Final accepted manuscript, subsequently published as
Hicklin, R. A., Ulery, B. T., Ausdemore, M., & Buscaglia, J. (2020). Why do latent fingerprint examiners differ in their conclusions?. *Forensic Science International*, 316, 110542.
https://doi.org/10.1016/j.forsciint.2020.110542

Why do latent fingerprint examiners differ in their conclusions?

R. Austin Hicklin,[a] Bradford T. Ulery,[a] Madeline Ausdemore,[a] and JoAnn Buscaglia [b*]

[a] Noblis, 2002 Edmund Halley Drive, Reston, Virginia 20191 (hicklin@noblis.org, ulery@noblis.org, madeline.ausdemore@noblis.org)

[b] Research and Support Unit, Federal Bureau of Investigation Laboratory Division, 2501 Investigation Parkway, Quantico, Virginia, 22135 (jbuscaglia@fbi.gov)

* Corresponding author

Declaration of interest

R. Austin Hicklin, Bradford T. Ulery, and Madeline Ausdemore are employees of Noblis, Inc. This work was funded under a contract award to Noblis, Inc. from the FBI Laboratory (DJF-17-1200-G-0007574). The work by J. Buscaglia (employee of the FBI Laboratory) was funded by the FBI Laboratory.

Keywords: *Latent fingerprint examination; Reproducibility; Repeatability; Fingermark; ACE-V; Biometrics*

Highlights

- *Differing conclusions explained in terms of image effects and examiner effects*
- *Variability among examiners due to implicit individual decision thresholds*
- *Innovative method of comparing performance of human forensic examiners*
- *3-level conclusion scale does not precisely represent examiners' conclusions*

Abstract

Forensic latent print examiners usually but do not always reproduce each other's conclusions. Using data from tests of experts conducting fingerprint comparisons, we show the extent to which differing conclusions can be explained in terms of the images, and in terms of the examiners. Some images are particularly prone to disagreements or erroneous conclusions; the highest and lowest quality images generally result in unanimous conclusions. The variability among examiners can be seen as the effect of implicit individual decision thresholds, which we demonstrate are measurable and differ substantially among examiners; this variation may reflect differences in skill, risk tolerance, or bias. Much of the remaining variability relates to inconsistency of the examiners themselves: borderline conclusions (i.e., close to individual decision thresholds) often were not repeated by the examiners themselves, and tended to be completed more slowly and rated difficult. A few examiners have significantly higher error rates than most: aggregate error rates of many examiners are not necessarily representative of individual examiners. The use of a three-level conclusion scale does not precisely represent the underlying agreements and disagreements among examiners. We propose a new method of quantifying examiner skill that would be appropriate for use in proficiency tests. These findings are operationally relevant to staffing, quality assurance, and disagreements among experts in court.

1 Introduction

Legal outcomes can be strongly influenced by the conclusions of a single forensic examiner. The accuracy of these conclusions is obviously important, but reproducibility — the variability of these conclusions among examiners — is also a fundamental concern for the criminal justice system, and within the forensic science disciplines themselves. In order to improve current practice, it is important to understand the extent and nature of disagreements among examiners, and how they arise.

Forensic examiners compare ***latents*** (fingerprints or palmprints from crime scenes) to ***exemplars*** (prints collected from known subjects) to determine whether the latents can be attributed to specific subjects, making subjective conclusions based on their expertise. In current practice, latent fingerprint conclusions are reported categorically, as an ***identification*** (*"**ID";*** Appendix SI-1, Glossary), ***exclusion***, ***inconclusive***, or ***no value*** (if the quality of the latent print is inadequate for comparison). When using data collected under controlled conditions, we know definitively the source of each fingerprint and, therefore, some conclusions are provably erroneous: identification conclusions on fingerprints from different subjects (***nonmated image pairs***) are false positive errors, and exclusion conclusions on fingerprints from the same subject (***mated image pairs***) are false negative errors. However, even when we know the source of each fingerprint we cannot definitively state whether a given conclusion is "correct": there is no ground-truth basis for determining whether a given comparison should result in a conclusion (ID or exclusion) vs inconclusive or no value, and therefore we must rely on consensus among examiners.

Examiners usually but do not always agree on conclusions: we previously observed that a second examiner reproduced 87% of identification decisions on mated pairs, 80% of exclusion decisions on nonmated pairs, and 84% of value decisions [1]. Most differences in conclusions do not involve errors per se, but are instead disagreements regarding whether the information in the fingerprints being compared is sufficient to differentiate between certain categorical conclusions: value vs. no value, identification vs. inconclusive, or exclusion vs. inconclusive. Disagreements due to erroneous exclusion conclusions are less common, and disagreements due to erroneous IDs much less common.

Overall reproducibility rates are notably affected by data selection. Even in operational casework, some latents are clear and undistorted, and can be expected to result in unanimous or near-unanimous conclusions (i.e., few or no inconclusive or no value determinations). Conversely, extremely poor-quality latents can be expected to result in unanimous no value determinations. The proportions of such outcomes can be expected to vary among agencies due to factors such as crime types and data collection policies.

In operational casework, disagreements between examiners may or may not be detected. Procedures for laboratories in the U.S. require a second examiner to verify identification conclusions, but only some laboratories conduct verifications of conclusions other than identification. When the verifier disagrees with the original examiner, laboratories use a variety of conflict management procedures to determine which conclusion to report.

In this paper, we use data from multiple tests of experts conducting latent print comparisons to show how the reproducibility of conclusions is associated with image characteristics, examiner-specific tendencies towards certain conclusions, and the (intra-examiner) repeatability of conclusions — as well as the relation of reproducibility and errors in conclusions.

2 Methods & Materials

In order to assess why and how examiners disagree, we present results from a new dataset, and new analyses of datasets from three previously reported datasets:

- Eye-tracking (ET) dataset — In the Latent Print Examiner Eye-Tracking Study, 121 practicing latent print examiners performed 1,444 latent-exemplar comparisons, and 550 exemplar-exemplar comparisons, with an average of 32 examiners per latent-exemplar image pair. This paper presents results from analyses of the examiner's conclusions, which have not been previously published. (1KHz eye-tracking data extracted during the fingerprint comparisons is not analyzed in this report; a portion of the eye-tracking results were published in [2], and other eye-tracking results are intended for future publication.)
- Black Box (BB) dataset — We conducted new analyses of data from our Latent Print Examiner Black Box Study [3], in which 169 examiners performed 17,121 latent-exemplar comparisons, with an average of 23 examiners per image pair.
- Black Box Repeatability (BBR) dataset — We conducted new analyses of repeatability data from our Latent Print Examiner Black Box Repeatability and Reproducibility Study [1], in which image pairs from BB were subsequently reassigned to participants to assess intra-examiner repeatability of conclusions: there were 2,303 reassignments of latents (340 latents, 168 participants), and 1663 reassignments of image pairs for comparison (632 image pairs, 72 participants).

- White Box (WB) dataset — We conducted new analyses of data from our Latent Print Examiner White Box Study [4], in which 170 examiners performed 3,730 latent-exemplar comparisons, with an average of 11.7 examiners per image pair.

Use of these four datasets allows us to combine new data and reanalysis of prior data in light of the information gleaned from the new data. The image pairs in the BB dataset were selected to be broadly representative of casework, whereas the latent-exemplar image pairs in the ET dataset were specifically selected for low reproducibility of examiner conclusions (see Appendix SI-2 for details). We use analyses of the BB and BBR datasets to show overall effects and trends, and analyses of the ET dataset to focus on image pairs with low levels of agreement. See Appendix SI-3 for summaries of these datasets.

3 Image effects

Examiners generally agree on value determinations (for a given image) and comparison conclusions (for a given image pair). In the BB dataset, individual examiners agreed with the majority of examiners in about 90% of trials, or about 80% if images resulting in unanimous determinations were omitted ([1], summarized here in Appendix SI-4).

Determinations were often unanimous (even with an average of 23 examiners per image pair in the BB dataset): 31% of latents were unanimously assessed as Value for ID (VID) and 19% were unanimously not VID (i.e., 50% resulted in varying levels of disagreement over value assessments); all examiners reached a comparison conclusion (identification or exclusion) on 18% of image pairs, and no examiners reached a conclusion (i.e. no value or inconclusive) on 22% of image pairs. However, unanimous conclusions are less likely when image pairs are assigned to more examiners: the ET dataset includes seven image pairs that resulted in unanimous conclusions when assigned to 9-15 examiners each in the Latent Print White Box Study [4]; in the ET study, each of these image pairs was assigned to an additional 25-34 examiners — and none remained unanimous.

The quality of the latent print is strongly associated with the proportion of examiners who assess the latent as VID, and with the proportion of examiners who reach a conclusion [5]. Using the LQMetric latent quality algorithm [6] as an objective quality metric on the BB dataset, of the latents in the highest quality quartile (LQMetric > 66), 97% of trials were assessed as VID, and 81% of trials resulted in conclusions during comparison; of the latents in the lowest quality quartile (LQMetric < 14), 21% of trials were assessed as VID, and 19% of trials resulted in conclusions during comparison. Unanimity is generally associated with the highest and lowest quality latent prints: 71% of image pairs that resulted in unanimous ID or exclusion conclusions were on latents in the highest quality quartile, and 52% of image pairs that resulted in no ID or exclusion conclusions were on latents in the lowest quality quartile. Therefore, the images and image pairs that result in disagreements among examiners are disproportionately of mediocre quality. (Details in Appendix SI-5.1.)

On review of the image pairs that resulted in the lowest rates of reproducibility, the latents generally have ambiguous ridge detail and/or discontinuous ridges. The ridge detail that is present has low specificity (i.e., is not particularly distinctive); several of the latents lack a clear focal area, such as a core or delta. Several of the latents contain (or appear to contain) multiple superimposed impressions. Several of the latents are either very light with low contrast, or contain dark, low-contrast areas. Most of the exemplars are clear, but in a few, the exemplar ridge detail is ambiguous. Examples of image pairs that resulted in notably low levels of reproducibility are included in Appendix SI-5.2.

4 Examiner effects

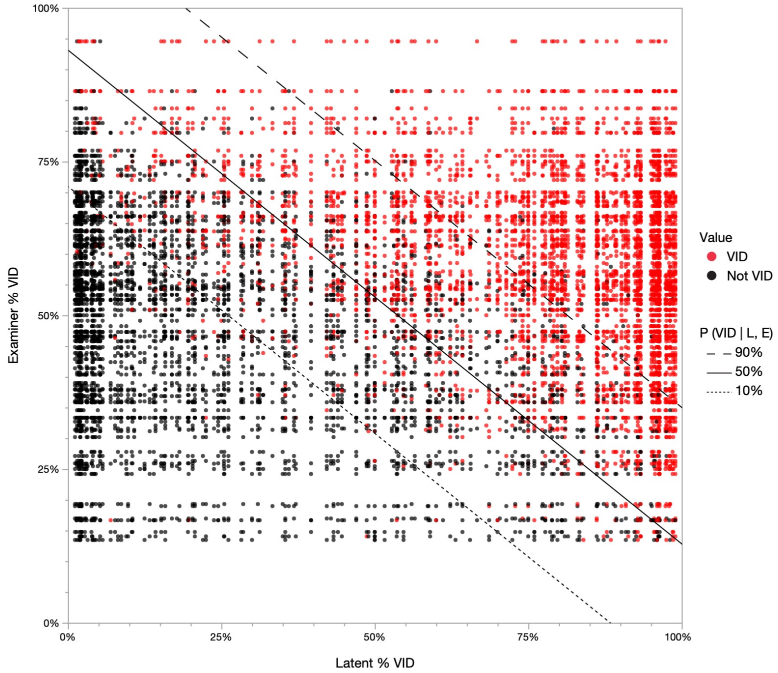

Fig. 1. Effect of latents and examiners on Value for ID (VID) assessments of latents: for each trial, Latent %VID (x axis) is the proportion of examiners who rated a given latent VID; Examiner %VID (y axis) is the proportion of non-unanimous latents that a given examiner rated VID; the colors represent the individual decisions. The diagonal lines represent {90%, 50%, 10%} probabilities of VID decisions as predicted by logistic regression on these two measures (Latent %VID, Examiner %VID). BB dataset, limited to trials where latent value determinations were not unanimous with respect to VID (56% of all trials). (n=9,552 assessments of 203 latents by 169 examiners)

Examiners vary in their tendencies regarding whether to reach specific determinations over others. Fig. 1 and Fig. 2 show how individual examiners' decisions are related to the collective assessments of latent value and sufficiency for comparison conclusions. Fig. 1 plots each value decision from the BB dataset in terms of that latent's VID rate (the proportion of all examiners who assessed that latent as VID), and that examiner's VID rate (the proportion of all latent prints that a given examiner assessed as VID). Each row corresponds to an examiner who participated in the study (sometimes superimposed or overlapping); each column corresponds to one of the latent prints presented for comparison in the study (limited to latent prints whose responses resulted in non-unanimous decisions; see discussion in Appendix SI-6.3). The diagonal lines represent probabilities of decisions as predicted by logistic regression based on the latent (L) and examiner (E) rates (P(VID|L,E)), performed on a leave-one-out basis: the outcome for each trial was omitted when calculating the two rates for that trial. Fig. 1 shows a wide variation in examiner VID rates (y axis), and suggests the presence of implicit individual decision thresholds: the x axis uses "votes" among multiple examiners (the relative proportions of determinations) to quantify the amount of useful information in each of the images; we interpret the examiners' individual assessments as individual estimates of whether the available information exceeded the examiners' implicit decision thresholds required to make each VID decision.

Because examiners' VID assessments agreed with the majority of other examiners 81% of the time (after omitting images resulting in unanimous determinations), a decision boundary at x=50% correctly classifies 81% of the decisions. The chart also demonstrates the effects of individual examiner tendencies (for example, examiners at the top and bottom of Fig. 1 disagreed with 90% or more of the other examiners). Many of the disagreements among examiners can be attributed to these individual tendencies: although predictions based only on the latent %VID rates result in a misclassification rate of 19%, taking both the latent and examiner VID rates into account results in a misclassification rate of 14% (the P(VID|L,E)=0.5 diagonal line correctly predicts 86% of the decisions).

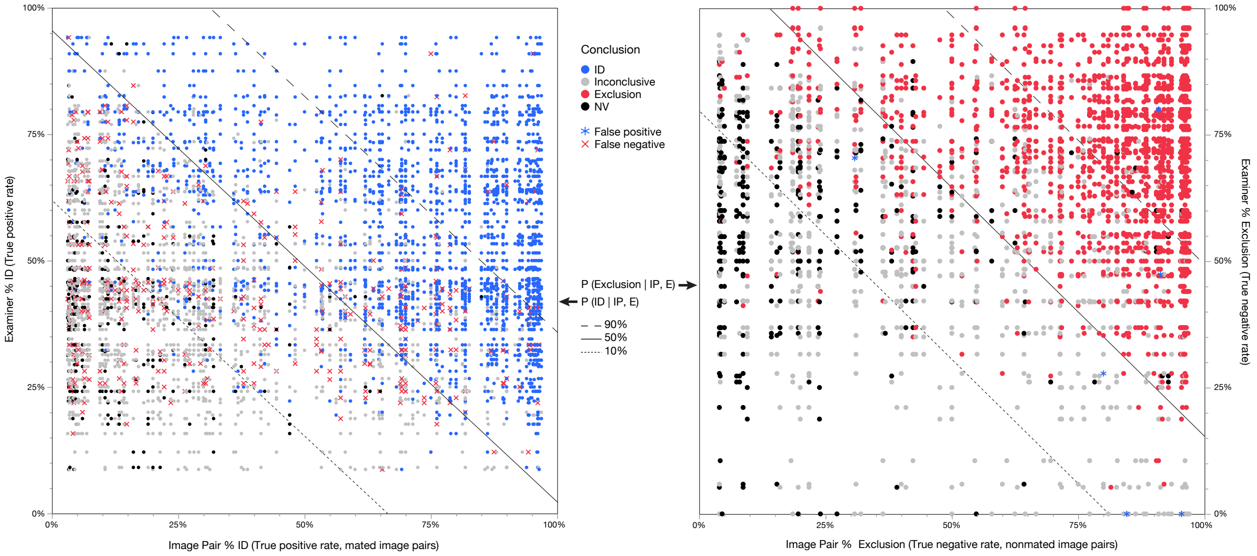

Fig. 2. Effect of image pairs (IP) and examiners (E) on ID and exclusion conclusions. Left: on mated image pairs, Image Pair True Positive Rate (x axis) is the proportion of examiners who made an ID conclusion on a given mated image pair; Examiner True Positive Rate (y axis) is the proportion of non-unanimous mated image pairs on which a given examiner concluded ID. Right: On nonmated image pairs, Image Pair True Negative Rate (x axis) is the proportion of examiners who made an exclusion conclusion on a given nonmated image pair; Examiner True Negative Rate (y axis) is the proportion of non-unanimous image pairs on which a given examiner concluded exclusion. The diagonal lines represent {90%, 50%, 10%} probabilities of ID or exclusion decisions as predicted by logistic regression on image pair and examiner rates). BB dataset, limited to trials where comparison conclusions were not unanimous (45% of mated trials; 69% of nonmated trials). (Left: n=5199 assessments of 224 mated image pairs by 169 examiners; Right: n=3845 assessments of 154 nonmated image pairs by 169 examiners)

Fig. 2 shows the results for analogous models predicting (left) ID vs. non-ID comparison decisions on mated image pairs, and (right) exclusion vs non-exclusion comparison decisions on nonmated image pairs. As in Fig. 1, each column corresponds to a specific image pair, and each row corresponds to a specific examiner. These models yield similar results to the VID model. Predicting that a given examiner would make an ID conclusion on a given mated image pair given that image pair's ID rates alone (i.e., based on the majority of other examiners' conclusions) has a misclassification rate of 20%; this reduces to 18% using the logistic regression model incorporating both image pair and examiner rates. The analogous model predicting exclusion on a given nonmated pair for a given examiner reduces misclassification from 20% to 15% (details in Appendix SI-6.2).

