesses. Currently, there is no generally accepted objective measure to assess the skill of latent print examiners. Skill is multidimensional and is not limited to error rates (FPR and FNR), but also includes TPR, true negative rate (TNR), VID and VEO rates, and conclusion rate (CR—the percentage of individualization or exclusion conclusions as opposed to no-value or inconclusive decisions). Any assessment of skill must consider these dimensions. Although most discussions of examiner skill focus on error rates (e.g., ref. 13), the other aspects of examiner skill are important not just to the examiner's organization, but to the criminal justice system as well; e.g., an examiner who is frequently inconclusive is ineffective and thereby fails to serve justice. Both individual examiners and organizations must strike a proper balance between the societal costs of errors and inappropriate decisions, and the operational costs of detection. Contradictory verification decisions, whether involving erroneous conclusions or inappropriate inconclusive decisions, should be internally documented and addressed through an organization's continual improvement processes.

We found that examiners differed substantially along these dimensions of skill, and that these dimensions were largely independent. Our study measured all of these dimensions with the exception of FPRs for individual examiners, which were too low to measure with precision (*SI Appendix*, section 3). Fig. 7 shows that examiners' conclusion rates (CR_{PRES}) varied from 15 to 64% (mean 37%, SD 10%) on mated pairs, and from 7 to 96% (mean

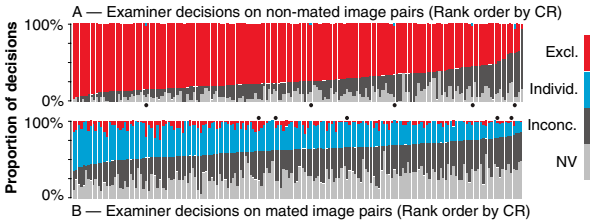

Fig. 7. Decision rates by examiner. Proportions of decisions for all 169 examiners on (*A*) nonmated and (*B*) mated image pairs. Examiners in each chart are sorted on CR. Each examiner was randomly assigned 51 to 74 mated image pairs (mean 69, SD 5) and 26 to 53 nonmated image pairs (mean 33, SD 7). In both, errors are shown in red. Column width indicates the number of image pairs. Examiners who made false positive errors are indicated with black dots (*SI Appendix*, Table S7).

71%, SD 14%) on nonmated pairs. The observed range in CRs may be explained by a higher level of skill (ability to reach more conclusions at the same level of accuracy), or it may imply a higher risk tolerance (more conclusions reached at the expense of making more errors).

Fig. 7 shows substantial variability in CR among examiners. These measured rates were based on an average of 69 mated presentations and 33 nonmated presentations. The limited number of presentations resulted in a wide margin of measurement error when evaluating the performance of an individual examiner (*SI Appendix*, Fig. S5). Although the estimates for each examiner are statistically unbiased, the sampling error in these estimates contributed substantially to the observed variability among examiners. The observed variability is a biased estimate that overstates the true variability (*SI Appendix*, Figs. S3*B* and S4).

Fig. 8 shows the relations between three of the skill dimensions measured for each examiner. Blue squares near the lower right of the chart represent highly skilled examiners: accurate (making few or no errors) and effective (high TNR and TPR, and therefore high CR). The red cross at the bottom left denotes an accurate (0% FNR_{VID}), but ineffective (5% TNR_{VID}, 16% TPR_{PRES}) examiner. The examiner denoted by the red cross at the top right is inaccurate (34% FNR_{VID}), and has mixed effectiveness (100% TNR_{VID}, 23% TPR_{PRES}). Attempting to compare the skill of any two examiners is a multidimensional problem. A combination of multiple dimensions into a single hypothetical measure of skill would require a weighting function to trade off the relative value of each dimension; such weighting might be driven by policy, based on the relative cost/benefit of each dimension for operational needs.

Tests could be designed to measure examiner skill along the multiple dimensions discussed here. Such tests could be valuable

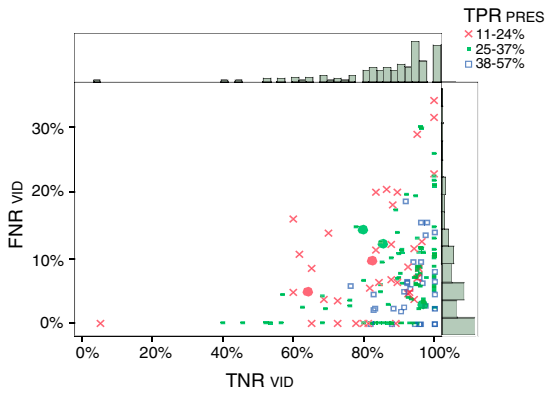

Fig. 8. Examiner skill. Each of the 169 examiners is plotted on three skill dimensions: TNR_{VID} (mean 88%, SD 13.6%), FNR_{VID} (mean 7.5%, SD 7.3%), and TPR_{PRES} (shown in color, with red crosses denoting the lowest quartile and blue squares the highest quartile; mean 32%, SD 9.4%). The five examiners who made false positive errors are indicated with bold filled circles.

not just as traditional proficiency tests with pass/fail thresholds, but as a means for examiners or their organizations to understand skills for specific training, or for tasking based on skills (such as selecting examiners for verification based on complementary skill sets).

Certified examiners had higher conclusion rates than noncertified examiners without a significant change in accuracy (significantly higher TPR_{VID} and TNR_{VID}; FNR_{VID} did not vary significantly) (*SI Appendix*, section 6). Length of experience as a latent print examiner did not show a significant correlation with TPR_{VID}, TNR_{VID}, or FNR_{VID} (*SI Appendix*, Table S9 and Fig. S2).

Examiners with a lower TPR_{VID} tended also to have a lower TNR_{VID}. Examiners with a higher FNR_{VID} tended to have a lower TPR_{VID}. Examiners with a higher TNR_{VID} tended also to have a higher FNR_{VID} (*SI Appendix*, Table S9 and Fig. S2).

Conclusions

Assessing the accuracy and reliability of latent print examiners is of great concern to the legal and forensic science communities. We evaluated the accuracy of decisions made by latent print examiners on difficult fingerprint comparisons in a computer-based test corresponding to one stage in AFIS casework. The rates measured in this study provide useful reference estimates that can inform decision making and guide future research; the results are not representative of all situations, and do not account for operational context and safeguards. False positive errors (erroneous individualizations) were made at the rate of 0.1% and never by two examiners on the same comparison. Five of the six errors occurred on image pairs where a large majority of examiners made true negatives. These results indicate that blind verification should be highly effective at detecting this type of error. Five of the 169 examiners (3%) committed false positive errors, out of an average of 33 nonmated pairs per examiner.

False negative errors (erroneous exclusions) were much more frequent (7.5% of mated comparisons). The majority of examiners (85%) committed at least one false negative error, with individual examiner error rates varying substantially, out of an average of 69 mated pairs per examiner. Blind verification would have detected the majority of the false negative errors; however, verification of exclusion decisions is not generally practiced in operational procedures, and blind verification is even less frequent. Policymakers will need to consider tradeoffs between the financial and societal costs and benefits of additional verifications.

Most of the false positive errors involved latents on the most complex combination of processing and substrate included in the study. The likelihood of false negatives also varied by image. Further research is necessary to identify the attributes of prints associated with false positive or false negative errors, such as quality, quantity of features, distortion, background, substrate, and processing method.

Examiners reached varied levels of consensus on value and comparison decisions. Although there is currently no objective basis for determining the sufficiency of information necessary to reach a fingerprint examination decision, further analysis of the data from this study will assist in defining quality and quantity metrics for sufficiency. This lack of consensus for comparison decisions has a potential impact on verification: Two examiners will sometimes reach different conclusions on a comparison.

Examiner skill is multidimensional and is not limited to error rates. Examiner skill varied substantially. We measured various dimensions of skill and found them to be largely independent.

This study is part of a larger ongoing research effort. To further our understanding of the accuracy and reliability of latent print examiner decisions, we are developing fingerprint quality and quantity metrics and analyzing their relationship to value and comparison decisions; extending our analyses to include detailed examiner markup of feature correspondence; collecting fingerprints specifically to explore how complexity of background, substrate and processing are related to comparison decisions; and measuring intraexaminer repeatability over time.

This study addresses in part NRC Recommendation 3 (12), developing and quantifying measures of accuracy and reliability for forensic analyses, and will assist in supporting the scientific basis of forensic fingerprint examination. The results of this study will provide insight into developing operational procedures and training of latent print examiners and will aid in the experimental design of future proficiency tests of latent print examiners.

ACKNOWLEDGMENTS. We thank the latent print examiners who participated in this study, as well as William Fellner, Jill McCracken, Keith Ward, Stephen Meagher, Calvin Yeung, Ted Unnikumaran, Erik Stanford, and William Chapman. This is publication number 10-19 of the FBI Laboratory Division. This work was funded in part under a contract award to Noblis, Inc. from the FBI Biometric Center of Excellence and in part by the FBI Laboratory Division. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the US government.

1. Evett IW, Williams RL (1995) A review of the 16 point fingerprint standard in England and Wales. *Fingerprint Whorld* 21.
2. Wertheim K, Langenburg G, Moenssens A (2006) A report of latent print examiner accuracy during comparison training exercises. *J Forensic Identification* 56:55–93.
3. Gutowski S (2006) Error rates in fingerprint examination: The view in 2006. *Forensic Bulletin* 2006:18–19 Autumn.
4. Langenburg G, Champod P, Wertheim P (2009) Testing for potential contextual bias effects during the verification stage of the ACE-V methodology when conducting fingerprint comparisons. *J Forensic Sci* 54:571–582.
5. Langenburg G (2009) A performance study of the ACE-V process. *J Forensic Identification* 59:219–257.
6. Cole SA (2005) More than zero: Accounting for error in latent fingerprint identification. *J Crim Law Criminol* 95:985–1078.
7. *United States v Mitchell* No. 96-407 (ED PA 1999).
8. *United States v Llera Plaza* Cr. No. 98-362-10, 11, 12 (ED PA 2002).
9. *Maryland v Rose* No. K06-0545 (MD Cir 2007).
10. Office of the Inspector General (2006) *A Review of the FBI's Handling of the Brandon Mayfield Case* (US Department of Justice, Washington, DC).
11. Budowle B, Buscaglia J, Perlman RS (2006) Review of the scientific basis for friction ridge comparisons as a means of identification: Committee findings and recommendations. *Forensic Sci Commun* 8:1.
12. National Research Council (2009) *Strengthening Forensic Science in the United States: A Path Forward* (National Academies Press, Washington, DC).
13. Koehler JJ (2008) Fingerprint error rates and proficiency tests: What they are and why they matter. *Hastings Law J* 59:1077–1110.
14. Mnookin JL (2008) The validity of latent fingerprint identification: Confessions of a fingerprinting moderate. *Law Probability and Risk* 7:127–141.
15. Haber L, Haber RN (2008) Scientific validation of fingerprint evidence under Daubert. *Law Probability and Risk* 7:87–109.
16. Cole S (2006) Is fingerprint identification valid? Rhetorics of reliability in fingerprint proponents' discourse. *Law Policy* 28:109–135.
17. Scientific Working Group on Friction Ridge Analysis, Study and Technology (2011) Standard terminology of friction ridge examination, Version 3., Available at http://www.swgfast.org/documents/terminology/110323_Standard-Terminology_3.0.pdf.
18. Dror I, Mnookin J (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 and Risk* 9:47–67.
19. Huber RA (1959) Expert witness. *Criminal Law Quarterly* 2:276–296.
20. Ashbaugh D (1999) *Quantitative-Qualitative Friction Ridge Analysis: An Introduction to Basic and Advanced Ridgeology* (CRC Press, New York).
21. Scientific Working Group on Friction Ridge Analysis, Study and Technology (2002) Friction ridge examination methodology for latent print examiners, Version 1.01., Available at http://www.swgfast.org/documents/methodology/100506-Methodology-Reformatted-1.01.pdf.
22. Champod C (1995) Edmond Locard—Numerical standards and "probable" identifications. *J Forensic Identification* 45:136–155.
23. Neumann C, et al. (2007) Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *J Forensic Sci* 52:54–64.
24. Mnookin JL (2008) Of black boxes, instruments, and experts: Testing the validity of forensic science. *Episteme* 5:343–358.
25. Budowle B, et al. (2009) A perspective on errors, bias, and interpretation in the forensic sciences and direction for continuing advancement. *J Forensic Sci* 54:798–809.
26. Saks M, Koehler J (2005) The coming paradigm shift in forensic identification science. *Science* 309:892–895.
27. Stoney DA (1991) What made us ever think we could individualize using statistics? *J Forensic Sci Soc* 31:197–199.
28. Grieve DL (1996) Possession of truth. *J Forensic Identification* 46:521–528.