These results show that much of the disagreement in examiners' determinations can be attributed to subjective variation from examiner to examiner in what we describe as an implicit individual decision threshold, which can be quantified using the collective opinion of many examiners as a reference metric. This measure (percentage of examiners making a given determination) avoids the need to separately account for factors such as number of minutiae, distortion, and contrast. This examiner-specific implicit threshold defines for that examiner what constitutes a sufficient basis for a given determination (value vs. no value, ID vs. inconclusive, or exclusion vs inconclusive).

5 Borderline decisions and comfort zones

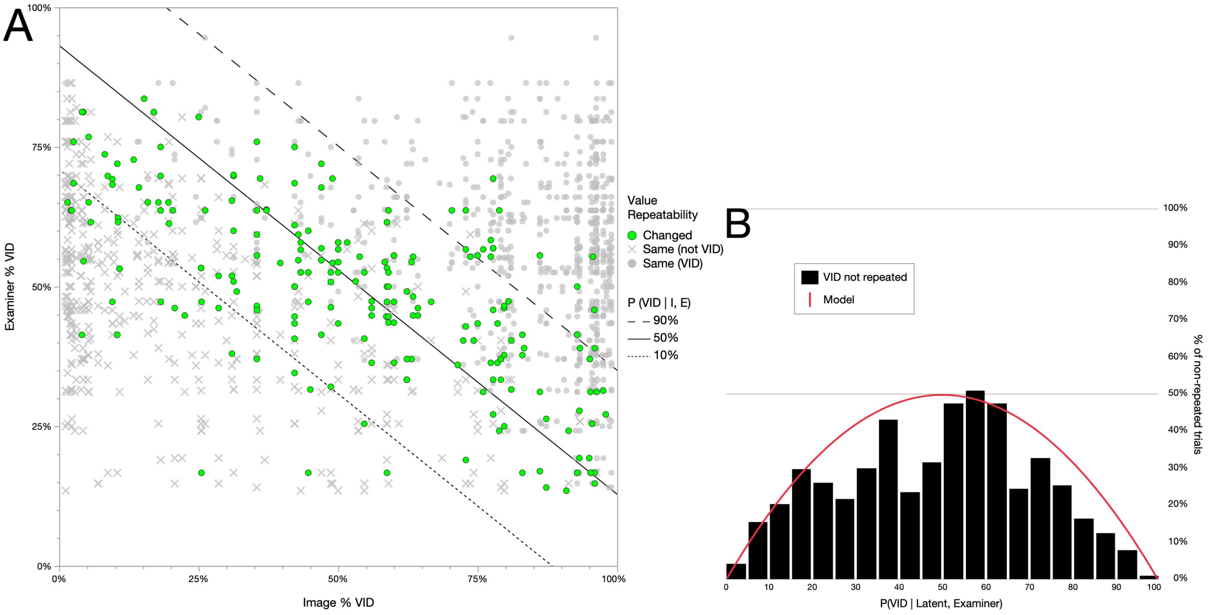

Fig. 3. Repeatability as a function of predicted VID determination. (A) Effect of image and examiner rates on repeatability of value determinations, on the subset of data in Fig. 1 for which examiners were presented the same image twice. Green points indicate instances in which the initial decision was not repeated on an image's second presentation to an examiner. (B) Black: Observed proportions of latent value decisions that were not repeated by the same examiner when retested at a later date as a function of the P(VID | L,E) model. Red: Expected proportions of non-repeated latent value decisions as a function of the P(VID | L,E) model. (BBR data: n=2754 trials with 2 trials for each latent-examiner combination (1377 paired assessments) of 196 latents by 164 examiners, limited to non-unanimous data; 40% of trials were unanimous with respect to VID).

Examiners do not always repeat their determinations when retested at a later date, and the specific images for which examiners do not repeat their determinations tend to be the same images that are associated with low reproducibility, as demonstrated in [1]. Here, we refine that observation by showing that low repeatability tends to occur on borderline decisions, near examiner-specific decision thresholds. Fig. 3 demonstrates that much of the uncertainty in the predictions of the model presented in Fig. 1 can be accounted for by the lack of repeatability in examiner decisions. Fig. 3A shows in green the instances in which the initial decision was not repeated on the second presentation of that image to that examiner (i.e., VID changed to not VID, or not VID changed to VID). Fig. 3B summarizes this result, showing that the observations in Fig. 3A nearly match what would be expected if lack of repeatability were entirely explained by this simple regression model. Each column in Fig. 3B summarizes the repeatability of trials within a five percent interval of probability that the trial would result in a VID assessment, where the probabilities were conditioned on both the latent and the examiner; the height indicates the percentage of trials on which examiners changed their value assessments. The curve in Fig. 3B indicates the rate at which examiners would be expected to change their assessments based on the regression model, $1-[P(VID|L,E)^2+P(\overline{VID}|L,E)^2]$, with the squared terms corresponding to the probabilities of repeated outcomes. For example, if the probability of an examiner assessing a print as VID is 0.90, then the probability that the examiner does not repeat the assessment is calculated as follows: the probability of repeating a VID assessment on the second presentation of the image to the examiner is 0.90^2, the probability of repeating a not-VID assessment on the second presentation of the image to the examiner is 0.10^2, and the overall probability of ***not*** repeating the initial assessment is $1\text{-}(0.90^2+0.10^2)=0.18$. Beyond demonstrating a relation between our model of examiner-specific decision thresholds and repeatability, this result indicates that most of the uncertainty in our predictions of examiner decisions can be accounted for by examiners failing to repeat their own initial assessments. A similar relationship can be observed for repeatability as a function of predicted identification and exclusion decisions, but with much smaller sample sizes and therefore a less conclusive result (Appendix SI-7.1). Nevertheless, the interpretation remains the same: as examiners become less confident in their assessments of latent prints or pairs of prints, the probability of repeating a decision based on that assessment decreases.

Examination times and assessments of difficulty are associated with these examiner-specific probabilities of conclusions: borderline decisions (near personal thresholds) are associated with longer task durations, and increased examiner assessments of difficulty. Highly probable decisions (safely within the examiners' individual "comfort zones") are associated with faster task durations and decreased examiner assessments of difficulty. Rapid completion of Analysis is associated with highly probable determinations (i.e., rapid VID assessments with high P(VID|I,E), and rapid non-VID assessments with low P(VID|I,E)); Comparison durations show similar but weaker associations with highly probable conclusions (details in Appendix SI-7.2). Examiners' assessments of comparison difficulties were found to be strongly inversely related with the examiner-specific probabilities of conclusions: as the probability of a decision increased, the reported difficulty decreased. For mated image pairs, as the probability of identification increased, examiners tended to describe their ID decisions as easier, and their inconclusive decisions as more difficult. Similarly, for nonmated image pairs, as the probability of exclusion increased, examiners tended to describe their exclusion decisions as easier, and their inconclusive decisions as more difficult. For exclusion (and especially borderline exclusion) decisions, increased difficulty was associated with an increase in errors (details in Appendix SI-7.3).

6 Disagreement as a function of categorical answers

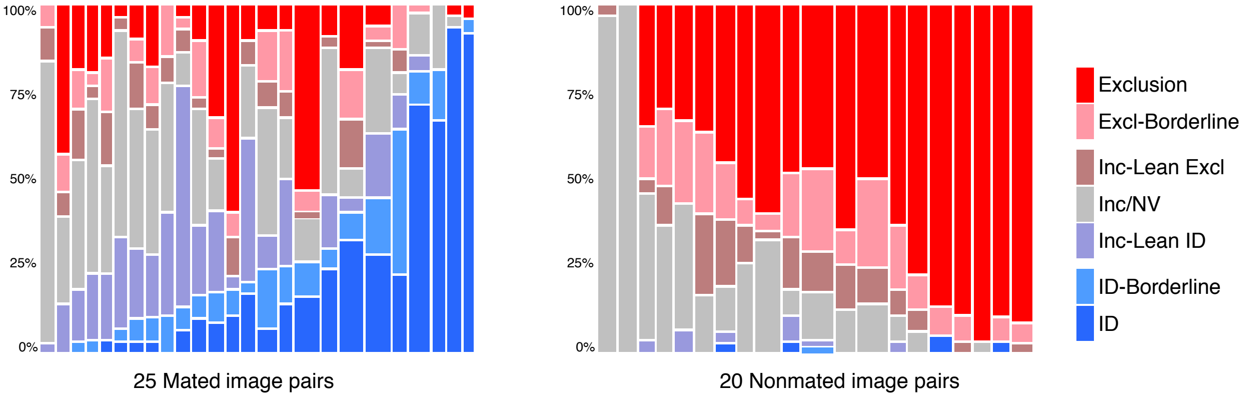

Fig. 4. Extent of disagreement, by conclusion subcategories. Each column summarizes responses on one image pair, shown sorted by responses with width proportionate to the number of responses; multiple colors in a single column indicate the mix of responses by all examiners on that image pair. (ET dataset: n=804 mated and 640 nonmated latent-exemplar trials).

The U.S. latent print examination community has proposed (but has not yet adopted) a conclusion standard that uses a five-level conclusion scale instead of the current three-level conclusion scale [7]. The psychology literature has shown that the use of categorical conclusion scales, and the number of categories in those scales, have notable effects on the measurement of reproducibility (see e.g. [8, 9]). Here we observe the specific effects of categorical conclusions on measured results. In the ET study, examiners selected subcategories within the typical three-level conclusion scale (ID, inconclusive, exclusion), rendering a seven-level conclusion scale (Fig. 4). Measuring reproducibility of conclusions using the three-level scale, examiners agreed exactly on 59% of comparison conclusions, and were diametrically opposed on 7% of comparison conclusions; using the seven-level scale, examiners agreed exactly on only 40% of comparison conclusions, and were diametrically opposed on only 3% of comparison conclusions (reproducibility was assessed using all pairwise combinations of conclusions made on each image pair, n=24,224 distinct pairs of conclusions). When two examiners reached different conclusions on the three-level scale, the extent of disagreement was often ***over***stated, because either or both of the examiners' decisions were borderline on the seven-level scale. Conversely, when two examiners reached the same conclusion on the three-level scale, the extent of disagreement was often ***under***stated, because either or both of the examiners' decisions were borderline on the seven-level scale. For example, if one examiner determines *inconclusive-leaning-ID*, and the other determines *ID-borderline-inconclusive*, the three-level scale treats the disagreement as inconclusive versus ID, overstating the extent of disagreement as compared to the seven-level scale; conversely, if one examiner determines *inconclusive-leaning-ID*, and the other determines *inconclusive-leaning-exclusion*, the three-level scale would consider it agreement (inconclusive), understating the extent of disagreement as compared to the seven-level scale. The three-level scale overstated disagreements in 20% of all conclusions, and understated disagreement in 23% of all conclusions. Thus, the underlying disagreements and agreements among examiners are more subtle than can be measured when using a three-level conclusion scale (Details in Appendix SI-8).

The number of levels in a conclusion scale also affect how error rates are measured. False negative rate (FNR) we measure as the proportion of mated image pairs that resulted in exclusions, which in the ET dataset is 22.1%. If FNR were to omit the *exclusion-borderline* decisions, that rate would drop to 14.6% (and TNR would drop from 67.8% to 55.0%), but if it were to include the *inconclusive-leaning-exclusion* decisions, that rate would rise to 28.2% (TNR would rise to 75.2%). Similarly, false positive rate (FPR) we measure as the proportion of nonmated image pairs that resulted in IDs, which in the ET dataset is 0.94% (limited to latent-exemplar comparisons). If FPR were to omit the *ID-borderline* decisions, that rate would drop to 0.78% (and TPR would drop from 31.1% to 22.3%), but if it were to include the *inconclusive-leaning-ID* decisions, that rate would rise to 2.19% (and TPR would rise to 46.4%). (Details in Appendix SI-3, Table S2).

7 Disagreements and erroneous conclusions

Disagreements between examiners are of particular concern when one of the conclusions is erroneous. The measured rates of erroneous conclusions are notably affected by the images selected for use in the study. In our previous studies [3, 4], erroneous IDs were rare (0.1% of nonmated comparisons in both BB and WB), and were never reproduced. The ET study, however, had a higher proportion of erroneous IDs (0.94% of latent-exemplar comparisons), several of which were reproduced. Erroneous exclusions were more common: for BB, the false negative rate (FNR) was 7.5%, and 15% of those erroneous exclusions were reproduced. The error rates for ET were much higher (22% FNR, and 36% of those erroneous exclusions were reproduced) because the ET dataset intentionally included image pairs that had previously resulted in errors in the BB and WB studies, as well as image pairs that were deemed likely to result in errors (for example, image pairs from the WB study that had very high numbers of corresponding minutiae marked, but that did not result in false positives). Some fingerprint image pairs are more likely to result in errors than others, and conditioning the selection of images on previous errors resulted in higher error rates and rates of reproduced errors. The ET dataset demonstrated that errors can be readily reproduced by reassigning images that had previously resulted in errors (as shown in [10, 11]). (See Appendix SI-9 for the images that resulted in erroneous IDs in the BB, WB, and ET studies.)

Measurement of false positive rates can be disproportionately affected by individual participants. In BB, five participants made six false positive errors, resulting in a false positive rate of 0.1% across all participants — even though each individual participant generally was assigned about 30 nonmated image pairs. If we assume that all participants are equivalent, this overall rate could be interpreted as applying equally to all participants, so that on a different test, different participants would have made erroneous IDs. However, the same overall error rates might have resulted if subsets of examiners had notably different individual error rates, and a significantly larger study would be necessary to compare the error rates of examiners who made one or two erroneous IDs to those who made no errors. Of the ten erroneous IDs made in the ET study, six were made by a single examiner (four of the six on exemplar-exemplar image pairs). If that examiner were not included, the false positive rate for latent-exemplar image pairs in the ET study would drop from 0.9% to 0.6% (and the exemplar-exemplar false positive rate would drop from 1.3% to 0%) — not a notable difference in absolute terms, but disproportionately affected by one individual.[*] Given that ***any*** erroneous ID is a serious concern, the

[*] *With such a high rate of error, we considered that the "outlier" participant may have not taken the test seriously. However, the eye-tracking data shows that this examiner performed detailed comparisons and took notably longer than the other participants on the same image pairs. The Institutional Review Board on human subject research that approved this research required that the participants remain anonymous, and that all cross-references between results and identities be destroyed. This participant was not IAI certified, works for an unaccredited*

fact that a practicing latent print examiner would make multiple errors is notable. Erroneous exclusions are common enough that individuals have little effect on the overall rate, but (for example) the three examiners with the highest false negative rates made 10% of the erroneous exclusions in the ET study, or 5% of the erroneous exclusions in the BB study.

The reproducibility of erroneous conclusions is of notable concern because quality assurance in practice relies on the review or verification by a second examiner. Lack of reproducibility is somewhat desirable during verification in that it would lead to detection of errors, and an error made by the initial examiner would not be reported out by the agency or laboratory. In [3], since no erroneous IDs were reproduced, we reported "This suggests that these erroneous individualizations would have been detected if blind verification were routinely performed." Over the course of the Black Box, White Box, and Eye-Tracking studies, we have now found four image pairs that have resulted in reproduced erroneous IDs, and therefore although reproduced erroneous IDs occurred very rarely, we can no longer make a blanket statement that erroneous IDs would have been detected by blind verification. In [3], we estimated the probability of reproduced erroneous exclusions to be 0.85%; although the ET dataset has a higher proportion of reproduced erroneous exclusions, we attribute that to deliberately selecting low-reproducibility images for use in the study, and therefore have no additional data to revise the BB estimate. For more details on the image pairs resulting in erroneous IDs and demographic information about the examiners who made erroneous IDs, see Appendix SI-9.

8 Quantifying examiner skill

The examiner tendencies toward conclusions we discussed above suggest an improved method of quantifying and evaluating examiner skill. As discussed in [3], an evaluation of skill needs to be multidimensional, including not just error rates, but the examiner-specific rates of a variety of decisions that examiners make. Here we build upon the approach used in Section 4 to propose a quantifiable measure of how much more or less likely a given examiner is than the mean to make a given type of decision, along four dimensions: true positive rates (TPR, rate of IDs on mated image pairs), true negative rates (TNR, rate of exclusions on nonmated image pairs), false negative rates (FNR, rate of erroneous exclusions on mated image pairs), and overall conclusion rates (CR, rate of IDs and exclusions, as opposed to inconclusives and no values, on all image pairs). For each of these dimensions, we calculate the actual vs. expected ratio for each examiner: the examiner's actual determination rate (shown as the y axes in Fig. 2) over the determination rate among all other examiners for the same set of image pairs (analogous to the x axes in Fig. 2, but adjusted to omit the current examiner). Image pairs that resulted in unanimous conclusions are omitted from the calculations (because they do not add any additional information to differentiate among examiners). False positive rates (erroneous IDs) are not included because false positive rates for individuals cannot be measured with adequate precision on a test of this size, as discussed above.

employer, and spends less than 50% of time doing latent comparisons. (N.B. 5 participants in the ET study meet these criteria, 2 of whom made erroneous IDs.)