A Study of the Accuracy and Reliability of Forensic Latent Fingerprint Decisions
Appendix: Supporting Information

1 Materials and Methods

1.1 Participants

Participation in this study was open to all practicing latent print examiners. Participation was solicited at the 2009 International Association for Identification (IAI) International Educational Conference, at SWGFAST, and by direct contact with various forensic organizations, some of which encouraged or required participation by their latent print examiners. A total of 169 latent print examiners participated; all volunteers were accepted. More examiners and organizations indicated interest in the study than actually participated and provided results, generally citing time constraints as the reason: 66.5% of those who indicated interest completed the test. Three of the participants sent in partial results; these examiners were not included in the analysis.

Although the participants were not randomly selected, the study successfully included a diverse participant group with respect to organization, training history, level of experience, and other factors. The latent print examiners were generally highly experienced: median experience was 10 years, and 83% were certified as latent print examiners. The impact of years of experience and certification on performance is discussed in SI-5 and SI-6.

The participants completed a background survey, the responses from which were used to assess experience and to understand the latent print examiners' standard operating procedures used in casework (see SI-1.4). Of the responding participants, 49% were certified as latent print examiners by the IAI, an additional 33% were certified by their employers or other organizations, and 18% were uncertified (Survey question #13). Forty-eight percent of the participants were employed by U.S. federal agencies, 23% by state agencies, 21% by local agencies, 7% by private organizations, and 1% by non-U.S. organizations (Survey question #4). The median amount of experience as a latent print examiner was 10 years (Survey question #9).

Six participants reported having made an erroneous individualization after training: one on a proficiency test; four on casework, detected during verification; one on casework, detected after reporting (Survey question #24). Fifty-four participants reported having made an erroneous exclusion after training: four on a proficiency test; forty-three on casework, detected during verification; ten on casework, detected after reporting; three examiners noted erroneous exclusions in two categories (Survey question #25).

1.2 Study description

Noblis developed custom, interactive software for this study that presented latent and exemplar fingerprint images, provided limited image processing capabilities, and recorded test responses. Image processing capabilities were deliberately limited in order to prevent differences in image processing skills from contributing substantially to examiner variability in the outcomes. The tests were distributed to participants on DVDs. Each participant was assigned a unique identifier, which was associated with a randomized sequence of image comparisons and was used to anonymize results. A set of practice data was included so that participants could become familiar with the software and procedures before starting the test itself.

Participants were instructed that the computer screen must accurately display the images presented by the software, and the software included a check for that purpose. The test environment was not controlled. Participants were given several weeks to complete the test, but there were not specific time requirements. They spent a median of 8 total hours on the test, over multiple sittings. Participants were not told what proportion of the image pairs were mated, nor was any feedback on the correctness of responses provided during the test. Participants were instructed that it was imperative that they conduct their analyses and comparisons in this study with the same diligence that they would use in performing casework. Participants were fully aware that they were participating in a study. Some may have been highly motivated to perform well in light of recent controversies concerning the accuracy of fingerprint identification, while others may not have taken the test as seriously as casework. The majority of the participants were volunteers, but some were encouraged or required by their employer/management to participate. We do not know the extent to which such factors may have altered performance. Motivational factors may have improved or worsened individual performance, or simply shifted their decision thresholds to be more or less conservative on borderline decisions.

For each image pair, the latent was presented for analysis, and the examiner was asked if the image was of value for individualization (VID); if the image was not VID, the examiner was asked if the image was of value for exclusion only (VEO); if the image was not VEO, the image was considered no value (NV). If the latent was NV, the exemplar was not presented for comparison; otherwise, the exemplar was presented and the examiner was required to make a decision of individualization, exclusion, or inconclusive. The specific wording used in the test and instructions is included in SI-1.5. The comparison decision corresponds directly to SWGFAST guidelines (*1*). The VEO decision is used operationally by a minority of participants (Survey question #29). Three questions were asked to provide additional insight into the examiner decisions: explanation of exclusion decisions, explanation of inconclusive decisions, and comparison difficulty. The inconclusive categorizations are in the forthcoming ANSI/NIST ITL-1 2011 standard (*2*). The categories for exclusions are based on the process for exclusion defined in the SWGFAST draft standards for conclusions (*3*): one can exclude on ridge flow alone, or sometimes must revert to level 2/3 detail. Examiners were able to review and correct their responses before proceeding to the next comparison, but could not revisit previous comparisons, or skip comparisons and return to them later. Examiners were also able to leave comments associated with specific test responses. All such comments were reviewed; a few data entry errors were excluded based on these comments.

The test was distributed to participants in multiple iterations. All testing was conducted in July-December 2009. An early version of the software used in 4% of the comparisons had intermittent image display problems. We found no indication that the display problems increased error rates or decreased conclusion rates. Early versions of the software used in 26% of the comparisons did not provide a final review of responses for each image pair; for those comparisons we cannot exclude the possibility of an undetected data entry error, although all comments were reviewed, and the few comparisons with comments that indicated data entry errors were excluded.

Prior to the tests, a pilot study was conducted to evaluate the test software prototype and to provide sizing information for the statistical design; the results of the pilot study are not included in our analysis. A repeatability test is in progress at the time of this writing.

Anonymity was part of the Institutional Review Board for Human Subject Research approval for this study. Participants were assured that

> *"Results will be anonymous. A blind coding system will ensure anonymity: participants receive a subject ID number from the study facilitator (at FBI Lab); the data sent by the study facilitator to the analysis team (at Noblis) only*

contains the subject ID number (not name or affiliation); the analysis team associates the subject ID with an anonymized code; the data and interpretation is based solely on the anonymized code; cross-references between the subject IDs and the anonymized codes will be destroyed after completion of the research. Therefore, the identities of participants will not be associated with the results at any point during analysis, and such association will not be possible subsequently, such as for discovery. The background survey does not request participants' names, employer, or any other personally identifying information. Reporting: Results will be aggregated across multiple examiners, based on categories of experience established in the background questionnaire. As noted, results will be anonymous; care will be taken so that the results are not aggregated in a way that compromises anonymity."

1.3 Fingerprint data

The fingerprints for the study were collected at the FBI Laboratory and at Noblis under controlled conditions to guarantee that the true mate relationships were known with certainty. The collection of fingerprint data from human subjects was approved by the FBI Institutional Review Board and the Noblis Institutional Review Board. Use of examiners in the study was approved by the FBI Institutional Review Board. Informed consent was obtained from all subjects. The overall fingerprint dataset included approximately 1,000 latents and an average of 13 exemplar impressions per finger position. Latents were captured electronically at 8-bit grayscale, uncompressed, at a resolution of 1,000 pixels per inch. Porous materials were processed using ninhydrin or physical developer, and were scanned using flatbed scanners. Latents on non-porous materials were processed using cyanoacrylate and light gray powder (captured using digital cameras), or using black powder, with lift cards scanned using flatbed scanners. The processing methods are summarized in Table S1. The exemplars included rolled and plain impressions, and were captured as inked prints on paper cards or using FBI-certified livescan devices; exemplars were 8-bit grayscale, 1,000 or 500 pixels per inch, and either uncompressed or compressed using Wavelet Scalar Quantization (*4*).

Processing method	Count	%
Black powder	126	35%
Ninhydrin	118	33%
Physical developer	55	15%
Cyanoacrylate with light gray powder	57	16%

Table S1: Distribution of processing methods for the 356 latents.

Deposition method	Count	%
Touch	166	47%
Roll	57	16%
Slide	45	13%
Twist	42	12%
(not recorded)	46	13%

Table S2: Distribution of deposition method for the 356 latents. Deposition method describes the motion with which the latent was deposited.

The latents and mated exemplars for the study were selected from the overall fingerprint dataset by subject matter experts to include a range of characteristics and quality, to be broadly representative of prints encountered in casework (Table S3). The term "broadly representative" is used because while any of the individual images could have been examples taken from casework, the overall distribution of the fingerprint data cannot as a whole be considered as statistically representative of operational data. The exemplar data included a larger proportion of poor-quality exemplars than is representative of operational IAFIS data, based on the NIST Fingerprint Image Quality (NFIQ) metric (*5*,Table S4). NFIQ was developed specifically to predict the likelihood that an exemplar will be correctly matched in an image-based search of a large-scale ten-print exemplar AFIS. NFIQ is used operationally in the U.S. Department of Homeland Security's US-VISIT IDENT system, and in the FBI's IAFIS system. There is no analogous means of comparing the latent data to operational casework: the difficulty and attributes of latents used in casework vary significantly among organizations, and nascent quality metrics have not been used to characterize operational casework. The data used in the study included many latents of relatively low quality, specifically to evaluate the consensus among examiners in making value decisions.