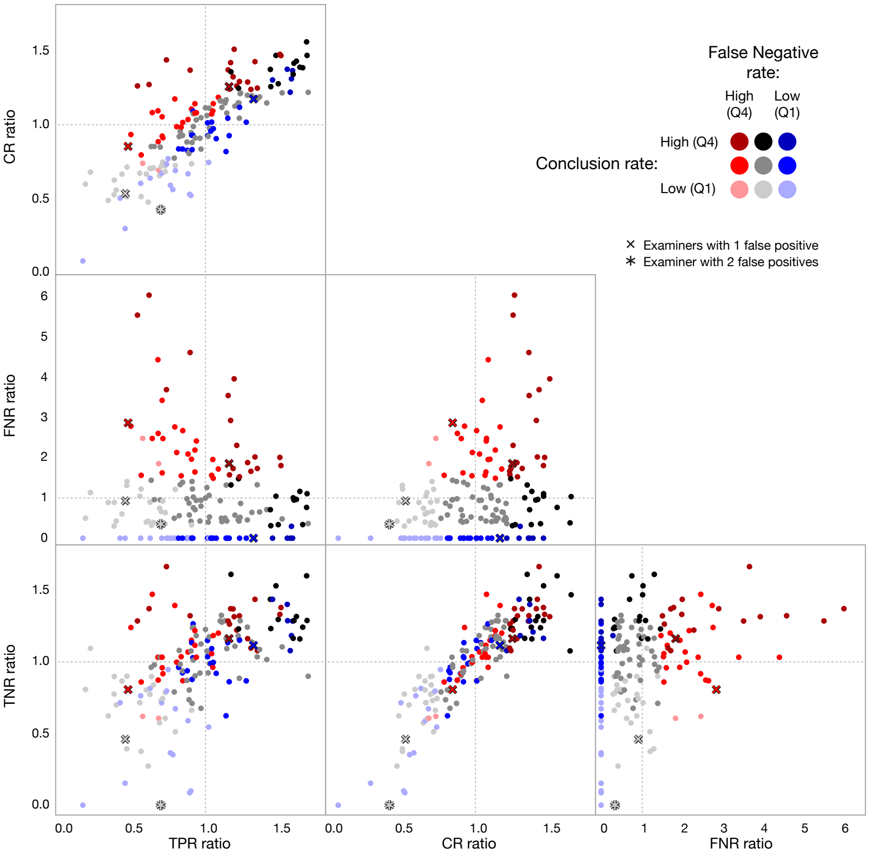

Fig. 5. Examiner tendencies toward conclusions. Each dimension indicates the relative frequency of an examiner's conclusions as compared to the other examiners on the same images. The highest and lowest quartiles of CR and FNR ratios are differentiated by color-coding. (BB data, n=169 examiners; derived from 5199 mated and 3845 nonmated non-unanimous trials)

Plotting these ratios against each other for each examiner allows us to compare these estimates of examiner tendencies using the BB dataset (Fig. 5). For each dimension, a ratio of 1.0 indicates that that examiner had the same rate as the average for the other examiners on the same image pairs, and (for example) a TPR ratio of 0.5 indicates that that examiner is half as likely as other examiners to make an ID, given the same mated pairs. For example, the examiner at the top right of the TNR-TPR chart has a TPR ratio=1.7 because that examiner had a 69% TPR on image pairs with a collective 41% TPR among other examiners; that examiner's TNR ratio=1.6 ($TNR_{examiner}$=95% / $TNR_{imagepair}$=59%), FNR ratio=0.8 (6%/8%), and CR ratio=1.6 (82%/53%). The examiner at the right of the FNR chart has TPR ratio=0.6 (30%/48%), TNR ratio=1.4 (90%/65%), FNR ratio=6.0 (33%/5.5%), and CR ratio=1.3 (74%/58%).

Examiners with high conclusion rates (dark colors) made fewer inconclusive or no value determinations than other examiners on the same image pairs: dark blue points indicate examiners with high conclusion rates and low false negative rates, which may be indicative of skill; dark red points indicate examiners with high conclusion rates but high false negative rates, which may be indicative of a tendency toward making conclusions at the risk of error. Pale colors indicate examiners who made fewer conclusions, valid or erroneous, which may be indicative of an excess of caution; the effectiveness of these examiners may be debatable. The examiner who made two erroneous IDs ("*") made unusually few exclusions (other than one false exclusion, that examiner only made exclusions on image pairs that were unanimously excluded).

These measures are affected by the other participants and image selection. An approach similar to this may be considered for broad-based proficiency tests, in which the pool of participants would be large and a balanced pool of samples would be consistently assigned to participants. This approach would provide a means of evaluating each examiner in comparison to examiners as a whole, using multiple dimensions of skill.

9 Conclusions

Why do examiners reach different conclusions on the same comparisons? We attribute these differences to several factors:

- *Image effects* — The extent of reproducibility is strongly associated with the specific images being compared. Although examiners usually reach the same conclusions, especially when comparing prints with very low or very high quality, some images are associated with low reproducibility of conclusions, and a small subset of images are associated with reproducibility of erroneous exclusions or erroneous IDs.
- *Examiner effects* — Examiners vary in what constitutes a sufficient basis for a given conclusion. This variation in implicit decision thresholds leads to a lack of reproducibility among examiners, particularly on marginal-quality prints.
- *Borderline decisions using categorical conclusion scales* — When examiners are forced to make categorical decisions near their own decision thresholds, they are often inconsistent — indeed, most of the non-reproduced conclusions not explained by image and examiner effects can be attributed to a lack of repeatability of the examiners' own decisions.
- *Granularity of categorical conclusion scales* — The measurement of reproducibility is affected by the specific conclusion scale used. We show how the use of a three-level categorical comparison conclusion scale (ID, inconclusive, exclusion) may overstate or understate the reproducibility of conclusions, when compared to a seven-level scale.

We see the latent print examination process as composed of three fundamental underlying decisions: whether the latent is of value for comparison (value vs. no value), whether there are sufficient differences to conclude the images are from different sources (exclusion vs. inconclusive), and/or whether there are sufficient similarities to conclude the images are from the same source (ID vs. inconclusive). One might expect that each of these decisions would have a generally accepted decision threshold used as a *de facto* standard, but this is not supported by the results. The examiner effects demonstrate the role of subjectivity in making these decisions, in the form of implicit examiner-specific thresholds. Note that we are not suggesting that examiners have explicit criteria used to define these thresholds; possibly not all criteria are even articulable. This subjectivity may result from a variety of factors, such as differences in skill and experience, variation in interpretation of the features within the images, lack of rigorous standards regarding the quality and quantity of features necessary for making each determination, differences in caution or risk aversion, or bias/preference for some conclusions over others. Making a borderline decision — close to that examiner's threshold — necessitates conscious or unconscious tradeoffs that can be described as a cost function: what does this examiner internalize as the relative benefit of a correct conclusion, versus the cost of an error, versus the cost of not making a conclusion? Costs and benefits could be considered in terms of the personal or professional cost/benefit to the examiner, the cost/benefit to the agency, or the cost/benefit to society. Standardizing conclusions would not just have to be defined in terms of features, but in terms of more explicit definition of these cost/benefit tradeoffs.

In practice, not all disagreements are substantive. For example, if disagreements arise within a laboratory during a conflict-resolution process, the examiners may agree that one conclusion was based on a mistaken interpretation, or come to a common agreement on how to report a borderline conclusion. Differences in skill are likely to result in some ongoing variability that can be accommodated if appropriate quality assurance procedures are used. Instances in which a single conclusion holds sway may be of especial concern, particularly for laboratories in which no value or inconclusive determinations are never verified.

We see several practical near-term steps that may mitigate some of the concerns raised here:

- *Revise the conclusion scale* — The effect of borderline decisions and the conclusion scale point to a common underlying issue: the three-level conclusion scale is not a precise representation of the examiners' underlying assessments when making comparisons — particularly for borderline conclusions. Representing the continuum of decision space as a categorical determination results in discretization error, which we see here as imperfect reproducibility and repeatability on borderline decisions. Imperfect repeatability and reproducibility rates should not necessarily be taken as a criticism of the examiners, but a criticism of the system: for the subset of comparisons that have no clear answer within the current limited conclusion scale, it is unreasonable to expect the answers to be consistent. The latent print discipline should consider the use of an approach to conclusions that better represents the subtleties in the examiners' decisions. Any revised approach to conclusions, however, should be carefully evaluated with respect to its effects on the accuracy, repeatability, and reproducibility of examiners' conclusions.

- *Improve proficiency testing* — The differences among the examiners in these anonymous studies raise the question of whether the examiners themselves, or their employers, know their abilities. The examiner who made six erroneous IDs on the ET study is (or claims to be) a practicing latent print examiner — as are the number of examiners who demonstrated high levels of accuracy. In BB, 65% of participants were unaware of ever having made an erroneous exclusion (during training, testing, or casework), but 72% of those same examiners made at least one erroneous exclusion on that test alone, unbeknownst to them. The obvious response is that individuals and their employers should know their capabilities through proficiency testing. However, the existing latent print proficiency tests have come under criticism, questioning their effectiveness and rigor [12]. The method of quantifying examiner skill proposed here suggests a path forward: proficiency tests could adopt these multidimensional metrics. The results of such rigorous, detailed, ongoing testing would provide laboratories with a greater understanding of the capabilities of their examiners, for targeted training and for targeted quality assurance — and would provide the broader forensic science and legal communities a more complete understating of latent print examination.
- *Limit the effects of individual conclusions* — Given the variability of examiners' conclusions, it cannot be assumed that a single examiner's conclusion would necessarily be reproduced by another examiner. For cases in which one latent print is the predominant evidence, the implications of an erroneous or debatable conclusion would be much more severe than in typical casework. In agencies that do not have a second examiner verify determinations of no value or inconclusive, inappropriate determinations will not be detected. Blind verification of all types of determinations may mitigate some of the effects of varying conclusions among examiners; this may be particularly important in cases that contain only a single latent print, or when the latents are poor quality.
- *Targeted training* — Training for examiners should include a focus on those images that result in disagreements among examiners. Individual agencies can collect examples of such images from casework, based on disagreements that were detected by verification. Although ground-truth images would be preferable, it is a significant effort to prepare and identify those image pairs that are likely to result in disagreements, which may be too burdensome for individual agencies.

10 Acknowledgments

We thank the latent print examiners who participated in these studies. This is publication number 20-67 of the FBI Laboratory Division. Names of commercial manufacturers are provided for identification purposes only and inclusion does not imply endorsement of the manufacturer or its products or services by the FBI. This work was funded under a contract award to Noblis, Inc. from the FBI Laboratory, with additional funding from the FBI CJIS Division Biometric Center of Excellence and the National Institute of Justice. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the U.S. Government.

References

[1] B. T. Ulery, R. A. Hicklin, J. Buscaglia and M. A. Roberts, "Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners," *PLoS ONE,* vol. 7, no. 3, 2012.

[2] R. A. Hicklin, B. T. Ulery, T. A. Busey, M. A. Roberts and J. Buscaglia, "Gaze behavior and cognitive states during fingerprint target group localization," *Cognitive Research: Principles and Implications,* vol. 4:12, 2019.

[3] B. T. Ulery, R. A. Hicklin, J. Buscaglia and M. A. Roberts, "Accuracy and reliability of forensic latent fingerprint decisions," *Proceedings of the National Academy of Sciences,* vol. 108, no. 19, pp. 7733-7738, 2011.

[4] B. T. Ulery, R. A. Hicklin, M. A. Roberts and J. Buscaglia, "Measuring what latent fingerprint examiners consider sufficient information for individualization determinations," *PLoS ONE,* vol. 9(11): e110179, 2017.

[5] B. T. Ulery, R. A. Hicklin, M. A. Roberts and J. Buscaglia, "Factors associated with latent fingerprint exclusion determinations," *Forensic Science International,* Vols. 275:65-75, 2017.

[6] N. D. Kalka, M. Michael Beachler and R. A. Hicklin, "LQMetric: A Latent Fingerprint Quality Metric for Predicting AFIS Performance and Assessing the Value of Latent Fingerprints," *Journal of Forensic Identification (in press),* 2020.

[7] Friction Ridge Subcommittee, OSAC, "Standard for Friction Ridge Examination Conclusions (OSAC Proposed Standard; v. 1.0, June 2018)," [Online]. Available: https://www.nist.gov/system/files/documents/2020/03/23/OSAC%20FRS%20CONCLUSIONS%20Document%20Template%202020_Final.pdf.

[8] D. V. Cicchetti, D. Showalter and P. J. Tyrer, "The effect of number of rating scale categories on levels of interrater reliability: A monte carlo investigation," *Applied Psychological Measurement,* Vols. 9, 31-36. doi:10.1177/0146621685009001, 2085.

[9] L. Lozano, E. García-Cueto and J. Muñiz, "Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales," *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences,* vol. 4, no. 2, p. 73–79, 2008.

[10] C. Neumann, C. Champod, M. Yoo, T. Genessay and G. Langenburg, "Improving the Understanding and the Reliability of the Concept of "Sufficiency" in Friction Ridge Examination," National Institute of Justice Report #244231, 2013.

[11] J. Koehler and S. Liu, "Fingerprint Error Rate on Close Non-Matches," in *American Academy of Forensic Science*, Anaheim, 2020.

[12] B. Max, J. Cavise and R. Gutierrez, "Assessing Latent Print Proficiency Tests: Lofty Aims, Straightforward Samples, and the Implications of Nonexpert Performance," *Journal of Forensic Identification,* vol. 69, no. 3, pp. 281-298, 2019.

[13] SWGFAST, "Standards for Examining Friction Ridge Impressions and Resulting Conclusions, Version 2.0," 2013. [Online]. Available: http://www.swgfast.org/documents/examinations-conclusions/130427_Examinations-Conclusions_2.0.pdf.

[14] SWGFAST, "Individualization / Identification Position Statement, Version 1.0," 2012. [Online]. Available: http://swgfast.org/Comments-Positions/120306_Individualization-Identification.pdf.

[15] SWGFAST, "Standard terminology of friction ridge examination, Version 3.0," 2011. [Online]. Available: http://swgfast.org/documents/terminology/110323_Standard-Terminology_3.0.pdf.

[16] FBI, "Universal Latent Workstation," [Online]. Available: https://www.fbibiospecs.cjis.gov/Latent/PrintServices.

[17] FBI, "“Black Box” Study Results," [Online]. Available: https://www.fbi.gov/services/laboratory/scientific-analysis/research-and-support/black-box-study-results. [Accessed 26 05 2020].

[18] B. T. Ulery, R. A. Hicklin, M. A. Roberts and J. Buscaglia, "Changes in latent fingerprint examiners’ markup between Analysis and Comparison," *Forensic Science International,* Vols. 247(2014):54-61, 2015.

[19] B. T. Ulery, R. A. Hicklin, J. Buscaglia and M. A. Roberts, "Interexaminer variation of minutia markup on latent fingerprints," *Forensic Science International,* vol. 264, p. 89–99, 2016.

Final accepted manuscript, subsequently published as
Hicklin, R. A., Ulery, B. T., Ausdemore, M., & Buscaglia, J. (2020). Why do latent fingerprint examiners differ in their conclusions?. *Forensic Science International*, 316, 110542.
https://doi.org/10.1016/j.forsciint.2020.110542

Why do latent fingerprint examiners differ in their conclusions?

Supplemental Materials

Contents

Appendix SI-1 Glossary

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

ACE-V	The prevailing method for latent print examination: Analysis, Comparison, Evaluation, Verification.
Analysis phase	The first phase of the ACE-V method. In these studies, the examiner assessed the latent and made a value determination before seeing the exemplar print.
BB	Latent Print Examiner Black Box Study [3], and the associated dataset. Described in Section 2, *Methods & Materials.*
BBR	The black box repeatability dataset from the Latent Print Examiner Black Box Repeatability and Reproducibility Study [1]. Described in Section 2, *Methods & Materials.*
Comparison phase (Comparison/Evaluation phase)	The second and third phases of the ACE-V method. In this test, there was no procedural demarcation between the Comparison and Evaluation phases of the ACE-V method; hence, this refers to the single combined phase during which both images were presented side-by-side.
Comparison determination	The determination of individualization, exclusion, or inconclusive reached in the Comparison/Evaluation phase of ACE-V. SWGFAST [13] refers to this determination as the Evaluation Conclusion.
Conflict resolution	The process conducted when there is a difference of determinations or conclusions between examiners, generally when the initial examiner and verifier disagree.
Determination	The result of an examiner's decision: the Analysis phase results in a Value determination, and the Comparison/Evaluation phase results in a Comparison determination.
ET	Latent Print Examiner Eye-Tracking Study, and the associated dataset. Described in Section 2 (*Methods & Materials*), and Appendix SI-2 (*Latent print examiner eye-tracking study*).
Exclusion	The comparison determination that the latent and exemplar fingerprints did not come from the same finger.
Exemplar	A fingerprint from a known source, intentionally recorded.
False negative	An erroneous exclusion of a mated image pair by an examiner.