Overall, are the latent prints in the Black Box study representative of your casework?		
Strongly Agree	12	17%
Agree	47	66%
Disagree	9	13%
Strongly Disagree	3	4%
Overall, are the exemplar prints (knowns) in the Black Box study representative of your casework?		
Strongly Agree	11	15%
Agree	42	59%
Disagree	15	21%
Strongly Disagree	3	4%
How does the overall difficulty of the comparisons in the Black Box study correspond to your casework?		
Easier	6	8%
Similar	60	85%
Harder	5	7%

Table S3: Participant responses to three questions regarding how the Black Box data corresponded to actual casework. 71 of the participants responded; this group under-represented U.S. federal employees.

NFIQ	Black Box exemplars	Operational IAFIS exemplars
1 (Good)	24.9%	35.7%
2	12.1%	29.6%
3	28.8%	18.5%
4	13.0%	7.2%
5 (Poor)	21.1%	8.8%

Table S4: Comparison of Black Box exemplar quality to operational IAFIS exemplar quality, using the NIST Fingerprint Image Quality (NFIQ) metric. Operational IAFIS data is based on 13,709,420 criminal arrest transactions received prior to April 2010. Black box exemplars include a greater proportion of lower quality images (NFIQ={3,4,5}) than the operational IAFIS data: 62.9% of black box exemplars correspond to the worst 34.5% of operational IAFIS exemplars. Values are averages of all 10 finger positions.

The fingerprint data included 356 latents, from 165 distinct fingers of 21 distinct subjects; and 484 exemplars. These were combined to form 744 distinct image pairs, with each pair including one latent and one exemplar. There were 520 mated pairs and 224 non-mated pairs.

The proportion of mated to non-mated pairs reflects the desire to achieve an approximate balance in the number of comparisons performed on mated and non-mated pairs, and the desire for a large variety of image pairs, limited by the difficulty of selecting challenging non-mated pairs. The test set included more mated pairs to compensate for the higher proportion of poor quality latents among the mated pairs: while 70% of the image pairs were mated (68% of presentations), 59% of the resulting VID comparisons were of mated pairs. No proportion of mated to non-mated pairs would be representative of all situations: the actual base rate for casework can be assumed to vary between agencies and by the type of case.

Prints to be compared were selected for difficulty. Mated pairs were selected at random from the latents and exemplars just described, which were both disproportionately poor quality. The selection of non-mated pairs was designed to yield difficult comparisons comparable to those resulting from searches of the FBI's Integrated AFIS (IAFIS). At the time of data collection, IAFIS included exemplars from over 58 million persons with criminal records, or 580 million distinct fingers; the size of the IAFIS database continues to grow. For each latent, one of the corresponding exemplars was cropped to correspond to the portion of the print visible in the latent, and then searched against IAFIS as an image-based search (no human markup). The cropped exemplar was used solely to enable the search and was not used in the test. The cropped exemplar searches were conducted as image-based searches ("LFIS"), limited by finger position and pattern classifications in cases where classification was unambiguous. This approach was chosen, instead of searching the latents with marked features, to increase the likelihood that the non-mated candidates would be similar to the actual fingerprint, not just to the features marked in the latent. When a latent is used to search IAFIS, a list of twenty candidate exemplars is returned by default. Approximately one-

half of the non-mated pairs were selected by an experienced latent print examiner, who was not a participant, using one of two processes with the objective of maximizing the difficulty of comparisons: either the examiner selected from the twenty candidates returned by IAFIS the exemplar that would result in the most difficult comparison (18%), or the examiner selected an exemplar from the neighboring fingers from the correct subject (29%).[*] For the remainder of the non-mated pairs, the first exemplar in the list of IAFIS candidates was selected. The process of selecting challenging non-mated pairs was time-consuming and therefore was not pursued for latents that were considered to be of no value by the subject matter experts doing data selection; as a result of this, the latents in mated pairs included a greater proportion of poor-quality latents than did the non-mated pairs.

Each of the 169 examiners was initially assigned 100 image pairs. Of these 16,900 presentations, 27 were either not administered (due to inadvertent duplicate assignments excluded prior to the test) or identified by the examiner as data entry errors during the test (via a comment facility built into the test software). Some of the examiners participated in a follow-on test (a repeatability study, manuscript in preparation), which included a small number of comparable, randomized assignments; 248 test responses (from 42 examiners) from that test were included in this analysis, for a total of 17,121 presentations. Distributions with respect to latents, image pairs, and examiners are shown in SI-2.2.

After test results were received, image pairs resulting in false positive or false negative errors were reviewed as an additional check that there were no data integrity errors in the ground-truth mating of the test data; ancillary information was used in this process, including review of the multiple impressions taken at each encounter during collection of the fingerprints used in the test. All false positive errors were reviewed, as well as false negative errors on image pairs for which no examiner made an individualization decision.

1.4 Participant background survey responses

Survey responses were provided by 159 of the 169 examiners; percentages in the summary below are based on a total of 159 unless otherwise indicated.

Despite the wide variety of experience among the participants, analyses of the survey data yielded few explanations for the observed differences in examiner performance. The survey variables were highly confounded (e.g., years of experience, gender, and education were all strongly associated, with years of experience inversely related to level of education), and statistical power was limited by uneven subpopulation sizes and modest differences in outcomes among the subpopulations. Notable results with respect to certification and experience are presented in SI-5 and SI-6.

Demographics

	Count	*Percent*
1. Sex		
Female	76	48%
Male	83	52%
2. Age		
20-29	22	14%
30-39	59	37%
40-49	27	17%
50-59	37	23%
60-69	10	6%
70-79	1	1%
No response	3	2%
Minimum: 24; Median: 39; Maximum: 70; Mean: 41.8; St. Dev.: 11.3.		

[*] *An individualization of an incorrect finger from the correct subject is an error. Neighboring index, middle, or ring fingers from a subject often have similar fingerprint pattern classifications and therefore are more likely to be similar than two random fingerprints.*

	Count	*Percent*
3. Highest level of education achieved		
Less than High School Diploma	0	0%
High School Diploma/GED	7	4%
Associate Degree/some college	34	21%
Bachelor's degree	79	50%
Graduate degree/Professional degree	39	25%

Employer

4. Current employment		
Non-U.S. National Gov't.	0	0%
Non-U.S. State/Province/Regional Gov't.	0	0%
Non-U.S. City/Local Gov't	0	0%
Non-U.S. Private sector (non-Gov't.)	2	1%
U.S. Federal Gov't.	77	48%
U.S. State Gov't.	36	23%
U.S. City/County Gov't	33	21%
U.S. Private sector (non-Gov't.)	11	7%
5. National affiliation of employer (only if non-US)		
No response	154	97%
Invalid responses	5	3%
6. Has your agency received accreditation in latent prints?		
Don't know	3	2%
No	24	15%
Yes (for example by ASCLD/LAB (American Society of Crime Laboratory Directors/Laboratory Accreditation Board) or FQS (Forensic Quality Services))	132	83%

Experience

7. Total number of years employed as a 10-print examiner (Note: 10-print, not latent)		
Less than 5	25	16%
5-9	20	13%
10-19	6	4%
20-29	6	4%
30-39	1	1%
40-49	1	1%
None (80 reported 0 years; 20 no responses)	100	63%
Minimum: 0.5; Median: 5; Maximum: 43; Mean: 7.8; St. Dev.: 8.2.		
8. How long was your 10-print training? (Note: 10-print, not latent training)		
None	80	50%
Under 6 months	49	31%
6 months - 1 year	23	14%
1 or more years	7	4%

	Count	*Percent*
9. Number of years employed as a latent examiner		
Less than 5	33	21%
5-9	44	28%
10-19	38	24%
20-29	27	17%
30-39	13	8%
40-49	2	1%
No response	2	1%
Minimum: 0.25; Median: 10; Maximum: 44; Mean: 13; St. Dev.: 10.4.		
10. Type of latent training received (Check all that apply - may add to over 100%)		
No response	1	1%
Formal program of instruction	117	74%
Informal training (on the job, apprenticeship)	90	57%
Limited formal training (courses, workshops)	88	55%
11. How long was your total latent training (including initial and continuing, formal and on-the-job)?		
Under 6 months	8	5%
6 months - 1 year	21	13%
1 or more years	130	82%
12. Who provided your training? (Check all that apply - may add to over 100%)		
No response	1	1%
An agency other than my employer at the time of my training	41	26%
The agency I was/am employed with at the time of my training	150	94%
Other	22	14%
Private organization	41	26%
13. Have you been certified by a certifying authority? (Check all that apply - may add to over 100%)		
No response (i.e., not certified)	27	17%
A current or previous accredited employer	72	45%
International Association for Identification (IAI) Latent certification (CLPE)	78	48%
International Association for Identification (IAI) Tenprint certification (CTPE)	2	1%
A current or previous non-accredited employer	3	2%
National certification (non-US only)	1	1%
On question 13, responses were available for 161 of the participants (as opposed to 159 for all of the other questions), and therefore percentages are based on a total of 161; 18 of the IAI CLPE-certified examiners also indicated one or more of the other certifications (11%), among other combinations of certifications. Both of the examiners who indicated CTPE were also CLPE.		
14. Are your certifications current/active?		
Never certified	26	16%
No	1	1%
Yes	132	83%
15. Were you tested for competency in comparison as part of your certification process?		
No	5	3%
Not certified	22	14%
Yes	132	83%
16. Have you taken a proficiency test within the past 12 months?		
No	24	15%
Yes, a test from an outside source	84	53%
Yes, an internal test	21	13%
Yes, both internal and external tests	30	19%

	Count	*Percent*
17. Are you currently or have you previously been employed as a latent fingerprint examiner?		
Currently employed as a latent fingerprint examiner	153	96%
No	4	3%
Previously (but not currently) employed as a latent fingerprint examiner	2	1%
18. Are you currently conducting latent examinations on a regular basis (at least weekly over an extended period)?		
No	1	1%
No, but I have previously conducted latent examinations on a regular basis	20	13%
Yes	138	87%
19. What year did you begin conducting latent casework?		
1960-1969	3	2%
1970-1979	14	9%
1980-1989	25	16%
1990-1999	27	17%
2000-2009	84	53%
No response	6	4%
Minimum: 1965; Median: 2000; Maximum: 2009; Mean: 1996; St. Dev.: 11.2.		
20. Have you ever testified in court to a latent fingerprint identification?		
No	18	11%
Yes, more than one year ago	46	29%
Yes, within past year	95	60%
21. Are you a trainee?		
No	152	96%
Yes	7	4%
Three of the 7 "trainees" state they are IAI-Latent certified; two state they have 25 or more years latent experience. None of these seven committed a false positive error.		
22. What percentage of time have you spent over the last year doing latent comparisons?		
N/A: I am not performing comparisons	1	1%
Less than 10%	15	9%
10-25%	28	18%
25-50%	28	18%
50-75%	39	25%
75-100%	48	30%
23. Of the latent-to-exemplar comparisons you have performed over the last year, what proportion do you perform on computer screens, as opposed to looking at physical evidence/paper cards?		
N/A: I am not performing comparisons	1	1%
0% computer	18	11%
1-30% computer	79	50%
30-60% computer	27	17%
60-99% computer	25	16%
100% computer	9	6%

Known errors

	Count	*Percent*
24. Are you aware of ever having made an erroneous individualization (after training)? (Check all that apply - may add to over 100%)		
No response	3	2%
No	150	94%
Yes, on casework; detected after it was reported to contributor	1	1%
Yes, on a proficiency test only	1	1%
Yes, on casework; detected during verification	4	3%
On question 24, there was no overlap among the three categories of "Yes" answers.		
25. Are you aware of ever having made an erroneous exclusion (after training)? (Check all that apply - may add to over 100%)		
No response	2	1%
No	103	65%
Yes, on casework; detected after it was reported to contributor	10	6%
Yes, on a proficiency test only	4	3%
Yes, on casework; detected during verification	43	27%
On question 25, one examiner indicated "Yes" both on a proficiency test and on casework detected during verification. Two examiners indicated "Yes" both on casework detected after reporting and on casework detected during verification.		

Standard operating procedures

26. If you have determined that a latent is of value and have started to conduct comparisons, are you permitted to change your mind and relabel the latent as no value? (Given the standard operating procedures that you/your agency currently use)		
No	13	8%
Yes	146	92%
27. Does your organization permit an official conclusion of less than individualization, more than inconclusive, such as "limited match" or "qualified identification"? (Given the standard operating procedures that you/your agency currently use)		
No	152	96%
Yes	7	4%
28. If the latent and exemplar are both of value, include a large potentially corresponding area, no other latent or exemplars images are available, and you already have all processing information related to the latent, are you permitted to make an inconclusive determination? (Given the standard operating procedures that you/your agency currently use)		
Inconclusive determinations are discouraged but possible in this case	31	19%
Inconclusive determinations are freely accepted in this case	77	48%
Inconclusive determinations are not permitted in this case	51	32%
29. In determining the value/sufficiency of a latent impression, how do you define an impression that is not suitable for individualization but could potentially be used for exclusion? (Given the standard operating procedures that you/your agency currently use)		
It has its own category used in standard practice, such as "Of value for exclusion only" or "Limited value"	27	17%
It has its own category, such as "Of value for exclusion only" or "Limited value" – but only used upon request	21	13%
No value	88	55%
Of value	23	14%

	Count	*Percent*
30. How often in casework do you make a conclusion that a latent and the exemplars provided definitively did not come from the same source? (Given the standard operating procedures that you/your agency currently use)		
Never	5	3%
Used only on request	4	3%
Rarely	16	10%
Often	134	84%
31. How do you use the term "exclusion" as a conclusion? (Given the standard operating procedures that you/your agency currently use)		
Any comparison that is not an individualization is an exclusion	7	4%
Exclusion means that the latent did not come from any finger for that subject, but could have come from other friction ridge skin (e.g., palm) from that subject	16	10%
Exclusion means that the latent did not come from any friction ridge skin for that subject	81	51%
Exclusion means that the latent did not come from the source of the exemplar (e.g., a specific finger), but could have come from another finger from that subject	18	11%
Not used	37	23%

1.5 Instructions to participants

The following are from the instructions given to the participants. The wording of the determinations is identical to the wording in the software.

Summary of Test Procedure

In the Black Box Study, you will be asked to perform a series of 100 comparisons. For each comparison, you will go through these steps:

- You will be presented a latent image for analysis, and asked to make a determination whether the latent is of value for individualization and/or exclusion. (See the Analysis and Comparison Determinations section for detail)
- If you determine that the latent is of value, you will be presented with a single exemplar image for comparison, and asked to make a comparison determination. (See the Analysis and Comparison Determinations section for detail)
- You may optionally make image enhancements, zoom in or out, or mark features in the latent or exemplar images. These are for your benefit, and may be used or ignored at your discretion. (See the General Guidance section for detail)
- At the end of each comparison, you have a chance to review the determinations you made for that comparison before you save your results for that comparison.

Analysis and Comparison Determinations

It is imperative that you conduct your analyses and comparisons in this study with the same diligence that you use in performing casework.

This section explains the determinations you will be asked to make, with explanations of the options. Note that which determinations you are asked to make depends on your previous answers for that comparison.

- #1 — Value for individualization
- #2 — Value for exclusion (only if #1 is "No Value")
- #3 — Comparison determination (only if #1 or #2 is "Of Value…")
- #4 — Reason for inconclusive (only if #3 is "Inconclusive")
- #5 — Reason for exclusion (only if #3 is "Exclusion")
- #6 — Difficulty (only if #1 or #2 is "Of Value…")

The determinations as defined here may not correspond precisely to you/your organization's current procedures. Please try to follow the guidance in this section as closely as possible so that responses from different people/organizations are comparable.

#1 — Value for individualization

Assess the value of the latent image for individualization. You may (optionally) mark points in the latent image; these points are solely for your reference.

1a	Of value for individualization	The impression is of value and is appropriate for potential individualization if an appropriate exemplar is available.
1b	Not of value for individualization	The impression is not of value for individualization.

#2 — Value for exclusion

Note: only applies if "Not of value for individualization" (1b)

Given that the latent image is NOT of value for individualization, assess the value of the latent image for exclusion.

2a	May be of value for exclusion	The impression contains some friction ridge information (level 1 and/or level 2) that may be appropriate for exclusion if an appropriate exemplar is available.
2b	Not of value for exclusion	The impression contains no usable friction ridge information.

#3 — Comparison determination

Note: only applies if "Of value for individualization or exclusion" (1a or 2a)

Compare the latent and exemplar images. You may (optionally) mark points in the latent and/or exemplar images; these points are solely for your reference.

3a	Individualization	The two fingerprints originated from the same finger.
3b	Inconclusive	Neither individualization nor exclusion is possible.
3c	Exclusion of finger	The two fingerprints did not come from the same finger.

#4 — Reason for inconclusive

Note: only applies if "Inconclusive" (3b)

Explain the reason for the inconclusive determination.

4a	Inconclusive due to no overlapping area	Corresponding areas (or potentially corresponding areas) of friction ridge detail are absent.
4b	Inconclusive due to insufficient information	Potentially corresponding areas are present, but they contain insufficient corresponding or contradictory data. This option should be selected if the exemplar is not of value.
4c	Inconclusive, but with corresponding features noted	No conclusive determination can be made. Corresponding features are present, and no substantive contradictory features are present. The correspondence of features is supportive of the conclusion that the two impressions originated from the same source, but not to the extent sufficient for individualization.

#5 — Reason for exclusion

Note: only applies if "Exclusion" (3c)

Explain the reason for the exclusion.

5a	Pattern class/ridge flow alone	The exclusion could be made based on pattern class/ridge flow/level-1 information alone. The exclusion did not require review of minutiae and/or Level-3 information.
5b	Minutiae and/or level 3	The exclusion determination required comparison of Level-2 and/or Level-3 information.

#6 — Difficulty

Note: only applies if "Of value for individualization" or "Of value for exclusion" (1a or 2a)

How easy or difficult was it to make the comparison determination?

6a	Very Easy/Obvious	The comparison determination was obvious.
6b	Easy	The comparison was easier than most latent comparisons.
6c	Moderate	The comparison was a typical latent comparison.
6d	Difficult	The comparison was more difficult than most latent comparisons.
6e	Very Difficult	The comparison was unusually difficult, involving high distortion and/or other red flags.

Comments

For each question, a comment box is available; this is intended to allow examiners to provide comments on the test process to the test administrators. Examples include software issues, data entry errors, an exceptional image whose inclusion in the test might be inadvertent, any problems taking the test. The comment box is ***not*** intended to routinely capture your thought process in reaching conclusions: comments should be reserved for exceptional circumstances.

General Guidance

When conducting analyses and comparisons, please note:

- You may assume that the images provided are the only images available, and that physical evidence, lift cards, fingerprint cards, and additional exemplars are not available.
- You may assume that every impression is a fingerprint, not a palmprint or lower joint.
- For an inconclusive determination, it is assumed that additional exemplars would have been requested; it is not necessary to state this in a comment.
- For the "Of value…" determinations, make the determination for the latent with the assumption that a good-quality exemplar with a large corresponding area may be available. The specific exemplar shown for comparison may or may not meet these criteria, but that should have no bearing on the "Of value…" determinations for the latent.
- No images have already been claimed "Of value." In a few images, impressions were marked to indicate which impressions were to be scanned; such marks do not indicate that another examiner necessarily determined that the print is of value.
- If you believe that the exemplar is not of value, the comparison should be noted as "Inconclusive due to insufficient information" (#3b & #4b).

2 Decision rates

Comparison Decision	Latent Value	Total	Mates	Non-mates	% of mated pairs			% of non-mated pairs		
					PRES	CMP	VID	PRES	CMP	VID
(not compared)	NV	3,947	3,389	558	29.3%			10.1%		
Exclusion	VEO	486	**161**	325	**1.4%**	**2.0%**		5.9%	6.5%	
Exclusion	VID	4,072	**450**	3,622	**3.9%**	**5.5%**	**7.5%**	65.3%	72.7%	88.7%
Inconclusive	VEO	2,596	2,019	577	17.4%	24.7%		10.4%	11.6%	
Inconclusive	VID	2,311	1,856	455	16.0%	22.7%	31.1%	8.2%	9.1%	11.1%
Individualization	VEO	40	40	**0**	0.3%	0.5%		**0.0%**	**0.0%**	
Individualization	VID	3,669	3,663	**6**	31.6%	44.7%	61.4%	**0.1%**	**0.1%**	**0.1%**
Totals		17,121	11,578	5,543	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Total comparisons	Value (either)	13,174	8,189	4,985						
Total comparisons	VID	10,052	5,969	4,083						

Table S5: Decisions by 169 examiners on 17,121 presentations of 744 image pairs. False positive and false negative counts and rates are highlighted in yellow. Rates of comparison decisions are calculated as a percentage of all presentations, including latents of no value (PRES), including latents of no value (NV); of comparisons where the latent was of value for individualization (VID); or of all comparisons (CMP), including comparisons where the latent was of value for exclusion only (VEO) as well as VID. Figure 2 in the main report depicts the six rightmost columns.

Comparison Decision	VEO			VID		
	Total	Mates	Non-mates	Total	Mates	Non-mates
Exclusion	486	161	325	4,072	450	3,622
- *Minutiae and/or level 3*	*231*	*53*	*178*	*3,529*	*384*	*3,145*
- *Pattern class/ridge flow alone*	*255*	*108*	*147*	*543*	*66*	*477*
Inconclusive	2,596	2,019	577	2,311	1,856	455
- *Corresponding features noted*	*396*	*304*	*92*	*620*	*500*	*120*
- *Insufficient information*	*489*	*440*	*49*	*499*	*454*	*45*
- *No overlapping area*	*1,711*	*1,275*	*436*	*1,192*	*902*	*290*

Table S6: Reasons given for exclusion and inconclusive decisions. For specific wording, see SI-1.5, questions #4 and #5. Posterior probabilities for exclusion decisions depending on exclusion reason are shown in Fig. S8. The mated pairs included a greater proportion of poor quality prints than the non-mated pairs. Whereas poor-quality latents result in NV decisions (Table S5), the poor-quality exemplars contribute to an increase in the proportion of inconclusive decisions.