False positive	An erroneous individualization of a nonmated image pair by an examiner.
ID (Identification, Individualization)	The comparison determination that the latent and exemplar fingerprints originated from the same source. According to SWGFAST, the term "individualization" is synonymous with "identification" — both are defined as: "the decision by an examiner that there are sufficient discrimination friction ridge features in agreement to conclude that two areas of friction ridge impressions originated from the same source. Individualization of an impression to one source is the decision that the likelihood the impression was made by another (different) source is so remote that it is considered as a practical impossibility." [14, 15, 13] Our earlier studies (BB,WB) used the term "individualization"; ET used "identification", in order to use the terms prevalent in the discipline at the time the research was conducted.
Inconclusive	The comparison determination that neither individualization nor exclusion is possible.
Latent (or latent print)	An image of a friction ridge impression from an unknown source. In North America, "print" is used to refer generically to known or unknown impressions [15]. Outside of North America, an impression from an unknown source (latent) is often described as a "mark" or "trace," and "print" is used to refer only to known impressions (exemplars).
LQMetric	Latent Quality Metric (LQMetric) software automatically assesses the quality of latent fingerprint images [6]. LQMetric is included in the FBI's Universal Latent Workstation (ULW) software [16], release 6.5 and later.
Mated	A pair of images (latent and exemplar) known a priori to derive from impressions of the same source (finger). Compare with "ID," which is an examiner's determination that the prints are from the same source.
Missed ID	Failure by an examiner to individualize a mated pair that was individualized by any other examiners.
Nonmated	A pair of images (latent and exemplar) known a priori to derive from impressions of different sources (different fingers and/or different subjects).
NV (No value)	The impression is not of value for individualization and contains no usable friction ridge information. See also VEO and VID.
Reliability	Consistency of results, here differentiated into repeatability (c.f.) and reproducibility (c.f.)
Repeatability	Intraexaminer agreement: when one examiner provides the same response (annotation or determination) to a stimulus (image or image pair) on multiple occasions.
Reproducibility	Interexaminer agreement: when multiple examiners provide the same response (annotation or determination) to a stimulus (image or image pair).
Source	An area of friction ridge skin from which an impression is left. Two impressions are said to be from the "same source" when they have in common a region of overlapping friction ridge skin.
Sufficient	An examiner's assessment that the quality and quantity of information in a print (or image pair) justifies a specific determination (especially used with respect to individualization).
Value determination	An examiner's determination of the suitability of an impression for comparison: value for individualization (VID), value for exclusion only (VEO), or no value (NV). Agency policy often reduces the three value categories into two, either by combining VID and VEO into a value for comparison category or by combining VEO with NV into a "not of value for individualization" (Not VID) category (survey in [3]).
VEO	Value determination 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. See also NV and VID.
Verification	The final phase of ACE-V: the independent application of the ACE process by a subsequent examiner to either support or refute the conclusions of the original examiner.
VID	Determination 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. See also VEO and NV.
WB	Latent Print Examiner White Box Study [4], and the associated dataset. Described in Section 2, *Methods & Materials.*

Appendix SI-2 Latent print examiner eye-tracking study

The Latent print examiner eye-tracking (ET) study was conducted to assess how latent print examiners conduct analysis and comparison during examinations: 121 practicing latent print examiners performed 1444 latent-exemplar comparisons, and 550 exemplar-exemplar comparisons (not including 34 comparisons that resulted in corrupt or invalid data). Each participant was assigned a sequence of fingerprint comparisons, interspersed with three types of directed tasks (find-the-target, ridge following, and ridge counting). In addition to the value and comparison determinations, the eye-tracking dataset includes visual fixations extracted from 1KHz sampling data. The find-the-target eye-tracking data was described in a previous publication [2], but otherwise the eye-tracking study and data have not previously been published. Each participant who completed the study was assigned 15 latent-exemplar comparisons, and 6 exemplar-exemplar comparisons.

Participation was open to practicing latent print examiners who are currently doing casework or have done casework within the last year. Participants gave informed consent after reviewing a human subject consent form approved by the Federal Bureau of Investigation Institutional Review Board prior to the start of the study. Participants were assured that their results would remain anonymous; a coding system was used to ensure anonymity during our analyses and in reporting. Cross-references between personal information and results were destroyed prior to publication: the identities of participants were not associated with the results during analysis, and such association will not be possible subsequently, such as for discovery. Each participant completed a background survey, the results of which can be found in [2].

Testing occurred in June-August 2016 in six locations in Ohio, Indiana, Virginia, Kentucky, and Georgia. Participants were provided with written instructions prior to the test. An experimenter then verbally summarized the instructions and answered any questions. Participants were requested to perform the assigned tasks for two hours or until all of the assigned tasks were completed; however, participants were permitted to stop early or continue after the two-hour time period.

The ET dataset included 25 mated and 20 nonmated latent-exemplar image pairs that had previously been used in the BB or WB studies. The image pairs were specifically selected to assess reproducibility of examiner conclusions, and were explicitly not intended to be representative. More mates than nonmates were selected to focus on understanding "missed IDs" (disagreements on mated image pairs). Seven of the 45 image pairs were selected because they resulted in unanimous responses in the WB study; as noted in Section 3, only one remained unanimous after being assigned to more examiners in the ET study (). The remaining 38 latent-exemplar image pairs were selected based on low reproducibility and/or previous errors in the Black Box and White Box studies [3, 4].

The ET dataset also included 8 mated and 10 nonmated exemplar-exemplar image pairs, selected to assess how very easy comparisons are conducted. All of the exemplars were very high quality, from the "ULW Ground Truth" dataset (1000ppi scans of the same images as used in NIST's Special Database 27). The mates were expected to be obvious IDs. The nonmates were expected to be obvious exclusions: six of the nonmates were unrelated pattern classes (e.g., whorl vs. loop), and four were superficially similar pattern classes (e.g. left loop against left loop).

Summary of Eye-Tracking Study test instructions

Ed. Note: this section is taken verbatim from the instructions provided to the eye-tracking study participants, but is limited to the content relevant to this paper (i.e. omitting directions regarding feature markup and eye tracking).

In this study, you will be asked to perform a series of friction ridge impression examinations. Eye-tracking cameras will record the position of your head and eyes, in order to measure where on each image you are looking. Most of the test involves performing comparisons of two fingerprints. In addition, there are a few "directed tasks". It is important that you apply the same diligence that you use in casework when performing comparisons.

Analysis & Latent Value

For each examination, the software will first present a fingerprint at the left of the screen for analysis: generally a latent, but occasionally an exemplar.

Once you complete the analysis stage, indicate the value of the print:

- *Of value for identification — The impression is of value and is appropriate for potential identification and/or exclusion if an appropriate exemplar is available.*
- *Of value for exclusion only — The impression is NOT of value for identification. The impression contains some friction ridge information that may be appropriate for exclusion if an appropriate exemplar is available.*
- *No Value — The impression does not contain sufficient friction ridge information to reach an identification or exclusion conclusion.*

Comparison/Evaluation Conclusion

At the end of comparison, you must make one of these conclusions: (say the conclusion aloud)

- *Identification — The two fingerprints originated from the same finger.*
- *Exclusion — The two fingerprints did not come from the same finger.*
- *Inconclusive — Neither identification nor exclusion is possible.*

- *Comparison not completed — You may choose this option if an examination is taking excessive time. You will be reminded of this option after 20 minutes, and asked to stop after 30 minutes.*

Borderline conclusion

Indicate whether your conclusion was a borderline decision, defined in this way:

- *If another examiner performed blind verification on this image pair and reached a different conclusion than you, how surprised would you be?*
    - *Not borderline — You would be very surprised if another examiner disagreed: you would expect almost every qualified examiner to reach the same conclusion (say "NOT BORDERLINE")*
    - *Borderline — You would not be very surprised if another examiner disagreed: you would expect other examiners might disagree (say "BORDERLINE"). For inconclusives, indicate "borderline ID" or "borderline exclusion."*

[Why we ask: When assessing differences in eye behavior and differences in determinations, we want assistance in recognizing these borderline cases. For example, if you make an ID that is right on the edge of inconclusive, we want to be able to flag that as different from an ID you would expect every examiner to make.]

Difficulty

For each comparison, say how difficult the comparison was. Routine comparisons should be indicated as "Moderate".

- *Very Easy/Obvious — The comparison determination was obvious.*
- *Easy — The comparison was easier than most latent comparisons.*
- *Moderate — The comparison was a typical latent comparison.*
- *Difficult — The comparison was more difficult than most latent comparisons.*
- *Very Difficult — The comparison was unusually difficult, involving high distortion and/or other red flags.*

Appendix SI-3 Summary of conclusion rates

This section provides a summary of the conclusion rates for the BB, BBR, WB, and ET datasets, to allow comparison of the three studies. More complete descriptions of the BB, BBR, and WB datasets, may be found in [3, 1, 4].

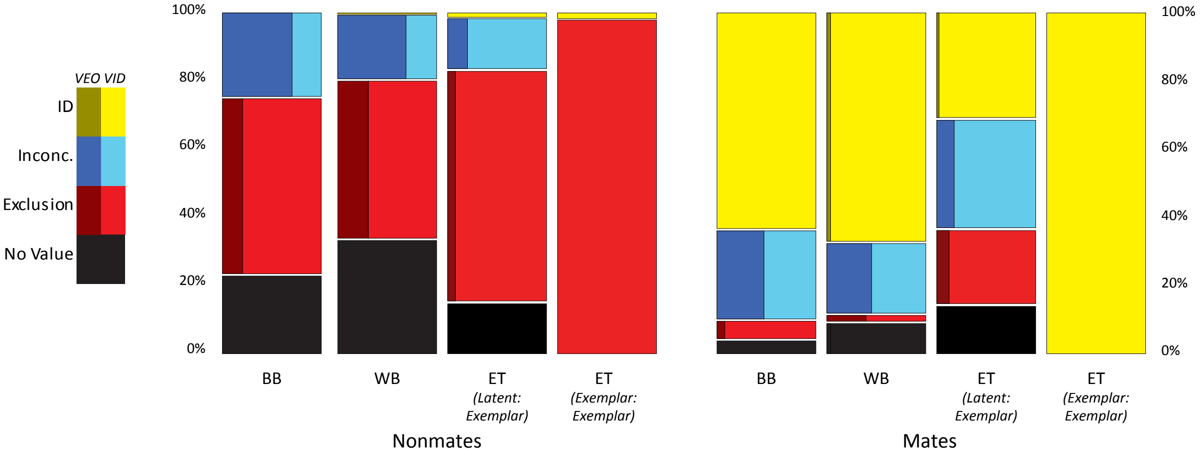

Fig. S1. Comparison of distributions of determinations in Black Box, White Box, and Eye-tracking studies. BB and WB graphs were previously published [3, 5], included here for ease of comparison. (BB: n=17,121 determinations, WB: n=3,730, ET(L:E): n=1,444, ET(E:E): n=550)

Table S1 provides summary counts and rates for the three studies. Regarding the erroneous ID (false positive) rate for the ET dataset, note that if the examiner who made six erroneous IDs in ET were not included, the exemplar-exemplar false positive rate would have been 0%, and the latent-exemplar false positive rate would have been 0.6%.

	Black Box dataset				*White Box dataset*				*Eye-tracking dataset*							
	Latent-Exemplar				*Latent-Exemplar*				*Exemplar-Exemplar*				*Latent-Exemplar*			
	Mates		*Nonmates*		*Mates*		*Nonmates*		*Mates*		*Nonmates*		*Mates*		*Nonmates*	
NV	3,389	*29.3%*	558	*10.1%*	485	*16.8%*	259	*30.5%*	0	*0.0%*	0	*0.0%*	119	*14.8%*	101	*15.8%*
Excl.	611	*5.3%*	3,947	*71.2%*	131	*4.5%*	430	*50.7%*	0	*0.0%*	303	*98.7%*	178	*22.1%*	434	*67.8%*
Inc.	3,875	*33.5%*	1,032	*18.6%*	567	*19.7%*	158	*18.6%*	0	*0.0%*	0	*0.0%*	257	*32.0%*	99	*15.5%*
ID	3,703	*32.0%*	6	*0.1%*	1,699	*59.0%*	1	*0.1%*	243	*100.0%*	4	*1.3%*	250	*31.1%*	6	*0.9%*
Total	11,578		5,543		2,882		848		243		307		804		640	

Table S1: Summary of responses by conclusion in Black Box, White Box, and Eye-tracking studies. (BB: conclusions by 169 examiners on 520 mated and 224 nonmated image pairs. ET: conclusions by 115 examiners on 8 mated and 10 nonmated exemplar-exemplar image pairs; conclusions by 121 examiners on 25 mated and 20 nonmated exemplar-exemplar image pairs)

	Exemplar-exemplar				*Latent-exemplar*			
	Mates		*Nonmates*		*Mates*		*Nonmates*	
NV	0	*0.0%*	0	*0.0%*	119	*14.8%*	101	*15.8%*
Ex	0	*0.0%*	303	*98.7%*	117	*14.6%*	352	*55.0%*
Ex-Bord	0	*0.0%*	0	*0.0%*	61	*7.6%*	82	*12.8%*
Inc-LeanExcl	0	*0.0%*	0	*0.0%*	49	*6.1%*	47	*7.3%*
Inc-NoTime	0	*0.0%*	0	*0.0%*	19	*2.4%*	6	*0.9%*
Inc	0	*0.0%*	0	*0.0%*	66	*8.2%*	38	*5.9%*
Inc-LeanID	0	*0.0%*	0	*0.0%*	123	*15.3%*	8	*1.3%*
ID-Bord	1	*0.4%*	0	*0.0%*	71	*8.8%*	1	*0.2%*
ID	242	*99.6%*	4	*1.3%*	179	*22.3%*	5	*0.8%*
Total	243		307		804		640	

Table S2: Summary of responses by conclusion subcategory in the eye-tracking study. During analyses, "Comparison not completed" (Inc-NoTime) is considered a category of Inconclusive. (ET dataset: conclusions by 115 examiners on 8 mated and 10 nonmated exemplar-exemplar image pairs; conclusions by 121 examiners on 25 mated and 20 nonmated exemplar-exemplar image pairs)

		Second assignment		
		Not VID	*VID*	*Total*
First assign.	*Not VID*	702	119	821
	VID	95	1,387	1,482
	Total	797	1,506	2,303

Table S3: Summary of (intra-examiner) repeatability of value determinations in the Black Box Repeatability study. (BBR dataset: n=2,303 assignments of latents on two occasions (4,606 total trials); 340 latents, 168 participants)

			Second assignment				
			ID	*Inconc*	*Excl*	*NV*	*Total*
First assignment	*Mates*	*ID*	237	20	9	0	266
		Inconc	35	208	12	26	281
		Excl	47	97	68	14	226
		NV	1	26	3	215	245
		Total	320	351	92	255	1,018
	Nonmates	*ID*	0	0	0	0	0
		Inconc	0	67	42	15	124
		Excl	0	39	426	5	470
		NV	0	10	5	36	51
		Total	0	116	473	56	645

Table S4: Summary of (intra-examiner) repeatability of comparison conclusions in the Black Box Repeatability study. (BBR dataset: 1,663 assignments of image pairs on two occasions (3,326 total trials); 422 mated and 210 nonmated image pairs, 72 participants. Note this is a subset of the value repeatability data shown in Table S3, which also includes reassignments of the same latents but with different exemplars.)

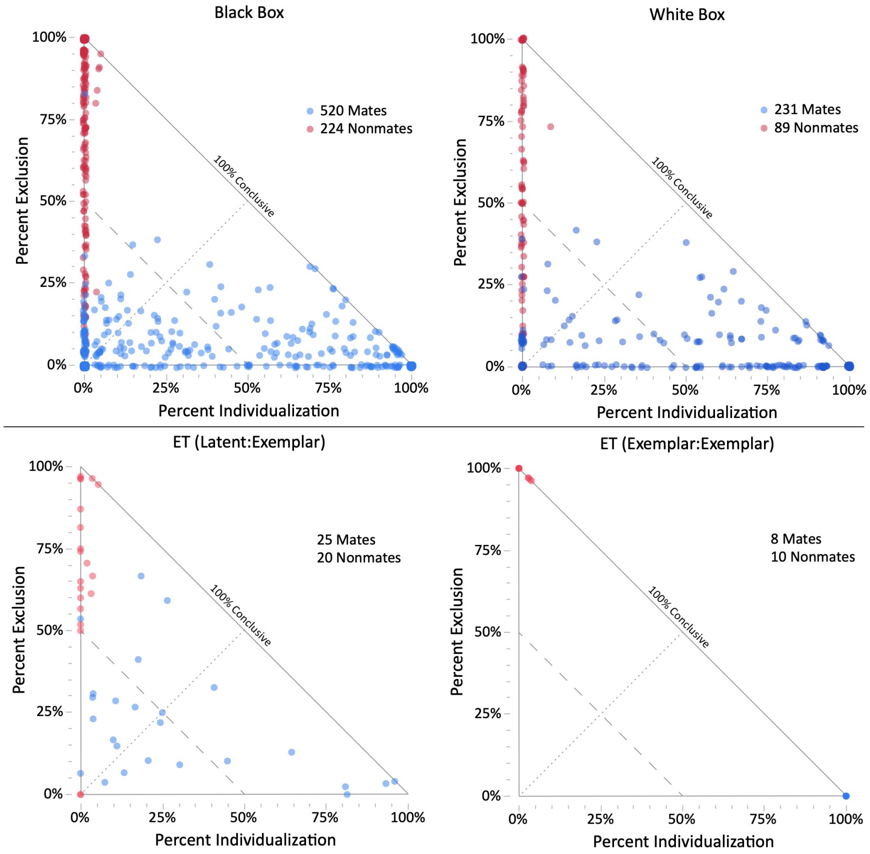

Fig. S2: Comparison of determination rates for each image pair in Black Box, White Box, and Eye-tracking studies. BB and WB graphs were previously published [3, 5], included here for ease of comparison. Mean examiners per image pair: 23 (BB); 12 (WB); 32 (ET L:E); 31 (ET E :E). Points at the origin (0,0) represent image pairs that examiners agreed unanimously could neither be excluded nor individualized; points at the bottom right were unanimous individualizations; points at the top left were unanimous exclusions. Image pairs above and right of the dashed line had more conclusions than inconclusive and NV. Image pairs above and left of the dotted line had more exclusions than individualizations.

Appendix SI-4 Summary of reproducibility rates

This section summarizes the reproducibility rates for the BB and ET datasets.

			Examiner 2							
			Black Box dataset				*Eye-tracking dataset*			
			NV	*VEO*	*VID*	*Total*	*NV*	*VEO*	*VID*	*Total*
Examiner 1	*Mates*	*NV*	59,146	13,526	4,299	76,971	1,018	502	2,193	3,713
		VEO	13,526	22,408	15,233	51,167	502	476	1,521	2,499
		VID	4,299	15,233	118,858	138,390	2,193	1,521	16,500	20,214
		Total	76,971	51,167	138,390	266,528	3,713	2,499	20,214	26,426
	Nonmates	*NV*	8,828	4,768	2,290	15,886	1,774	226	1,088	3,088
		VEO	4,768	9,588	9,105	23,461	226	238	1,291	1,755
		VID	2,290	9,105	90,954	102,349	1,088	1,291	13,540	15,919
		Total	15,886	23,461	102,349	141,696	3,088	1,755	15,919	20,762

Table S5: Inter-examiner reproducibility of value determinations. Counts of all pairwise combinations of decisions (on the same latents). (BB: 408,224 inter-examiner decision pairs derived from 17,121 decisions. ET: 47,188 inter-examiner decision pairs derived from 1,444 decisions.)

			Examiner 2									
			Black Box dataset					*Eye-tracking dataset*				
			NV	*Excl*	*Inc*	*ID*	*Total*	*NV*	*Excl*	*Inc*	*ID*	*Total*
Examiner 1	*Mates*	*NV*	59,146	2,028	14,710	1,087	76,971	1,018	610	1,297	788	3,713
		Excl	2,028	2,052	6,057	3,723	13,860	610	2,198	1,812	1,457	6,077
		Inc	14,710	6,057	55,108	13,450	89,325	1,297	1,812	3,128	1,765	8,002
		ID	1,087	3,723	13,450	68,112	86,372	788	1,457	1,765	4,624	8,634
		Total	76,971	13,860	89,325	86,372	266,528	3,713	6,077	8,002	8,634	26,426
	Nonmates	*NV*	8,828	2,415	4,643	0	15,886	1,774	964	349	1	3,088
		Excl	2,415	86,092	10,577	116	99,200	964	10,976	2,135	170	14,245
		Inc	4,643	10,577	11,220	27	26,467	349	2,135	708	32	3,224
		ID	0	116	27	0	143	1	170	32	2	205
		Total	15,886	99,200	26,467	143	141,696	3,088	14,245	3,224	205	20,762

Table S6: Inter-examiner reproducibility of comparison conclusions. Counts of all pairwise combinations of decisions (on the same image pairs). (BB dataset: 408,224 inter-examiner decision pairs derived from 17,121 decisions.[†] ET dataset: 47,188 inter-examiner decision pairs derived from 1,444 decisions.)

	Black Box		*Eye-tracking*	
	Mates	*Nonmates*	*Mates*	*Nonmates*
Agreement	69.2%	74.9%	41.8%	64.9%
Disagreement (sufficiency)	28.0%	24.9%	47.2%	33.5%
Disagreement (error)	2.8%	0.2%	11.0%	1.6%

Table S7: Proportions of inter-examiner disagreements regarding sufficiency vs errors. Summary of Table S6, indicating disagreements involving errors (ID vs exclusion), and disagreements regarding sufficiency ([ID or exclusion] vs [no value or inconclusive]). No value and inconclusive are treated as equivalent here.

[†] *Note the counts in this table for the Black Box study are based on all trials in that dataset and, therefore, differ from Table S4b in (Ulery B. T., Hicklin, Buscaglia, & Roberts, 2012), which was limited to the 72 examiners who participated in the repeatability retest.*

			Examiner 2									
			No Value	*Exclusion*		*Inconclusive*				*ID*		*Total*
			(NV)	*Ex*	*Ex-Bord*	*Inc-LeanExcl*	*NoTime*	*Inc*	*Inc-LeanID*	*ID-Bord*	*ID*	
Examiner 1	*Mates*	*NV*	1,018	331	279	268	121	311	597	308	480	3,713
		Ex	331	1,314	349	266	79	355	426	327	606	4,053
		Ex-Bord	279	349	186	169	62	165	290	223	301	2,024
		Inc-LeanExcl	268	266	169	110	36	126	248	137	213	1,573
		NoTime	121	79	62	36	24	53	88	59	99	621
		Inc	311	355	165	126	53	258	353	149	287	2,057
		Inc-LeanID	597	426	290	248	88	353	928	336	485	3,751
		ID-Bord	308	327	223	137	59	149	336	340	606	2,485
		ID	480	606	301	213	99	287	485	606	3,072	6,149
		Total	3,713	4,053	2,024	1,573	621	2,057	3,751	2,485	6,149	26,426
	Nonmates	*NV*	1,774	740	224	118	22	186	23	1	0	3,088
		Ex	740	7,512	1,493	796	84	544	110	24	116	11,419
		Ex-Bord	224	1,493	478	293	31	227	50	12	18	2,826
		Inc-LeanExcl	118	796	293	162	19	135	23	6	10	1,562
		NoTime	22	84	31	19	2	14	3	0	1	176
		Inc	186	544	227	135	14	104	24	6	5	1,245
		Inc-LeanID	23	110	50	23	3	24	4	1	3	241
		ID-Bord	1	24	12	6	0	6	1	0	0	50
		ID	0	116	18	10	1	5	3	0	2	155
		Total	3,088	11,419	2,826	1,562	176	1,245	241	50	155	20,762

Table S8: Inter-examiner reproducibility of subcategories of comparison conclusions. Counts of all pairwise combinations of decisions on the same image pair. (ET dataset: 47,188 inter-examiner decision pairs derived from 1,444 decisions.)

Appendix SI-5 Image effects

This section provides support for Section 3, Image effects.

Appendix SI-5.1 Image quality and unanimity of conclusions

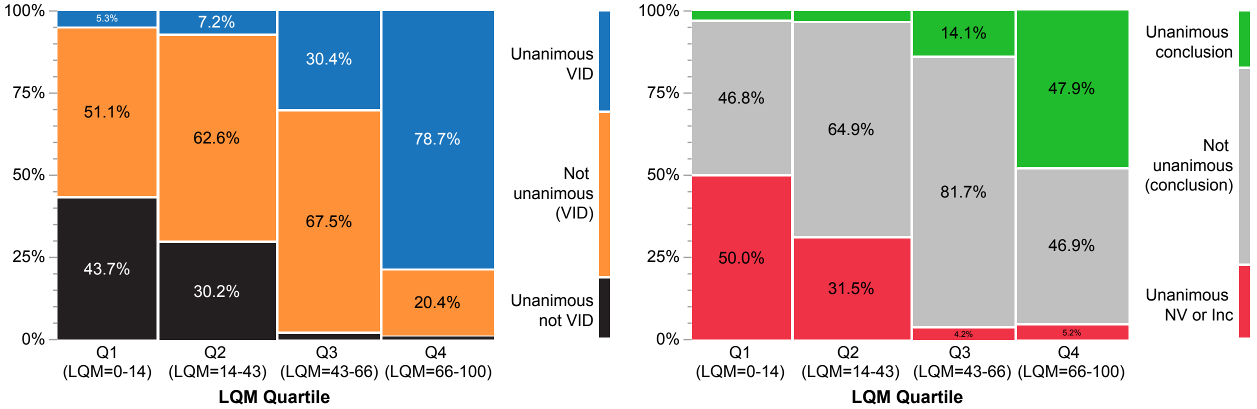

Fig. S3. Association of latent quality (x axis) with rates of unanimity (y axis) for value (left) and conclusions (right). The quality of each latent in each image pair is assessed in terms of LQMetric quartiles. Latents on quartile boundaries are included in both quartiles. (BB dataset: 17,121 trials; 744 image pairs)

LQM Quartile	*Image pairs*	*%VID*	*Conclusion Rate*
Q1 (0-14)	190	21%	19%
Q2 (14-43)	222	36%	30%
Q3 (43-66)	191	77%	61%
Q4 (66-100)	202	97%	81%

Table S9. Association of LQMetric quartiles with VID and conclusion rates. Latents on quartile boundaries are included in both quartiles (61 image pairs). (BB dataset: 17,121 trials; 744 image pairs)

VIDgroup	*Unanimous not VID*	*Unanimous VID*	*Not unanimous by VID*	*Unanimous NV or Inc*	*Unanimous ID or Excl*	*Not unanimous by conclusion*
Q1 (0-14)	53%	4%	24%	52%	4%	18%
Q2 (14-43)	43%	6%	34%	38%	6%	30%
Q3 (43-66)	3%	23%	32%	4%	19%	32%
Q4 (66-100)	1%	66%	11%	6%	71%	20%
Image pairs	156	245	404	183	139	483

Table S10. Association of unanimity categories and LQMetric quartiles: same data as Fig. S3 but percentages by column rather than row.

Appendix SI-5.2 Examples of low-reproducibility image pairs

The following figures show examples of mated latent-exemplar image pairs that resulted in notably low levels of reproducibility. Counts of assignments and conclusions include the totals for the image pair across the BB, WB, and/or ET studies.

The images below are reproduced at the same resolution. For journal reproduction, histogram equalization was used to adjust the grayscale values for journal reproduction, and images were cropped to reduce background area.

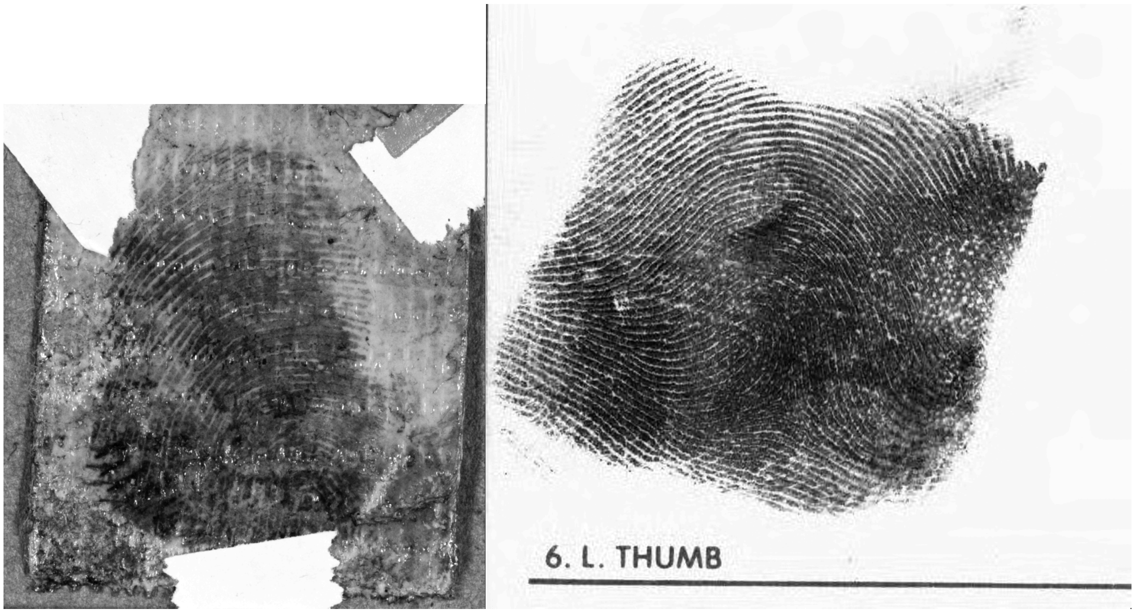

Fig. S4. Low reproducibility mated image pair. 49 assignments: 20 ID (41%), 16 exclusion (33%), 13 inconclusive or NV (27%).

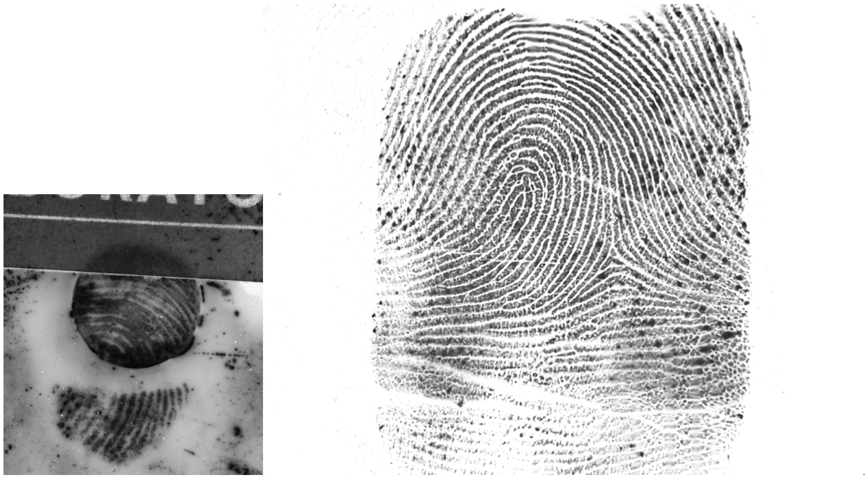

Fig. S5. Low reproducibility mated image pair. 49 assignments: 22 ID (45%), 5 exclusion (10%), 22 inconclusive or NV (45%).

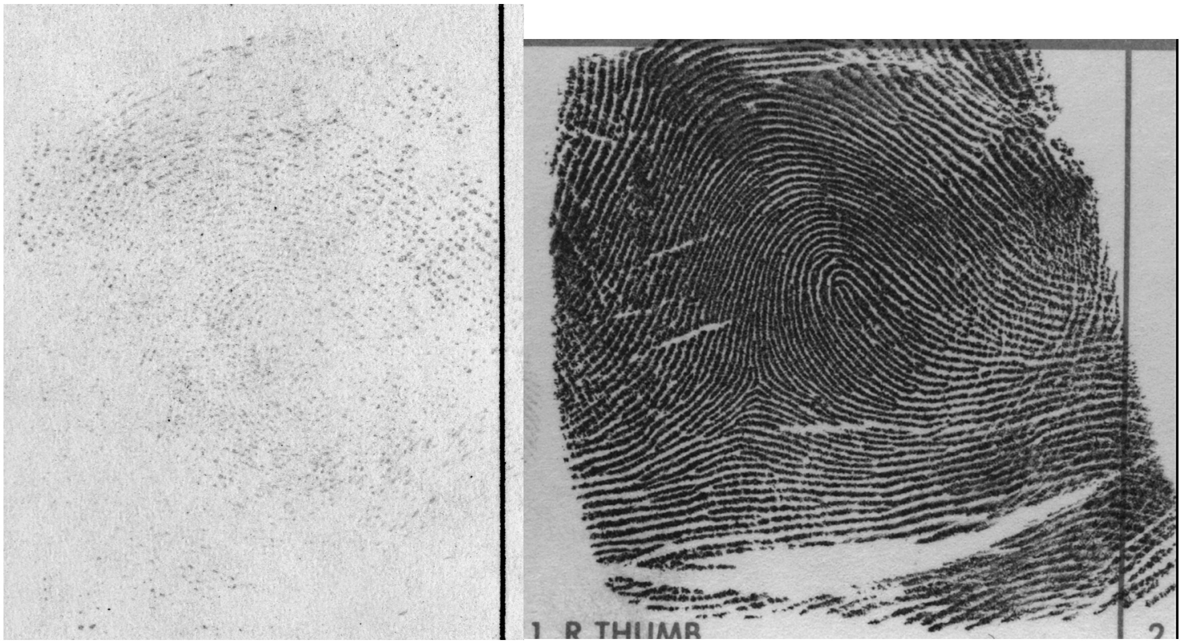

Fig. S6. Low reproducibility mated image pair. 41 assignments: 10 ID (24%), 9 exclusion (22%), 22 inconclusive or NV (54%).

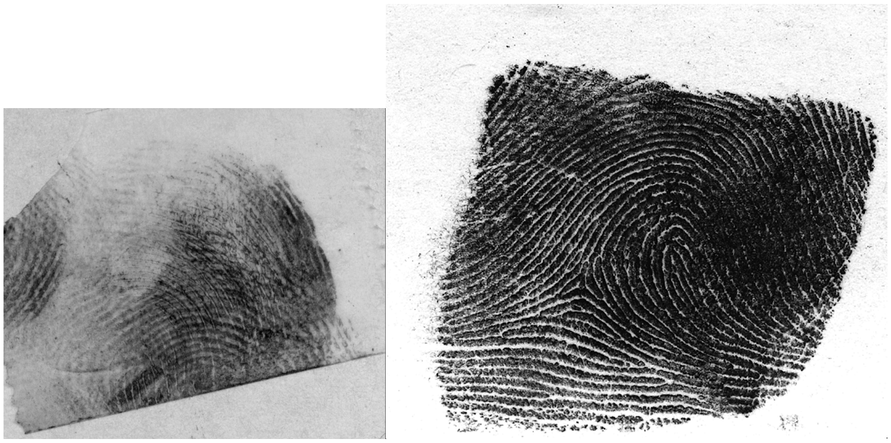

Fig. S7. Low reproducibility mated image pair. 33 assignments: 10 ID (30%), 3 exclusion (9%), 20 inconclusive or NV (61%). All superimposed impressions in the latent are from the same finger.