2.1 Confidence intervals

We report confidence intervals for two of our rate estimates in order to provide an indication of the reliability of these estimates. These intervals should be interpreted under the assumption that the sample data are representative of decisions by a similar group of examiners performed on similar images – not the universal population of examiners and images. Additionally, the estimate we use, the Agresti-Coull, like most other confidence interval methods, assumes independence among the decisions. Because our data include many commonalities of examiners and image pairs, we expect the confidence intervals presented here to be somewhat narrower than appropriate for the data.

Six false positives occurred among 4,083 non-mated comparisons where the latent was VID, for a false positive rate FPR_{VID} = 0.15%, with a 95% confidence interval of 0.06% to 0.3%. Similarly, 450 false negatives occurred among 5,969 mated VID comparisons, for a false negative rate FNR_{VID} = 7.5%, with 95% confidence interval of 6.9% to 8.2%.

2.2 Distributions

Image pairs were randomly assigned to examiners (through random resampling of the pool of available image pairs). This method resulted in a large variance in the number of times each pair was presented.

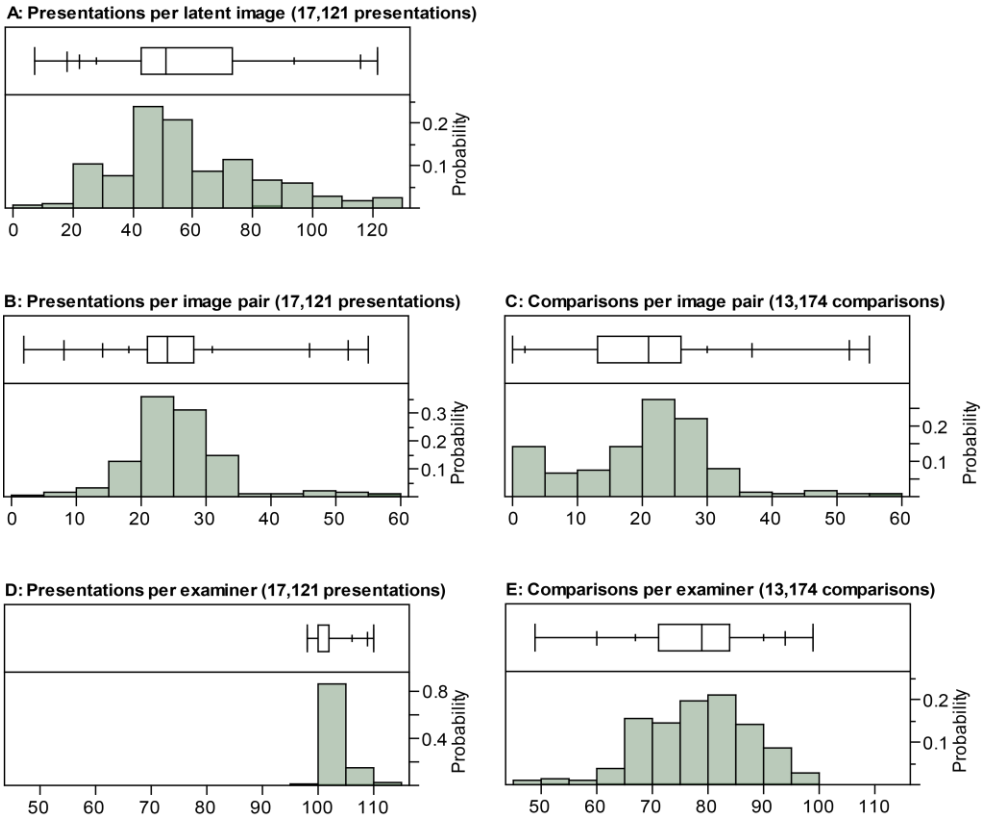

Fig. S1: A: Number of presentations of each latent image (17,121 presentations of 356 latent images). B/C: Number of presentations and comparisons (both VID and VEO) varied substantially by image pair (n=744 image pairs). D/E: Number of presentations and comparisons (both VID and VEO) of image pairs by examiner (n= 169 examiners).

3 Individual examiner error and success rates

	CR			TPR			TNR			FNR		
	PRES	CMP	VID	PRES	CMP	VID	PRES	CMP	VID	PRES	CMP	VID
Min	13%	19%	22%	11%	18%	29%	7%	8%	5%	0%	0%	0%
Median	49%	63%	79%	31%	44%	62%	74%	82%	92%	4%	6%	6%
Max	73%	94%	100%	57%	91%	94%	96%	100%	100%	24%	32%	34%
Mean	48%	63%	77%	32%	46%	61%	71%	79%	88%	5%	7%	7%
s.d.	11%	12%	11%	9%	12%	12%	14%	15%	14%	5%	7%	7%

Table S7: Summary statistics for individual examiner success and error rates.

Because false positive errors occurred infrequently (0.1% overall), and due to the limited number of non-mated pairs assigned to each examiner (26 to 53 non-mated image pairs per examiner, mean 33, s.d. 7), it would not be realistic to conclude that the actual FPR of an examiner who committed one or two false positive errors can be distinguished from the actual FPR of an examiner who committed no false positive errors. The difficulty of comparing FPRs of individual examiners is compounded by the fact the each examiner saw a different set of images.

4 False positives

Table S8 shows details for the six false positive errors (erroneous individualizations) encountered in this study, labeled A-F. The images for pairs B & C are shown in Figure 3. Pairs C & D both resulted in false positives by the same examiner; D & F involved the same latent, but the exemplars were from different subjects. The method of selection of exemplar included the rank 1 candidate from IAFIS search; examiner-selected most difficult comparison from top 20 candidates from IAFIS search;

neighboring finger from the actual subject. Latent B was processed with black powder (BP); the others were processed with cyanoacrylate and light gray powder (CA+GP) and were partially or fully tonally reversed. The highlighted table cells correspond to the responses of the examiner who committed each error. The examiner who made two false positive errors made them on consecutive comparisons.

Image pair label	A	B	C	D	E	F
Examiner certifications	Other	IAI	IAI	IAI	NA	None
Method of selection of exemplar	Selected	Neighbor	Neighbor	Rank 1	Selected	Selected
Latent processing	CA+GP	BP	CA+GP	CA+GP	CA+GP	CA+GP
Tonal reversal	Partial	None	Partial	Full	Full	Full
Latent value						
No Value (NV)						
Exclusion only (VEO)	13			1		1
Value for Individualization (VID)	13	30	26	97	22	97
Comparison decision						
Exclusion	8	24	22	21	20	21
Inconclusive (no overlap)	14	2	2			1
Inconclusive (insufficient)	2	2				
Inconclusive (corresponding)	1	1	1		1	
Individualization	1	1	1	1	1	1
Comparison difficulty						
Obvious			1	6	2	
Easy	1	3		9	10	8
Medium	14	19	14	7	9	11
Difficult	9	7	10		1	4
Very Difficult	2	1	1			

Table S8: Summary description of the 6 false positive errors. Each column describes one of the six image pairs on which a false positive error occurred. The counts indicate for each image pair the distribution of all examiners responses. The counts for latent value include all presentations of this latent image, some of which occurred in pairings with a different exemplar image. The response of the specific examiner who committed the error is indicated by highlighting the cell for that response.

5 Correlations between experience and examiner performance

Length of experience as a latent print examiner did not show a significant correlation with any of the following performance indicators: TPR_{VID}, TNR_{VID}, and FNR_{VID}.

Examiners with lower TPR_{VID} tended also to have lower TNR_{VID}. Examiners with higher FNR_{VID} tended to have lower TPR_{VID}. Examiners with higher TNR_{VID} tended also to have higher FNR_{VID}.

	Years of Latent Experience	TNR_{VID}	FNR_{VID}
TPR_{VID}	ρ: 0.017 (p-value = 0.835) $R_{x,y}$: 0.131 (p-value = 0.103) R^2: 0.00028	ρ: 0.333 (p-value < 0.0001) $R_{x,y}$: 0.286 (p-value = 0.0003) R^2: 0.111	ρ: -0.346 (p-value < 0.0001) $R_{x,y}$: -0.279 (p-value = 0.0004) R^2: 0.120
FNR_{VID}	ρ: 0.149 (p-value = 0.063) $R_{x,y}$: 0.143 (p-value = 0.075) R^2: 0.022	ρ: 0.310 (p-value < 0.0001) $R_{x,y}$: 0.374 (p-value < 0.0001) R^2: 0.140	
TNR_{VID}	ρ: 0.169 (p-value = 0.035) $R_{x,y}$: 0.165 (p-value = 0.039) R^2: 0.028		

Table S9: Correlations among examiner performance measures and years of latent experience. Three measures of association were analyzed: ρ, Pearson correlation coefficient; $R_{x,y}$, Spearman rank correlation coefficient; R^2, Coefficient of determination.

One examiner was excluded from this correlation analysis as an outlier due to the extremely low true negative rate (5.3%) and low true positive rate (31.4%).

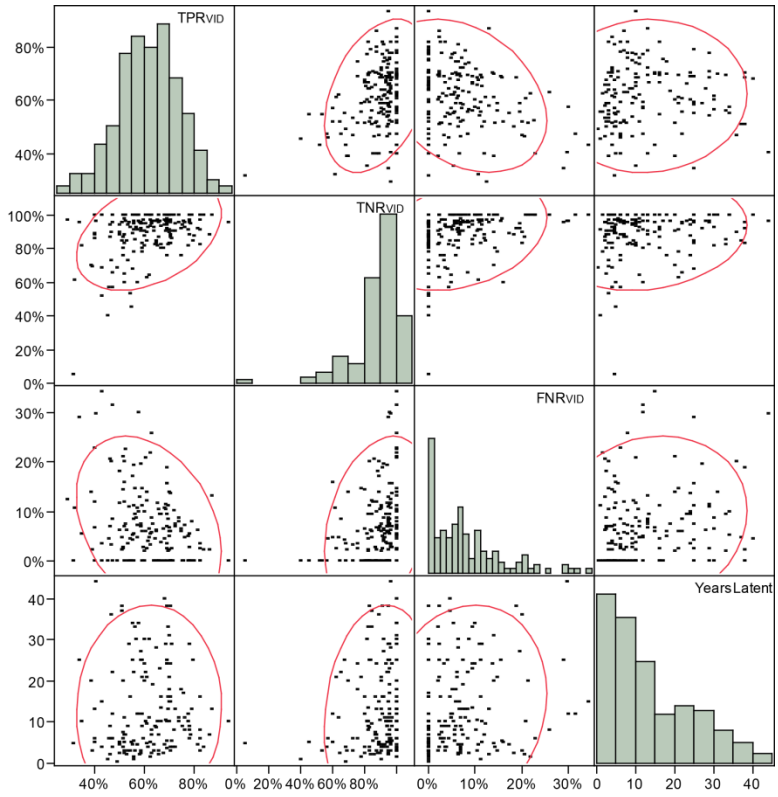

Fig. S2: Scatterplot of examiner performance measures and years of latent experience

6 Correlations between certification and examiner performance

Certified examiners had significantly higher TPR_{VID} and higher TNR_{VID} than non-certified examiners; FNR_{VID} did not vary significantly with certification.