Appendix SI-6 Predicting determinations based on image and examiner rates

This appendix provides support for Section 4, Examiner effects.

Appendix SI-6.1 Predicting VID vs. VCMP

Examiners in the BB and ET studies assessed value on a three-category scale: value for ID (VID), value for exclusion only (VEO), or no value (NV). For the models discussed in this paper, we have reported this as a two-state decision, VID vs. not VID (combining VEO and NV in not VID); alternatively, the three-category scale could be reduced to VCMP vs. NV (combining VID and VEO in "Value for Comparison", VCMP). Fig. S8A shows the same data as Fig. 1, but color-coded to differentiate VEO. Fig. S8B shows how the model is affected by using VCMP rather than VID as a basis. We use VID (instead of VCMP) as it provides more of a basis for differentiating among examiners, and because it is the approach most frequently used (the background survey of BB participants showed that 55% do not differentiate between VEO and NV).

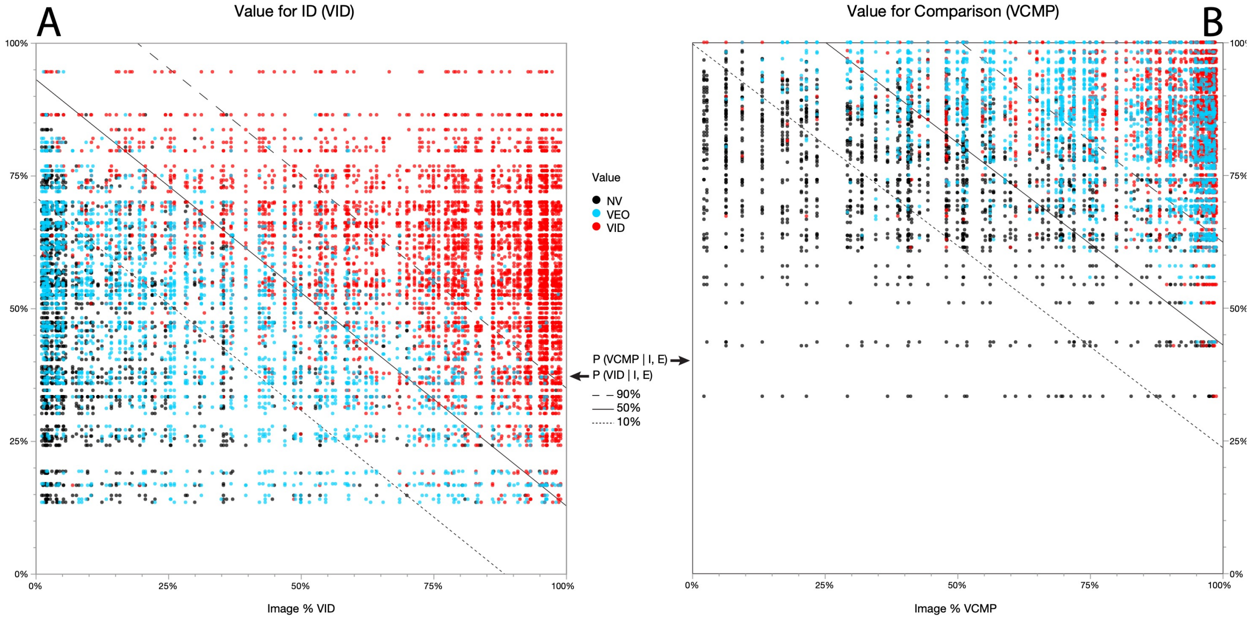

Fig. S8. Effect of images and examiners on value assessments of latent images, showing the effects of VEO determinations. A is identical to Fig. 1, but color-coded to differentiate three categories of value. (BB dataset. A: n=9,552 trials, limited to 203 latents on which VID determinations were not unanimous; B: n=6,505 trials, limited to 138 latents on which VCMP determinations were not unanimous)

Appendix SI-6.2 Relative effects of examiners vs. images

Table S11 shows the relative contributions of image effects and examiner effects for each of the models discussed in Section 4, Examiner effects. For each model, the relative effect of the image (or image pair) vs examiner can be shown by comparing nominal logistic regression models based on the examiner alone, the image (or image pair) alone, or the image and examiner. For example, for VID, this compares predicting VID for each trial using the examiner's VID rate (omitting the current trial) and/or the image's VID rate (omitting the current trial).

	AUC			Misclassification rate		
	Examiner	Latent / Image pair	Latent / Image pair & Examiner	Examiner	Latent / Image pair	Latent / Image pair & Examiner
VID	0.66	0.89	0.94	0.37	0.19	0.14
Exclusion (TN, Nonmates)	0.68	0.84	0.91	0.30	0.20	0.15
ID (TP, Mates)	0.66	0.85	0.90	0.36	0.20	0.18

Table S11. Relative contributions of image effects and examiner effects for the regression models P(VID|L,E), P(Exclusion|IP,E), P(ID|IP,E). For example, for VID, this compares P(VID|E), P(VID|L), and P(VID|L,E). AUC = area under the (receiver operating characteristic) curve. (BB dataset. VID: n=9,552 assessments of 203 latents by 169 examiners. TP: n=5199 assessments of 224 mated image pairs by 169 examiners. TN: n=3,845 assessments of 154 nonmated image pairs by 169 examiners. Limited to non-unanimous trials)

Appendix SI-6.3 Modeling based on non-unanimous decisions

The extent to which examiners agree or disagree, and estimates of examiner decision rates are sensitive to data selection. The Black Box datasets spanned a wide range of images, including many assignments (latents and image pairs) on which decisions were unanimous [3]. The models presented here were constructed using only those assignments that did not result in unanimous decisions. Limiting to this subset of assignments standardizes how the measurements are performed, facilitating comparability of results across studies where data selection may differ substantially. Additionally, although BB assignments were randomized and balanced as described in [3], examiners were not all given the same assignments: omitting assignments that resulted in unanimous decisions further reduces biases in our estimated examiner decision rates, making the estimates for different examiners more comparable.

For this purpose, images (or image pairs) were omitted if they were unanimous for the specific determination being modelled. For example, for P(ID), image pairs were considered unanimous if 100% or 0% of examiners reported ID, so an image pair would be treated as unanimous with respect to ID if it was 50% inconclusive, 40% NV, and 10% exclusion.

Appendix SI-7 Borderline decisions and comfort zones

This appendix provides support for Section 5, Borderline decisions and comfort zones.

Appendix SI-7.1 Repeatability with respect to predictive models

Fig. 3 in the main paper showed the extent to which the P(VID|I,E) model explained the repeatability of value determinations. Fig. S9 shows the corresponding charts for the P(ID|IP,E) and P(Exclusion|IP,E) models. For this purpose, conclusions were considered “repeated” with respect to the specific determination being modelled. For example, for P(ID), conclusions were considered not repeated if the examiner concluded ID on one occasion and any determination other than ID on the other occasion (so a conclusion would be treated as repeated with respect to ID if the examiner made one inconclusive and one NV determination).

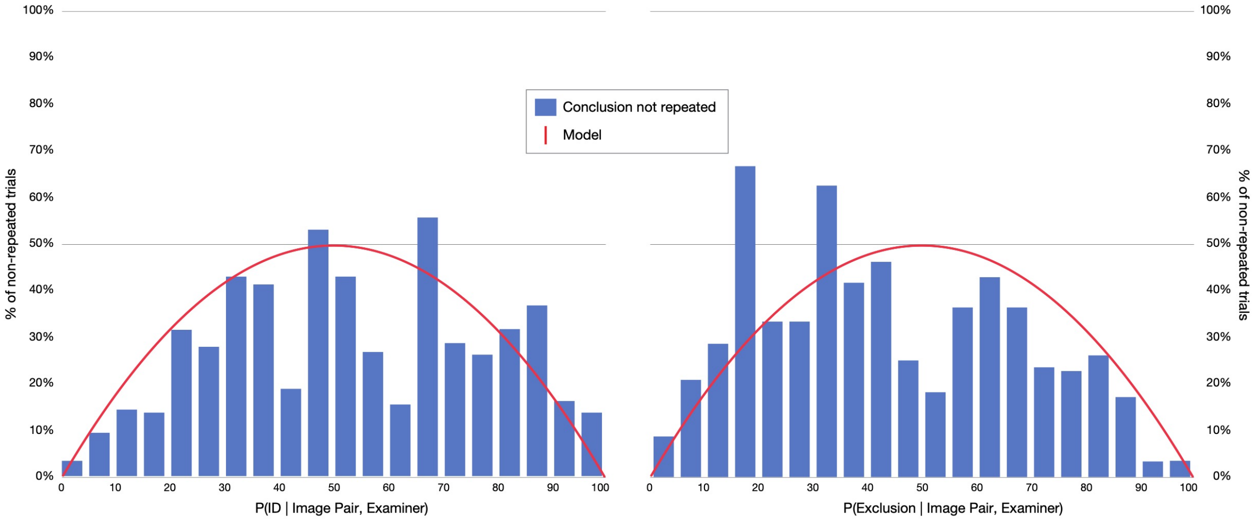

Fig. S9: Repeatability as a function of predicted conclusions. Bars indicate observed proportions of conclusions that were not repeated by the same examiner when retested after approximately 4-7 months, as a function of the P(ID) and P(Exclusion) models. Red: Expected proportions of non-repeated conclusions as a function of the P(ID) and P(Exclusion) models. See Fig. 3B for corresponding P(VID) chart. (BB repeatability dataset, limited to non-unanimous image pairs. Left: n=483 presentations of 194 distinct mated image pairs on two different occasions to 72 examiners (966 total trials); Right: n=464 presentations of 147 distinct mated image pairs on two different occasions to 72 examiners (928 total trials))

Fig. S10 shows more detail on the repeatability of conclusions with respect to the P(ID|IP,E) and P(Exclusion|IP,E) models. The color-coding indicates the determinations that were/were not repeated: for example, "Excl-IncNV" indicates those image pairs on which an examiner made an exclusion conclusion on one trial but an inconclusive or no value on a different trial, about four to seven months apart. Shading indicates those image pairs that always resulted in ID or exclusion conclusions, or never resulted in ID or exclusion conclusions (omitted from Fig. S8).

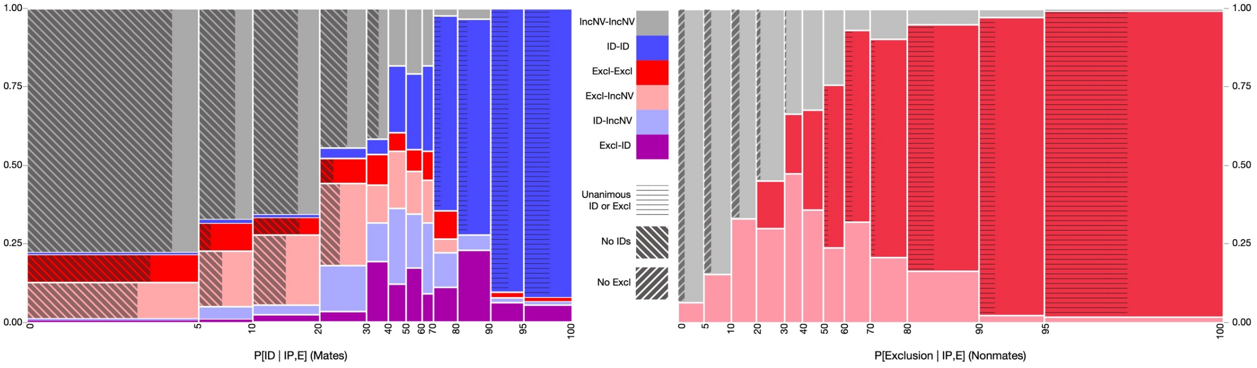

Fig. S10. Repeatability of conclusions by P(ID|IP,E) and P(Exclusion|IP,E). Conclusions are by the same examiner on the same image pairs, after approximately 4-7 months. Image pairs that resulted in unanimous conclusions are indicated by shading. There were no erroneous IDs (false positives) in the BB repeatability dataset. (BBR dataset. Left: n=1018 presentations of 422 distinct mated image pairs on two different occasions to 72 examiners (2036 total trials); Right: n=645 presentations of 210 distinct mated image pairs on two different occasions to 72 examiners (1290 total trials))

Appendix SI-7.2 Analysis and Comparison time vs. probability of conclusions

The eye-tracking study (ET) allowed precise measurement of Analysis and Comparison phase durations, which was not possible in the Black Box study. Rapid completion of Analysis is associated with highly probable determinations (i.e., rapid non-VID assessments with low P(VID|I,E), and rapid VID assessments with high P(VID|I,E). Comparison durations show similar but weaker associations with P(ID | IP,E) and P(Exclusion | IP,E).

Fig. S11 shows that rapid non-VID assessments tend to be associated with low P(VID|I,E), and rapid VID assessments tend to be associated with high P(VID|I,E). (Notice the asymmetry at the extremes, in that some VID assessments are lengthy even when P(VID|I,E) is very high: a plausible explanation is that Analysis is complete after an non-VID decision, but after a VID determination examiners may spend additional time preparing for Comparison.)

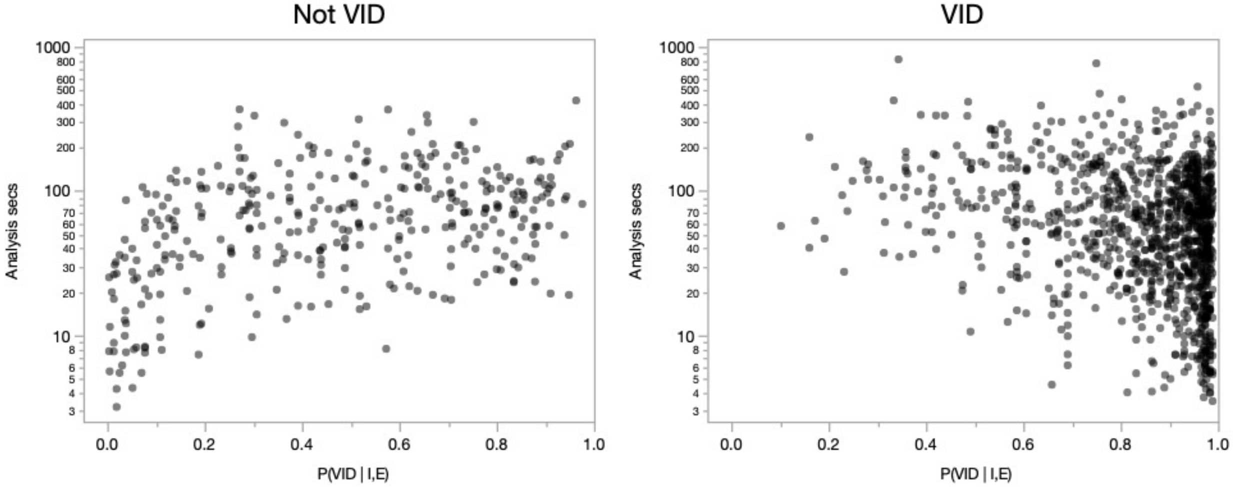

Fig. S11. Analysis phase duration by P(VID|I,E), for VID and non-VID assessments of latent images. (ET dataset. n=352 non-VID trials; 1,092 VID trials; 121 examiners on 45 distinct latents.)

Fig. S12 and Fig. S13 show the corresponding associations for Comparison durations. Note that the ET dataset had far fewer assignments per examiner than the BB dataset, affecting how well P(ID|IP,E) and P(Exclusion|IP,E) are estimated.

Fig. S12 shows the associations between Comparison durations and P(ID|IP,E) for mated pairs. Rapid inconclusives and erroneous exclusions are weakly associated with low P(ID|IP,E); rapid IDs are weakly associated with high P(ID|IP,E).

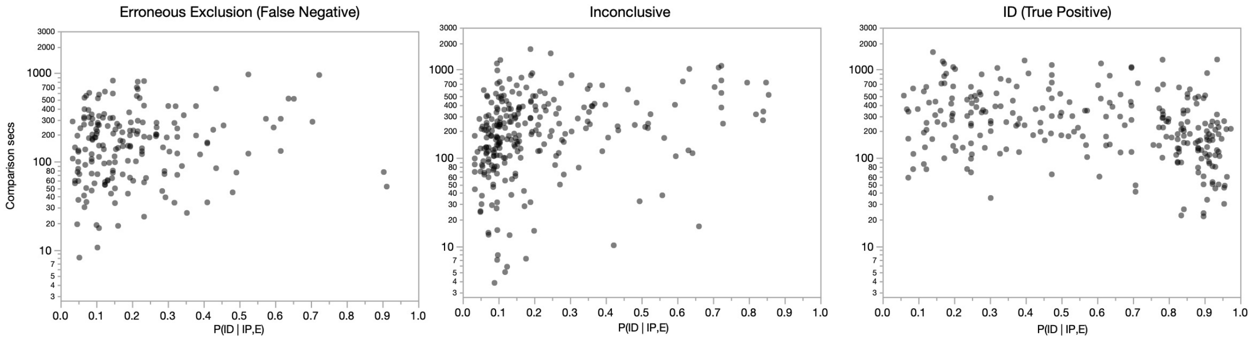

Fig. S12. Comparison phase duration by P(ID|IP,E), by conclusion on mated latent-exemplar pairs. (ET dataset. n=178 exclusions, 257 inconclusives, 250 IDs; 121 examiners on 25 mated image pairs.)

Fig. S13 shows the associations between Comparison durations and P(Exclusion|IP,E) for nonmated pairs. Rapid exclusions are associated with high P(Exclusion|IP,E).

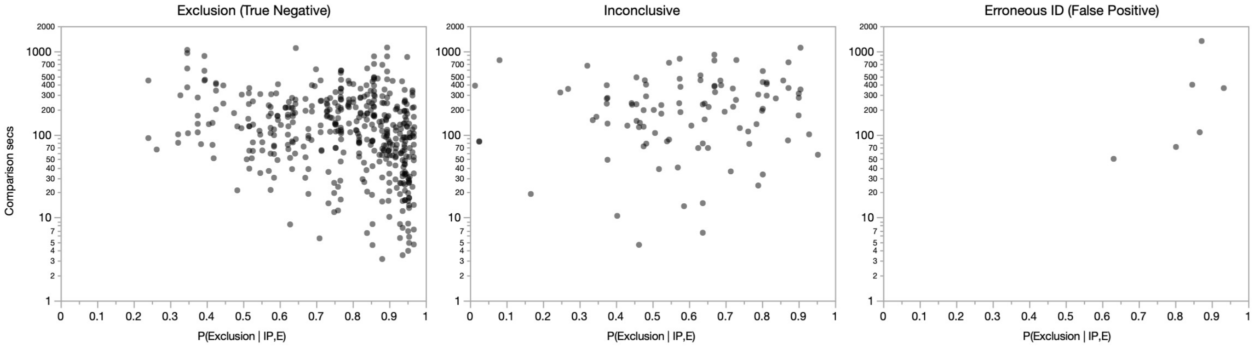

Fig. S13. Comparison phase duration by P(Exclusion|IP,E), by conclusion on nonmated latent-exemplar pairs. (ET dataset. n=434 exclusions, 99 inconclusives, 6 IDs; 121 examiners on 20 nonmated image pairs.)

Appendix SI-7.3 Difficulty and probability of conclusions

Fig. S14 shows the associations between probabilities of conclusions and the examiners' assessments of difficulty, by conclusion. Examiners' assessments of the difficulty of conclusions were strongly associated with the probability of conclusions:

- For mated image pairs, as the examiner-specific probability of ID (P(ID|I,E)) increased, examiners described their ID decisions as notably easier, and their inconclusive decisions as somewhat more difficult. Erroneous exclusions do not show a clear corresponding effect.
- For nonmated image pairs, as the examiner-specific probability of exclusion (P(Exclusion|IP,E)) increased, examiners described their exclusion decisions as notably easier, and their inconclusive decisions as somewhat more difficult.

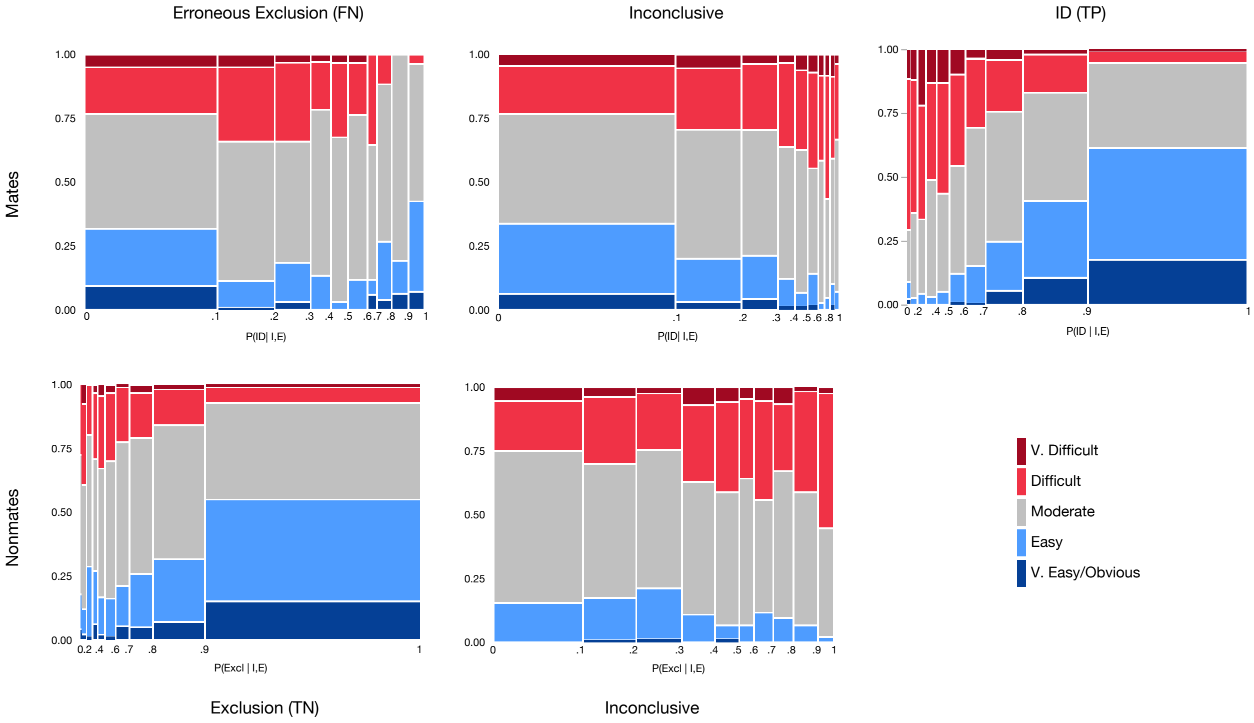

Fig. S14. Comparison difficulty by probabilities of conclusions. (BB dataset. Mates: n= 611 exclusions (false negatives), 3875 inconclusives, 3703 IDs (true positives). Nonmates: n=3947 exclusions (true negatives), 1032 inconclusives. (6 IDs on nonmates not shown))

Table S12 shows associations between difficulty and erroneous conclusions in the ET dataset. The proportion of exclusions that are erroneous is associated with increased difficulty, and therefore lower reproducibility would be associated with increased difficulty. This effect is more pronounced for the borderline exclusion subcategory.

Conclusion	*Conclusion subcategory*	*Difficulty*	*#trials*	*N(Mates)*	*N(Nonmates)*	*Percent mates*	*Reweighted to 50:50*	
							Mates	*Nonmates*
Exclusion	Ex	VeryEasy	15	2	13	13%	11%	89%
		Easy	125	28	97	22%	19%	81%
		Moderate	264	64	200	24%	20%	80%
		Difficult	62	22	40	35%	31%	69%
		VeryDifficult	3	1	2	33%	29%	71%
	Ex-Bord	Easy	1	0	1	0%	0%	100%
		Moderate	45	12	33	27%	23%	77%
		Difficult	80	38	42	48%	42%	58%
		VeryDifficult	17	11	6	65%	59%	41%
Inconclusive	Inc-IncEX	Moderate	25	11	14	44%	39%	61%
		Difficult	53	28	25	53%	47%	53%
		VeryDifficult	18	10	8	56%	50%	50%
	Inc	VeryEasy	1	1	0	100%	100%	0%
		Easy	5	4	1	80%	76%	24%
		Moderate	28	19	9	68%	63%	37%
		Difficult	48	30	18	63%	57%	43%
		VeryDifficult	20	10	10	50%	44%	56%
	Inc-IncID	Easy	5	5	0	100%	100%	0%
		Moderate	30	30	0	100%	100%	0%
		Difficult	76	70	6	92%	90%	10%
		VeryDifficult	20	18	2	90%	88%	12%
ID	ID-Bord	Moderate	21	20	1	95%	94%	6%
		Difficult	40	40	0	100%	100%	0%
		VeryDifficult	11	11	0	100%	100%	0%
	ID	VeryEasy	8	8	0	100%	100%	0%
		Easy	53	52	1	98%	98%	2%
		Moderate	91	87	4	96%	95%	5%
		Difficult	30	30	0	100%	100%	0%
		VeryDifficult	2	2	0	100%	100%	0%

Table S12. Mating by conclusion subcategories and difficulty. Reweighting accounts for the proportion of mated image pairs in the dataset (55.6% of ET image pairs were mated). Errors and conclusions contrary to ground truth are highlighted. (ET dataset: 1197 trials, omitting NV and NoTime trials, and 2 Inc trials for which the difficulty was missing)

Appendix SI-8 How much do examiners disagree?

This appendix provides support for Section 6, Disagreement as a function of categorical answers.

In the ET study, examiners selected subcategories within the typical three-level conclusion scale (ID, inconclusive, exclusion), rendering a seven-level conclusion scale. We can use the 7-level scale to better characterize the extent of disagreement among examiners..

If we assign values to the 3-level scale (Exclusion=0; Inconclusive or No value = 0.5; ID=1), then comparing examiners' conclusions results in a difference of 0 (same conclusion) to 1 (opposite conclusions, i.e. ID vs exclusion). For the 7-level scale, we assign intermediate categories within the same range: (Exclusion=0; Exclusion-Borderline=$^1/_6$; Inconclusive leaning toward Exclusion=$^1/_3$; Inconclusive or No value = $^1/_2$; Inconclusive leaning toward ID=$^2/_3$; ID-Borderline=$^5/_6$; ID=1).

In the ET dataset, we cross-compared all pairs of responses on each image pair (n=23,594 distinct pairs of responses from 1,444 individual responses). Fig. S15 shows the differences between 3-level and 7-level disagreements. Fig. S16 shows the distribution of disagreements in the ET dataset compared with BB data.

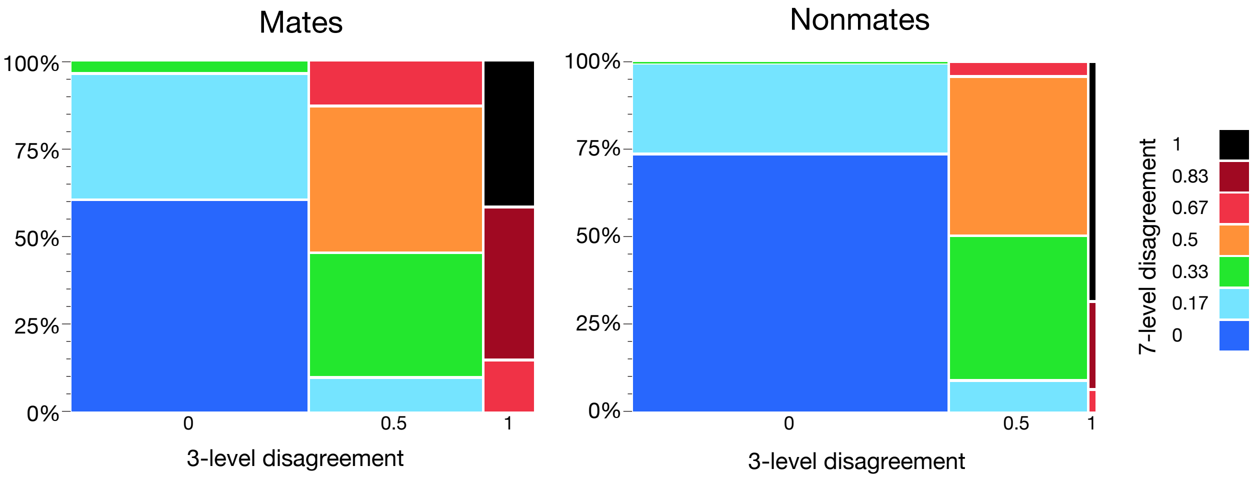

Fig. S15. Extent of disagreement as measured using two different conclusion scales. (ET dataset; n=23,594 distinct pairs of conclusions)

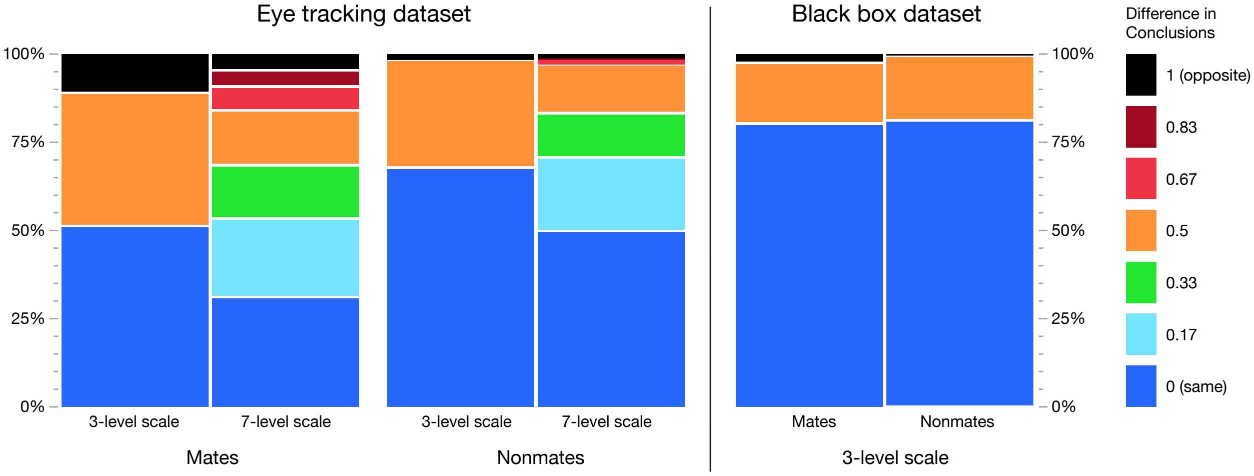

Fig. S16. Effect of conclusion scale on extent of disagreement. (ET dataset: n=23,594 distinct pairs of conclusions; BB dataset: n=204,112 distinct pairs of conclusions)

Appendix SI-9 Erroneous IDs

This appendix provides support for Section 7, Disagreements and erroneous conclusions.

Table S13 summarizes the responses for the image pairs that resulted in any erroneous IDs in the Black Box, White Box, and Eye-tracking studies. The following sections show the fingerprint images.

Type	*Image Pair*	*Latent*	*Exemplar*	*Assigns*	*ID (FP)*	*Inconclusive Leaning ID*	*Other Inconclusive*	*NV*	*Exclusion (TN)*	*Studies*
Latent-Exemplar	M044052	L_044	E_052	57	2	2	12	0	41	BB; ET
	M134387	L_134	E_387	60	3	0	1	0	56	BB; ET
	M134386	L_134	E_386	22	1	0	0	0	21	BB
	M143447	L_143	E_447	50	2	1	0	0	47	BB; ET
	M153389	L_153	E_389	60	1	3	11	0	45	BB; WB; ET
	M191450	L_191	E_450	52	1	2	24	4	21	BB; ET
	M245381	L_245	E_381	69	2	1	15	1	50	BB; WB; ET
	M289238	L_289	E_238	57	1	2	8	0	46	BB; ET
Exemplar-Exemplar	CE001			35	1	0	0	0	34	ET
	CE008			26	1	0	0	0	25	ET
	CE009			29	1	0	0	0	28	ET
	CE010			34	1	0	0	0	33	ET

Table S13. Summary of conclusions for the nonmated image pairs that resulted in any erroneous IDs in the three studies. Image pair IDs correspond to the data in [17]. "Inconclusive Leaning ID" combines the BB "inconclusive with corresponding features" and ET "inconclusive, borderline ID" categories.

Table S14 summarizes the associations between background survey data and erroneous IDs, combining the data from all three studies. Note that examiners in almost every category made erroneous IDs.

Agency	*Participants*	*Participants with erroneous IDs*	*Total erroneous IDs*	*Total Nonmates*	*FPR*
US Federal	195	3	3	3,428	0.1%
Non-US	40	2	3	252	1.2%
Private	21	0	0	367	0.0%
US State/Local	193	5	6	2,665	0.2%
No response	10	1	1	304	0.3%
Total number of years employed as a latent examiner[‡]					
1-4 years	111	2	3	1,825	0.2%
5-9 years	59	0	0	293	0.0%
5-14 years	138	5	6	2,489	0.2%
10 or more years	70	1	1	340	0.3%
15 or more years	69	2	2	1,675	0.1%
No response	12	1	1	394	0.3%
Training					
Less than 1 year	123	5	6	1,337	0.4%
Over 1 year	326	5	6	5,375	0.1%
No response	10	1	1	304	0.3%
Percent of time spent conducting latent print comparisons					
less than 50%	152	6	8	2,947	0.3%
more than 50%	297	4	4	3,765	0.1%
No response	10	1	1	304	0.3%
Agency Accreditation					
Accreditation unknown	19	1	1	422	0.2%
Accredited	340	5	5	5,491	0.1%
Not accredited	100	5	7	1,103	0.6%
Certification					
IAI CLPE	166	5	5	2,260	0.2%
Not IAI CLPE	283	5	7	4,452	0.2%
No response	10	1	1	304	0.3%

Table S14. Summary of associations between background survey data and erroneous IDs, across all three studies (BB, WB, ET). Note that to preserve anonymity, participation in multiple studies cannot be cross-referenced, so individuals who participated in more than one study would be counted more than once. (Participants in each study: 169 BB, 169 WB, 121 ET)

[‡] *Categories could not be combined across all studies: WB used {5-9 years, 10+ years}, whereas BB and ET used {5-14 years,15+ years}.*

Appendix SI-9.1 Latent fingerprints resulting in erroneous IDs

Fig. S17 through Fig. S23 show all of the latent prints resulting in erroneous IDs from the BB, WB, and ET studies. Image pair IDs correspond to the data in [17]. Five of the seven latents have the same substrate (galvanized metal) and processing (cyanoacrylate and light gray powder), which tonally reversed the image so that portions of ridges were light rather than dark.