In many cases, the distributions were skewed and assumptions of homogeneous variances, equal sample size, and normally distributed errors were not met. Several nonparametric tests of location were used to test whether performance of the groups differed. Results of the Welch ANOVA test are reported below.

	FNR_{VID}		TNR_{VID}		TPR_{VID}	
	Not certified	Certified (any type)	Not certified	Certified (any type)	Not certified	Certified (any type)
Maximum	0.343	0.316	1.000	1.000	0.793	0.935
Decile (90%)	0.210	0.179	1.000	1.000	0.667	0.780
Quartile (75%)	0.129	0.105	0.952	0.965	0.610	0.703
Median	0.056	0.062	0.842	0.929	0.548	0.645
Quartile (25%)	0.000	0.024	0.684	0.860	0.457	0.553
Decile (10%)	0.000	0.000	0.444	0.765	0.398	0.482
Minimum	0.000	0.000	0.053	0.533	0.314	0.342
Mean	0.077	0.074	0.786	0.901	0.537	0.635
Std Dev	0.086	0.070	0.221	0.099	0.104	0.114
p-value (Welch)	0.871		0.013		< 0.0001	

Table S10: Correlations between certification and examiner performance. Of the participants; 27 were not certified, and 132 were certified (including all types of certification); 10 participants who did not respond to the background survey were excluded from this analysis.

7 Variability in test results

It is important to understand both the magnitude and source of variability in the test results. A wide variability in examiner skill might motivate increased training and certification. A wide variability in latent value decisions might motivate standardization of terminology or procedures. Variability in performance on image pairs might reveal the extent to which errors are tied to image attributes vs. careless mistakes or training issues.

Even if every examiner were equally skillful, we would expect some variability in the measured examiner rates due to the random image assignments and other random factors (such as interruptions during the test). A reasonable null hypothesis for the rate distributions might be a binomial distribution. Due to the complicated test structure, a binomial distribution is not quite the right basis for this comparison: examiners were not all presented mated and non-mated image pairs in exactly the same proportion; based on latent value decisions, examiners performed a different number of comparisons; and the number of examiners to which each image pair was presented varied substantially (SI-2.2).

In order to understand how the rate distributions (by examiner and by image pair) differed from a simple binomial expectation (null hypothesis), we produced probability-probability plots of observed distributions vs. simulated results (Fig. S3). The simulations retained the exact test structure, in particular the assignment of image pairs to examiners and the latent value decisions, but replaced the actual examiner comparison decisions by a uniform random variable parameterized to the rate measured across the entire test.

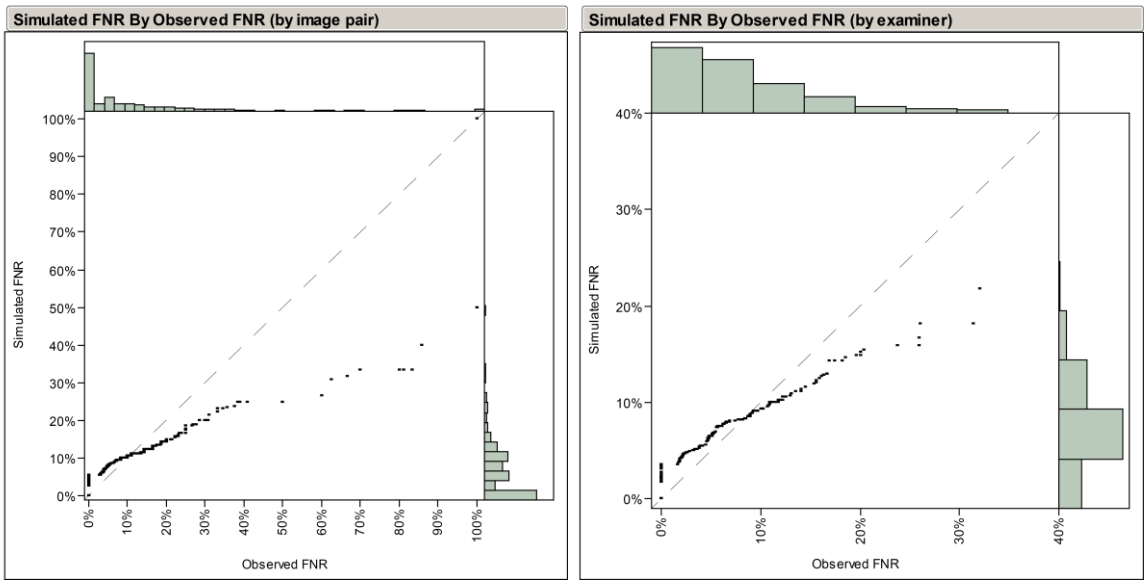

Fig. S3: Probability-Probability plots comparing the observed FNR_{CMP} to simulated FNR_{CMP}: (A) by image pair; (B) by examiner. Data is based on 17,121 presentations of 469 image pairs to 169 examiners. The simulation preserved the exact test structure (assignments of image pairs to examiners) and the actual latent value decisions, but replaced examiner comparison responses by random values with P(exclusion) = 7.5%, i.e., a simple weighted coin toss. In (A), the point at the origin represents 34% of data; another 14% lies directly above that point. In (B), the point at the origin represents 2% of the data; another 12% lies directly above that point.

The observed false negative rates have a heavier tail than the simulated rates, meaning that some image pairs (and some examiners) have a significantly higher than average rate. The observed rates have a greater concentration at zero, meaning that some image pairs (and some examiners) have a significantly lower than average rate. The model explains much of the observed variability, meaning that the true extent of variability by image pair (and by examiner) is much less than it first appears.

A similar analysis of latent value decisions (Fig. S4) reveals that many individual examiners have strong tendencies toward NV or VID decisions. Note that this analysis focuses on latent images on which value decisions were not unanimous.

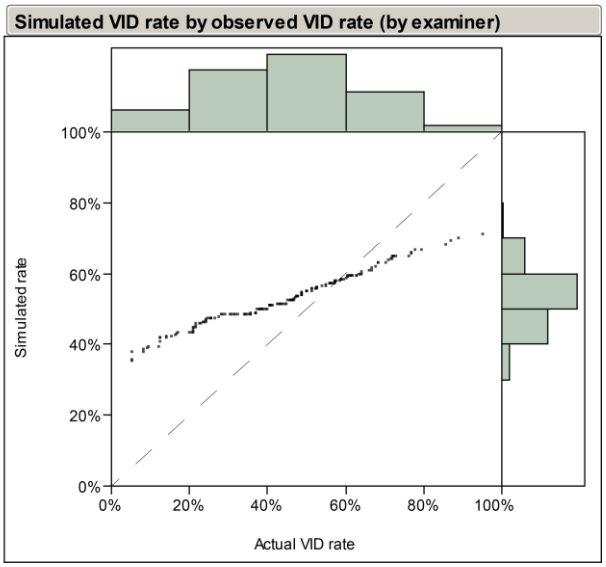

Fig. S4: Probability-Probability plot comparing the observed VID rate (proportion of latent images decided to be VID) to a simulated VID rate. The simulation preserved the exact test structure of assignments of latent images to examiners, but replaced actual VID decisions by random values with P(VID) = 53%, i.e., a simple weighted coin toss. Actual value decisions show strong individual examiner tendencies toward either NV or VID. Chart shows VID decision rates for 169 examiners based on 6,684 latent value decisions on 169 latent images on which value decisions were not unanimous.

Confidence intervals are provided in Fig. S5 as an indicator of measurement precision or reproducibility. For confidence intervals by examiner, we assumed that future measurements would involve the same examiners with similar sets of images. For FNR confidence intervals by mated image pair, we assumed that future measurements would involve the same mated pairs with similar groups of examiners.

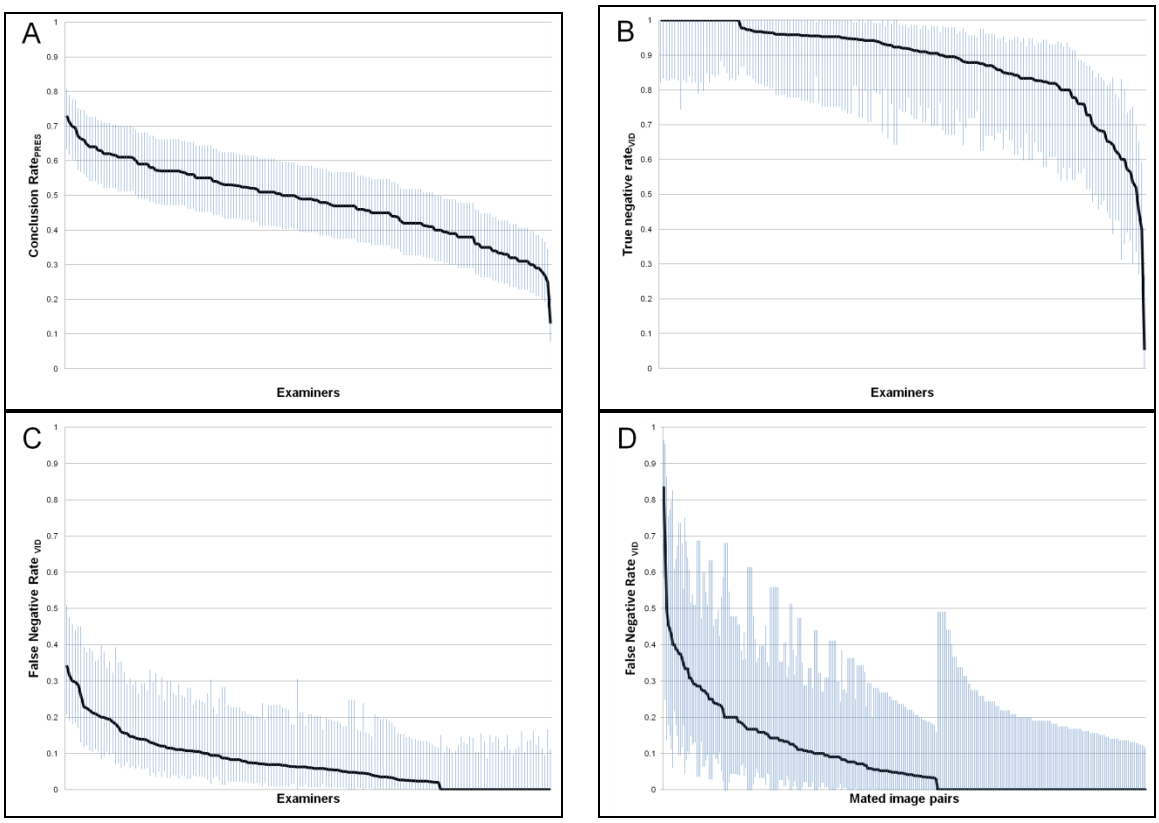

Fig. S5: (A) Conclusion rates (CR_{PRES}; 17,121 presentations of 744 image pairs to 169 examiners), (B) true negative rates (TNR_{VID}; 4,083 VID comparisons of 224 non-mated pairs by 169 examiners), and (C) false negative rates (FNR_{VID}; 5,969 VID comparisons of 520 mated pairs to 169 examiners) and their 95% Agresti-Coull confidence intervals by examiner. (D) False negative rates (FNR_{VID}; 5,969 VID comparisons of 446 mated pairs) and their 95% Wilson Score confidence intervals by mated image pair.[†] The Wilson Score interval was used due to the small sample sizes for many image pairs. 74 image pairs with 4 or fewer comparisons were excluded from this figure (D).

When asked whether latent images were VID, examiners agreed unanimously on 151/356 images. Each examiner was presented from 48 to 74 latent images on which the VID decisions were not unanimous (mean 60.6; sd 4.1).

One minor factor contributing to the lack of consensus is intra-examiner inconsistency. Within the tests, the same latent image was sometimes presented to one examiner on two different occasions (i.e., within a period of hours or days), but paired with different exemplar images. Through random assignments, 147 latent images were presented twice to the same examiner for a total of 900 observations. Examiners changed their VID decisions at a rate of 8%. Most of these were borderline cases: when the latent was not considered VID, it was considered VEO. In 10 cases (1%), a complete reversal was observed, between VID and NV. Additional data on intra-examiner variability in comparisons is being collected in a separate repeatability study (manuscript in preparation).

[†] *Image pairs in chart D are sorted first by FNR, then secondarily by confidence interval. The apparent discontinuity is an artifact of the sort order.*

8 Examiner consensus

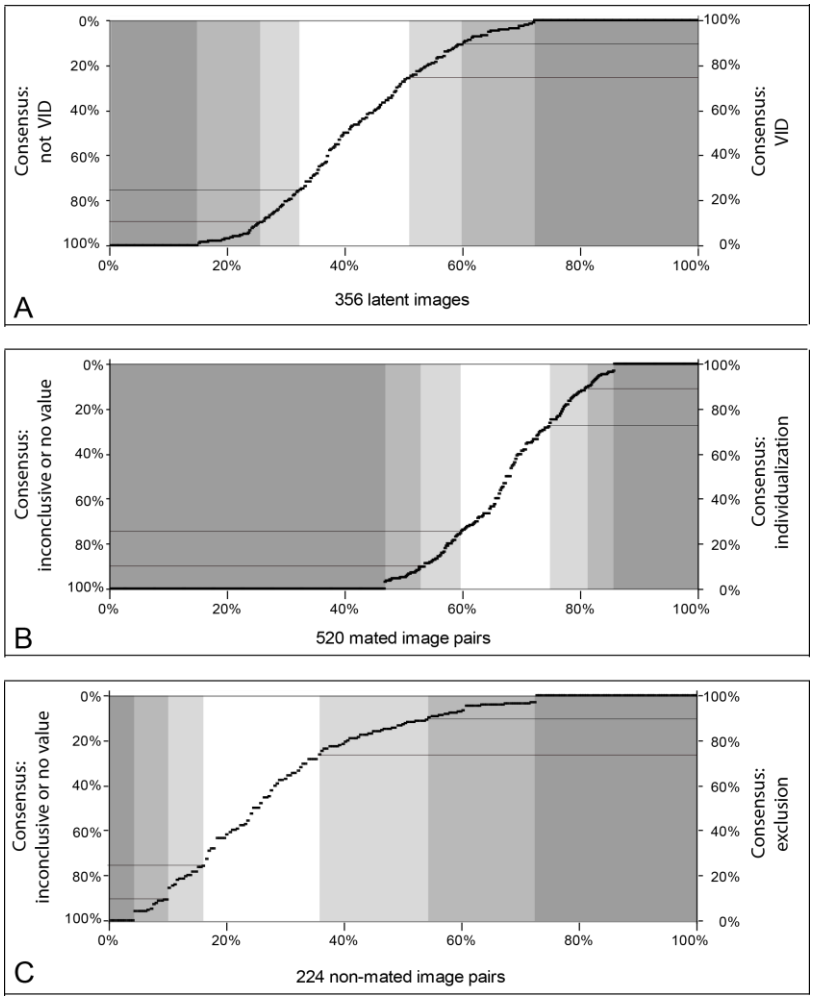

Fig. S6: Examiner consensus on decisions: (A) VID, (B) individualization of mated pairs (excluding false negatives), (C) exclusion of non-mated pairs (excluding false positives). Charts show the percentage of examiners reaching consensus (y-axis) on each image/image pair (x-axis). Areas of unanimous (100%), decile (10%,90%), and quartile (25%,75%) consensus are marked. For example, in (A), at a 90% level of consensus (y-axes), examiners agreed that 40% of the latents were VID (interval from 60% to 100% indicated by a horizontal line in upper right) (Table S11). Such measures of consensus may be useful in developing quantity and quality metrics. These distributions reflect the data selection and may be used to guide the design of future tests. Chart A is identical to Figure 6.

	Level of consensus		
	Unanimous (100% of examiners)	90% of examiners	75% of examiners
(A) VID (356 latents)	43%	66%	81%
(B) Individualization (520 mated image pairs)	61%	73%	85%
(C) Exclusion (224 non-mated image pairs)	32%	56%	80%

Table S11: Supporting data for Figure 6 and Fig. S6. Examiner consensus on decisions: (A) value for individualization (VID), (B) individualization of mated pairs (excluding false negatives), (C) exclusion of non-mated pairs (excluding false positives). Agreement includes both extremes, e.g., in (A), 100% of examiners agreed on the VID decision for 43% of latents (including both VID and not VID decisions). Data values indicate the percentage of images on which this level of agreement was reached. The extent of consensus reflects the selection of data in the study: e.g., a greater proportion of pristine prints would have increased the proportion of prints that examiners found to be of value.

9 Positive and negative predictive values

Positive and negative predictive values are functions of the prior prevalence of mated pairs among the examinations performed. These functions permit the estimation of the predictive values for other test mixes. We estimate Positive Predictive Value (PPV) as the observed rate True Positives/(True Positives + False Positives). We adjust this rate based on a prior prevalence of mated image pair comparisons performed using the following formula:

$$PPV_{MatePrevalence} = \frac{MatePrevalence * TPR}{(MatePrevalence * TPR) + (NonmatePrevalence * FPR)}$$

where

- MatePrevalence is the percentage of all comparisons that were performed on mated pairs,
- MatePrevalence + NonmatePrevalence = 100%, and
- TPR = Count of True Positives / Count of mate comparisons.

Similarly, we estimate Negative Predictive Value (NPV) as the observed rate True Negatives/(True Negatives + False Negatives). We adjust this rate based on a prior prevalence of mated image pair comparisons performed using the following formula:

$$NPV_{MatePrevalence} = \frac{NonmatePrevalence * TNR}{(NonmatePrevalence * TNR) + (MatePrevalence * FNR)}$$

where

- TNR = Count of True Negatives / Count of non-mate comparisons.

If comparisons are performed in a context where mated pairs are common, false negatives will be relatively more common: PPV will increase, and NPV will decrease. Conversely, if non-mate pairs are more common, false positives will be relatively more common, NPV will increase, and PPV will decrease. The prior prevalence of mated pair comparisons varies substantially among organizations, by case type, and by how candidates are selected. Mated pair comparisons are far more prevalent in cases where the candidates are suspects determined by non-fingerprint means than in cases where candidates were selected by an AFIS. An AFIS search is much more likely to return mated candidates for highly recidivist crimes, such as bank fraud, than when screening job applicants or travelers at border crossings.

Some sources use the term "false discovery rate," which is equal to 1-PPV (*6*).

10 Exclusion decisions

Exclusion decisions were more often correct when the latent image was VID ($NPV_{VID,59\%}$=89%) than when the latent image was VEO ($NPV_{VID,59\%}$=67%) (Fig S7D). This is not due to a difference in error rate on mated pairs, but primarily to a difference in conclusion rates on both mated and non-mated pairs (Fig S7).

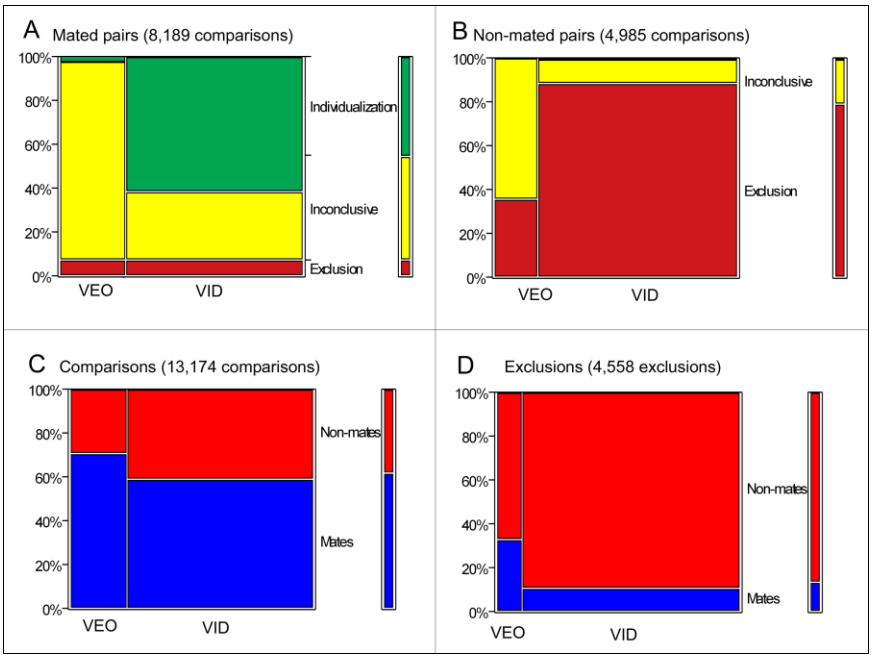

Fig. S7: Relation between latent value decisions and erroneous exclusions. The percentage of exclusions that were erroneous was 33% when latent was VEO vs. 11% when the latent was VID (D). This effect arises from differences in image pair sufficiency between mated and non-mate pairs (A and B), not from different error rates on mated pairs according to the value of the latent images (A).