These seven latents were collected from four individuals. The three latent prints shown in Fig. S19 through Fig. S21 were collected from one individual. Since other fingerprints from that subject are not unusual, the errors are presumably due to the complex combination substrate and processing. The two latent prints shown in Fig. S18 and Fig. S22 were collected from one individual.

The exemplars are also included for two image pairs. The exemplars for the other six image pairs are not releasable (fingerprints are protected as Personally Identifiable Information and public release requires permission from the subject, which could not be obtained for those exemplars); the exemplars that cannot be shown are all of typical to good quality for inked rolled exemplars.

The images below are reproduced at the same resolution. For journal reproduction, histogram equalization was used to adjust the grayscale values for journal reproduction, and images were cropped to reduce background area.

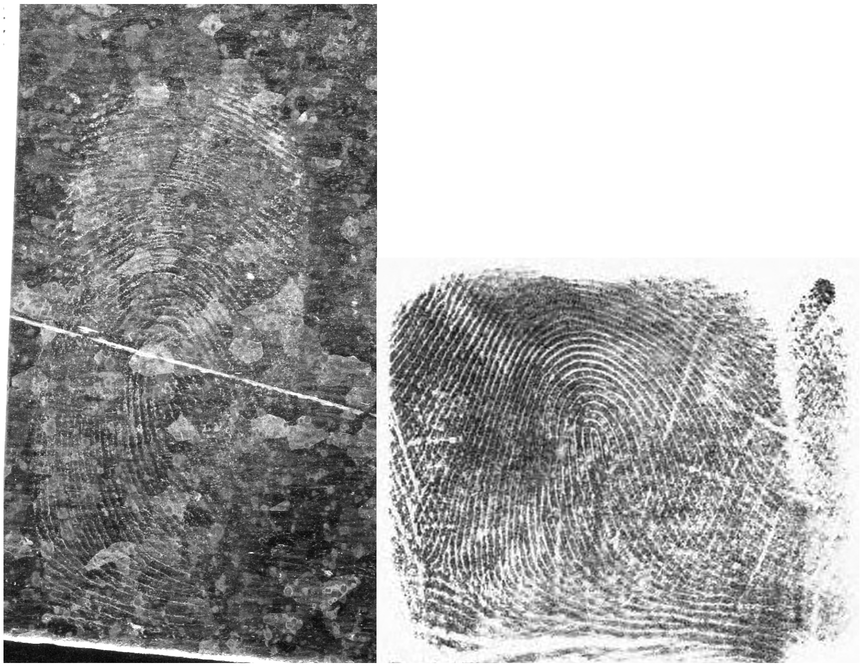

Fig. S17. Image pair M044052 (Latent L_044, Exemplar E_052). Erroneous IDs: 2 (57 assignments). (Previously shown in [3], Fig. 3)

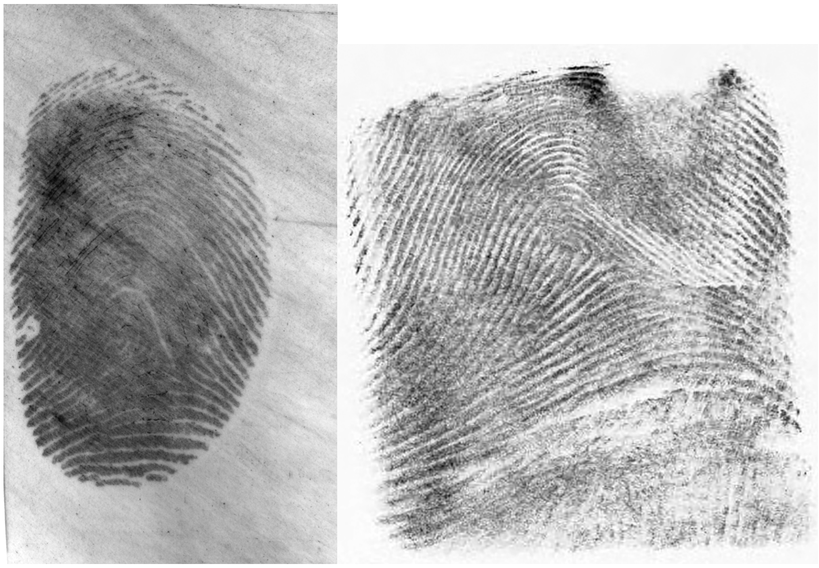

Fig. S18. Image pair M289238 (Latent L_289, Exemplar E_238). Erroneous IDs: 1 (57 assignments). (Previously shown in [3], Fig. 3)

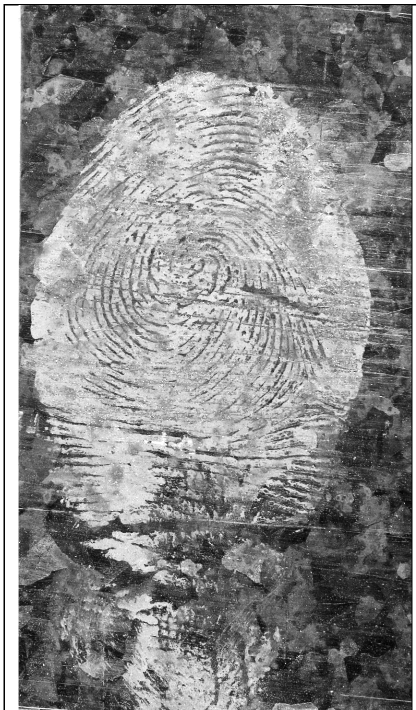

Fig. S19. Latent L_134. Used in two image pairs in BB resulting in erroneous IDs: Image pair M134387 (exemplar not releasable), erroneous IDs: 3 (60 assignments); Image pair M134386 (exemplar not releasable), erroneous IDs: 1 (22 assignments).

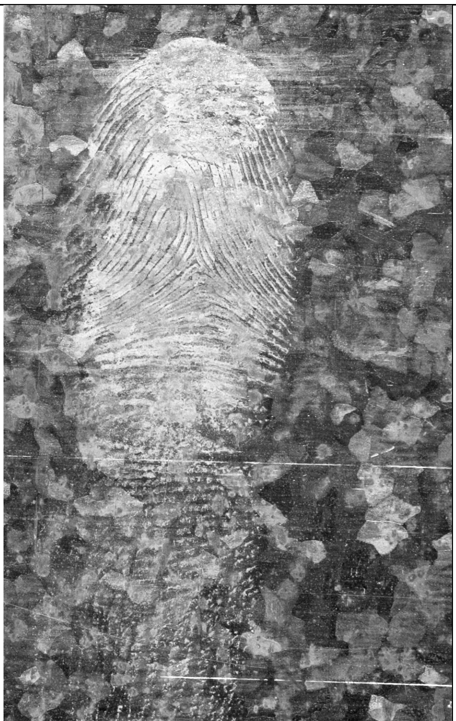

Fig. S20. Latent L_143 (from Image pair M143447, exemplar not releasable). Erroneous IDs: 2 (50 assignments)

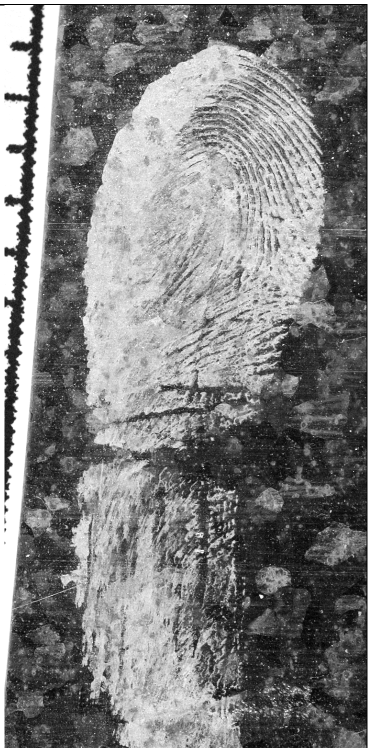

Fig. S21. Latent L_153 (Image pair M153389, exemplar not releasable). Erroneous IDs: 1 (60 assignments)

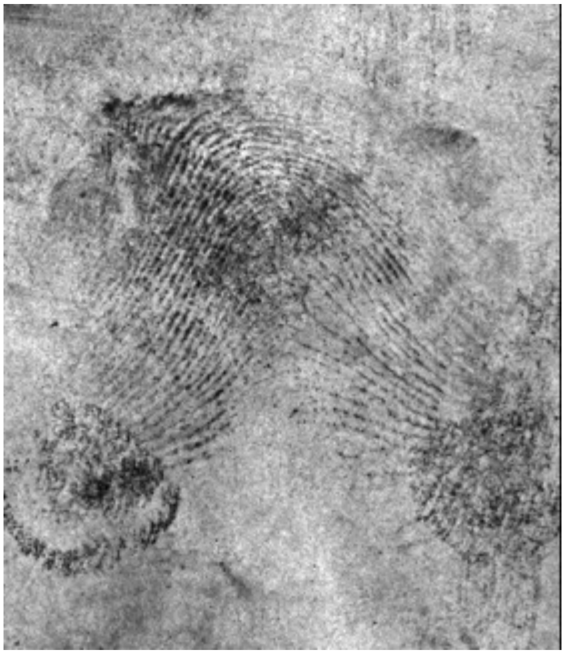

Fig. S22. Latent L_245 (Image pair M245381, exemplar not releasable). Erroneous IDs: 2 (69 assignments) (Previously shown in [18], Fig. 2)

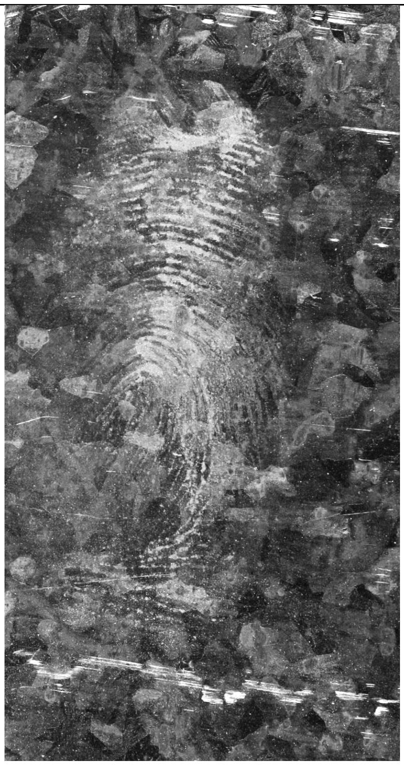

Fig. S23. Latent L_191 (Image pair M191450, exemplar not releasable). Erroneous IDs: 1 (52 assignments)

Appendix SI-9.2 Exemplars resulting in erroneous IDs

One participant in the eye-tracking study made four erroneous IDs on the four nonmated exemplar image pairs shown in Fig. S24 through Fig. S27 (as well as two erroneous IDs on latent-exemplar image pairs). Each of these was selected for the test as an obvious exclusion due to unrelated pattern classes, in order to evaluate eye tracking on easy comparisons. The examiner who made these erroneous IDs assessed the difficulty of each of these as “Moderate”; all other examiners assessed these as “Easy” or “Very Easy” and made conclusions of exclusion**.**

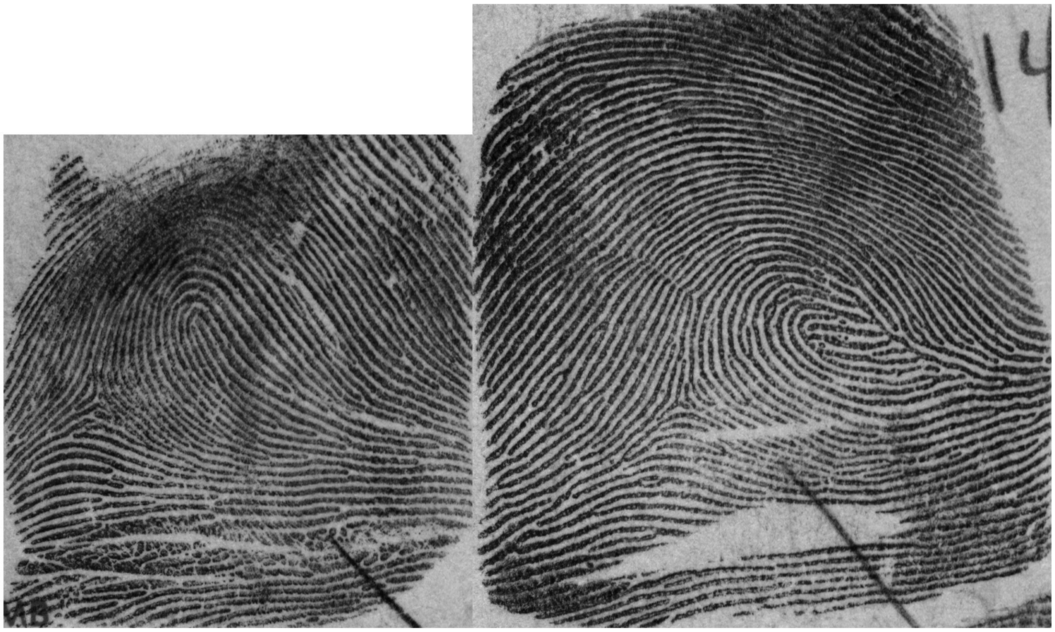

Fig. S24. Image pair CE001. Erroneous IDs: 1 (35 assignments)

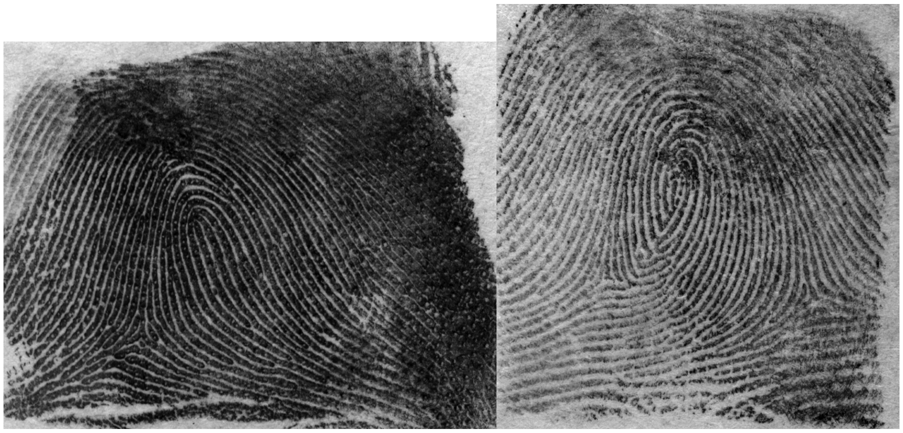

Fig. S25. Image pair CE008. Erroneous IDs: 1 (26 assignments)

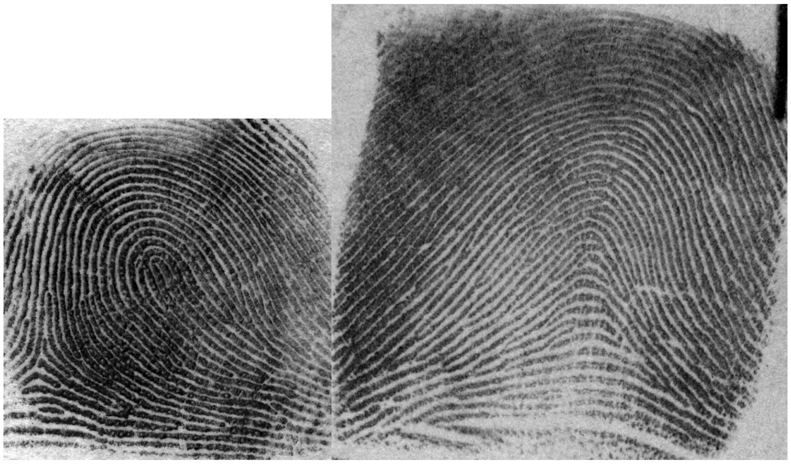

Fig. S26: Image pair CE009. Erroneous IDs: 1 (29 assignments)

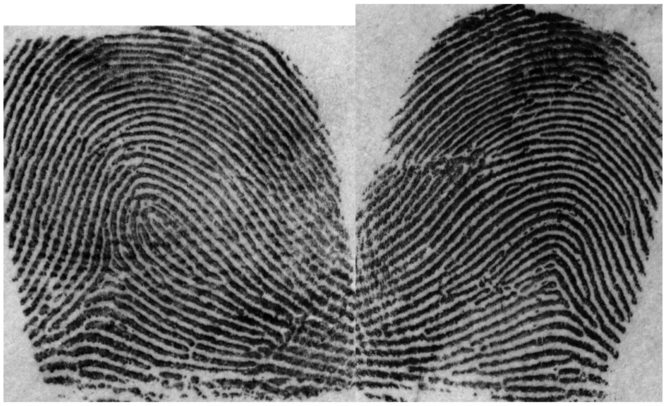

Fig. S27. Image pair CE010. Erroneous IDs: 1 (34 assignments)