After an examiner has made an exclusion decision, additional information is available to assess the posterior probability that the decision is erroneous (i.e., the image pair is mated) (Fig. S7). Note that these proportions reflect the prior probabilities inherent in this test design: 62% of test responses were on mated image pairs; and the mated pairs included a greater proportion of poor quality prints than the non-mated pairs.

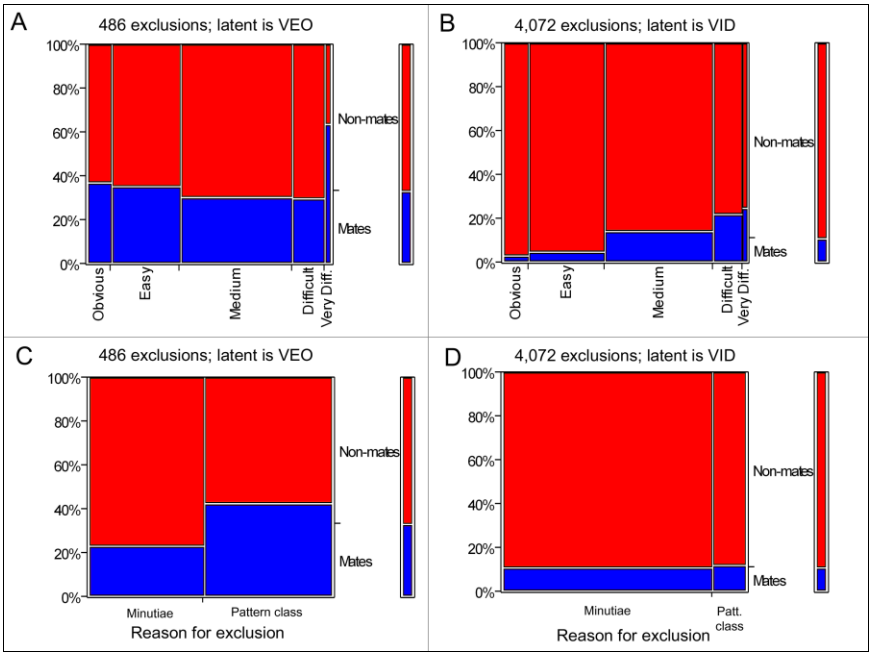

Fig. S9: Posterior probabilities for exclusion decisions depending on exclusion difficulty (from obvious to very difficult), and exclusion reason (minutiae, pattern class). NPV was much higher when latent was VID (B & D; n=4,072 exclusions) than when latent was VEO (A & C; n=486 exclusions). When latent was VID, error rate increased with comparison difficulty (B). Whether the exclusion decision was based on minutiae or pattern class was not a useful

indicator of errors when the latent was VID (D); when the latent was VEO, a greater proportion of erroneous exclusion decisions were based on pattern class (rather than minutiae) as the reason for the exclusion (C).

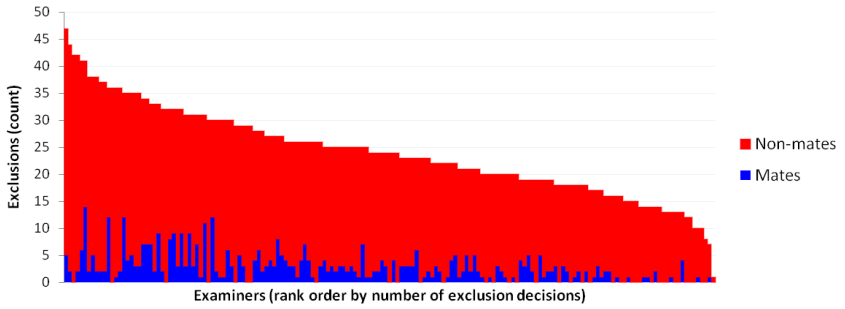

Fig. S10: Exclusion decisions, by examiner (4,072 exclusions by 169 examiners). NPV_{VID} varied by examiner from 100% to 60%.

11 Verification of exclusions

We were interested in the potential operational benefits of blind verification of exclusions. According to SWGFAST guidelines (*1*), verification is optional for exclusion decisions.

We considered the probability (p_v) that two randomly selected examiners would both exclude under the following assumptions: a mated image pair is selected at random; all casework is performed redundantly by two examiners.

The mean exclusion rate for a randomly selected mated image pair was 5.35%. This is the expected exclusion rate without verification. If exclusions were equally likely for all image pairs (independence assumption), we would estimate that exclusions by two examiners would occur at the rate p_v = 5.35% x 5.35% = 0.29%. However, the data shows that the independence assumption is not valid: some mated pairs are more likely to be excluded than others. We use the following method to revise our estimate taking into account the fact that the outcomes of blind verifications are not statistically independent, but depend on the image pairs. By this method, we estimate p_v = 0.85%.

Let p_i be the false negative rate for mated pair i. We estimate this as $\hat{p}_i = k_i/n_i$ where k_i is the observed number of false negatives on this pair, and n_i is the number of presentations of this pair. We estimate the probability that two examiners both exclude this mated pair, p_i^2 as $k_i^2/n_i^2 - (k_i/n_i * (1-k_i/n_i))/n_i$. That is, we use $(k_i/n_i * (1-k_i/n_i))/n_i$ as an estimate of variance(p_i) to correct the biased estimator k_i^2/n_i^2. We then average these estimates of p_i^2 over all mated pairs ($m=520$) in our test data:

$$\left(\sum_{i=1}^{m} k_i^2/n_i^2 - (k_i/n_i \times (1-k_i/n_i))/n_i \right) \Big/ m$$

This models a scenario where all casework is performed redundantly by two examiners and (impractically) presumes that the true mating of image pairs is known. However, it provides a rough estimate of the benefits of exclusion verifications considering the lack of independence inherent in blind verification.

12 Glossary/Acronyms

This section defines terms and acronyms as they are used in this paper.

ACE	The stages of ACE-V prior to verification: Analysis, Comparison, Evaluation
ACE-V	The prevailing method for latent print examination: Analysis, Comparison, Evaluation, Verification
AFIS	Automated Fingerprint Identification System
blind verification	The independent examination of one or more impressions by another competent examiner who has no expectation or knowledge of the conclusions of the original examiner.
CMP	Abbreviation for "comparisons": used as a subscript to indicate that a rate (percentage) was computed as a proportion of all comparisons performed, including both those where the latent was VEO and VID (omitting NV latents).
CR	Conclusion Rate: the percentage of individualization or exclusion conclusions (as opposed to no value or inconclusive decisions)
exclusion	The comparison/evaluation decision that two fingerprints did not come from the same finger. (Note: for our purposes, this is *exclusion of source*, which means the two impressions originated from different sources of friction ridge skin, but the subject cannot be excluded, whereas *exclusion of subject* means the two impressions originated from different subjects.)
exemplar	Fingerprint from a known source
false discovery rate	The percentage of individualization decisions that are false positives (i.e., made on non-mated image pairs). Equal to 1 – PPV.
false negative	An erroneous exclusion of a mated image pair by an examiner.
false positive	An erroneous individualization of a non-mated image pair by an examiner.
FNR	False negative rate: percentage of mated image pairs that resulted in (erroneous) exclusion decisions. This rate may be computed over various subsets of the data: $FNR_{VID} = \frac{\#\text{false negatives with VID latents}}{\#\text{comparisons with VID latents}}$ {omits VEO and NV latents} $FNR_{CMP} = \frac{\#\text{false negatives with VEO or VID latents}}{\#\text{comparisons with VEO or VID latents}}$ {omits NV latents} $FNR_{PRES} = \frac{\#\text{false negatives}}{\#\text{comparisons}}$ {includes VID, VEO, and NV latents}
FPR	False positive rate: percentage of non-mated image pairs that resulted in (erroneous) individualization decisions. This rate may be computed over various subsets of the data: $FPR_{VID} = \frac{\#\text{false positives with VID latents}}{\#\text{comparisons with VID latents}}$ {omits VEO and NV latents} $FPR_{CMP} = \frac{\#\text{false positives with VEO or VID latents}}{\#\text{comparisons with VEO or VID latents}}$ {omits NV latents} $FPR_{PRES} = \frac{\#\text{false positives}}{\#\text{comparisons}}$ {includes VID, VEO, and NV latents}
IAFIS	The FBI's Integrated Automated Fingerprint Identification System
IAI	International Association for Identification
inconclusive	The comparison/evaluation decision that neither individualization nor exclusion is possible

individualization	The comparison/evaluation decision that two fingerprints originated from the same finger
latent	Latent fingerprint
mated	A pair of images (latent and exemplar) known *a priori* to derive from impressions of the same source (finger)
NRC	National Research Council of the National Academies
non-mated	A pair of images (latent and exemplar) known *a priori* to derive from impressions of different sources (different fingers and/or different subjects)
NPV	Negative predictive value: the percentage of exclusion decisions that are true negatives (i.e., made on non-mated image pairs).
NV	The impression is not of value for individualization, and contains no usable friction ridge information. See also VEO, VID.
PPV	Positive predictive value: the percentage of individualization decisions that are true positives (i.e., made on mated image pairs). Equal to 1 – false discovery rate.
PRES	Abbreviation for presentation: used as a subscript to indicate that a rate (percentage) was computed as a proportion of all presentations (including those where no comparison was performed). Includes NV, VEO, and VID latents.
SWGFAST	Scientific Working Group on Friction Ridge Analysis, Study and Technology
TNR	True negative rate: percentage of non-mated image pairs that resulted in exclusion decisions. Also known as Specificity.
TPR	True positive rate: percentage of mated image pairs that resulted in individualization decisions. Also known as Sensitivity.
true negative	The exclusion of a non-mated image pair by an examiner.
true positive	The individualization of a mated image pair by an examiner.
VEO	Decision based on the analysis of a latent that the impression is of value for exclusion only, and contains some friction ridge information that may be appropriate for exclusion if an appropriate exemplar is available. Only applies to latents deemed not VID.
verification	The final stage of ACE-V: the review and analysis by a subsequent examiner of the conclusion of an original examiner. See also Blind verification.
VID	Decision based on the analysis of a latent that the impression is of value and is appropriate for potential individualization if an appropriate exemplar is available.

References for Supporting Information Appendix

1 SWGFAST (2001) Friction ridge examination methodology for latent print examiners. Ver.1.01.

2 ANSI/NIST (2011) Data format for the interchange of fingerprint, facial & other biometric information. ANSI/NIST-ITL 1-2011 (Draft).

3 SWGFAST (2010) Standards for examining friction ridge impressions and resulting conclusions. Ver.1.0 (Draft for comment).

4. FBI Criminal Justice Information Services (1997) Wavelet scalar quantization (WSQ) gray-scale fingerprint image compression specification. Ver.3.

5. Tabassi E, Wilson C, Watson C (2004) Fingerprint image quality. NIST Interagency Report 7151.

6. Koehler JJ (2008) Fingerprint error rates and proficiency tests: what they are and why they matter. *Hastings Law J.* **59 (5)**, 1077-1110.