Final accepted manuscript, subsequently published as
Ulery, B. T., Hicklin, R. A., Roberts, M. A., & Buscaglia, J. (2017). Factors associated with latent fingerprint exclusion determinations. *Forensic Science International, 275*, 65–75.
https://doi.org/10.1016/j.forsciint.2017.02.011

Factors associated with latent fingerprint exclusion determinations

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

[a] Noblis, Falls Church, Virginia, USA; [b] Latent Print Support Unit, Federal Bureau of Investigation Laboratory Division, Quantico, Virginia, USA; [c] Counterterrorism and Forensic Science Research Unit, Federal Bureau of Investigation Laboratory Division, Quantico, Virginia, USA

Corresponding author:

JoAnn Buscaglia, PhD
Counterterrorism and Forensic Science Research Unit
Federal Bureau of Investigation Laboratory
2501 Investigation Parkway
Quantico, Virginia 22135 USA
(703) 632-4553
(703) 632-4557 (fax)
joann.buscaglia@ic.fbi.gov

Abstract

Exclusion is the determination by a latent print examiner that two friction ridge impressions did not originate from the same source. The concept and terminology of exclusion vary among agencies. This study assesses the associations between a variety of factors and exclusion determinations. Although erroneous exclusions are more likely to occur on some images and for some examiners, they were widely distributed among images and examiners. Measurable factors found to be associated with the accuracy of exclusions include the quality of the latent, value determinations, analysis minutia count, comparison difficulty, and the presence of cores or deltas. The current lack of standard methods to precisely document the basis for exclusions limits our ability to discover exactly why examiners disagree and why they make errors. This research is intended to assist examiners in improving the examination process and provide information to the broader community regarding the accuracy, reliability, and implications of exclusion decisions.

1 Introduction

Historically, the latent print* [1-9] examination process was primarily focused on identifying (or individualizing) the person (subject) who left a latent print. Only in special circumstances did examiners need to make the distinction between not identifying the source of a latent print ("non-identification") and determining that a specific finger or palm from a subject was not the source of a latent print (exclusion). "Non-identification" is inherently ambiguous, as it does not differentiate between exclusions and inconclusive determinations: exclusions explicitly indicate that a subject was not the source of a latent, whereas inconclusives indicate that the examiner could not determine whether or not a subject was the source of a latent. This ambiguity came under criticism in the late 1990s and early 2000s as part of the accreditation of latent print units and crime laboratories. In response the Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST) guidelines were changed between 1997 and 2002, dropping non-identification as a determination, and adding inconclusive and exclusion determinations. Although SWGFAST guidelines changed, some laboratories and individual examiners continue to use the older non-identification determination [10]. The changing role of exclusion determinations in standard practice presents a new challenge for the latent print community, which is still adjusting to these changes.

SWGFAST defines the term "exclusion" to mean "the determination by an examiner that there is sufficient quality and quantity of detail in disagreement to conclude that two areas of friction ridge impressions did not originate from the same source" [11]. An examiner can exclude a specific anatomical area (such as a specific finger from a specific person), or a person ("if all relevant comparable anatomical areas are represented and legible in the known exemplars") [12].†

The term "exclusion" is not used consistently throughout the latent print community. In 2009, the latent print examiners who participated in our Black Box study [2] were asked to specify how they use the term "exclusion" as a conclusion in their standard operating procedures: examiners differed on whether exclusion means that the latent did not come from any friction ridge skin for that subject (51%), from any finger from the subject (10%), or from a specific exemplar (e.g., a specific finger) (11%) — 4% said that any comparison that is not an individualization is an exclusion, and 23% said they do not use the term. However, most survey respondents (84%) said that they often conclude that a latent and the exemplars provided definitively did not come from the same source; only 3% never make such a conclusion ([2], summarized in Appendix SI-2.4).

This shift in standards for reporting conclusions has given rise to a new type of error: erroneous exclusions. Under the identification vs. non-identification approach, an examiner could err by making a "missed ID," failing to individualize two fingerprints that other examiners individualize. Missed IDs include not only erroneous exclusions, but also inconclusives and no value determinations on comparisons on which other examiners made individualization determinations. Using SWGFAST terminology, an erroneous exclusion is an error, because it can be shown to be demonstrably wrong; a missed ID is a non-consensus decision in which examiners disagree regarding whether there is sufficient support for an individualization decision.

Explicitly dividing the old non-identification determination into inconclusive and exclusion determinations reduces ambiguity, but in operational casework the distinction is often not important. Occasionally, the distinction between an inconclusive and an exclusion may be important for exculpatory evidence, if the latent is of high probative value (e.g., on the handle of a knife), or if the latent indicates that another person was present at a crime scene. However, the probative value of an exclusion is usually minimal because excluding a person does not mean that the person did not touch an object. In most casework, an exclusion has the same

---

* *Regarding the use of terminology — "latent print" is the preferred term in North America for a friction ridge impression from an unknown source, and "print" is used to refer generically to known or unknown impressions. We recognize that outside of North America, the preferred term for an impression from an unknown source is "mark" or "trace," and that "print" is used to refer only to known impressions. We are using the North American standard terminology to maintain consistency with our previous and future papers in this series [1-9]. See Glossary, Appendix SI-1.*

† *Note that there are additional unrelated uses for the term "exclusion" occasionally used in forensic contexts: the positive identification of a latent to an elimination print (e.g., officer, family member, victim), and the inadmissibility of evidence in court. The term "elimination" is sometimes used as a synonym of exclusion.*

operational implications as an inconclusive, and an erroneous exclusion usually has the same operational implications as a missed ID.

Deciding whether or not to exclude can be straightforward if the prints being compared are high quality and there are notable differences in the pattern classes or overall ridge flow. However, deciding whether or not to exclude may be more challenging if either the latent or exemplar is unclear, distorted, or incomplete: features and ridge flow can be misinterpreted in unclear prints; distortion can lead to extreme dissimilarity in mated prints (from the same person) [12,13]; incomplete or partial prints are susceptible to being erroneously excluded as the result of incorrect anchoring or localization (comparing the wrong areas).

Deciding whether to make an exclusion determination requires assessing whether dissimilarities are in fact due to true discrepancies. The distinction between these terms is important: a dissimilarity is a difference in appearance between two friction ridge impressions, but a discrepancy is an examiner's assessment that a dissimilarity originates in the skin itself and cannot be explained as an artifact or distortion. In the "one discrepancy rule" [12,14], any discrepancy is sufficient to exclude; over-eager application of this rule may lead to errors [13,15,16]. SWGFAST states that "The term discrepancy is only used as a description of incompatibility between two impressions that has resulted in a conclusion of exclusion," [12] and therefore per that definition the examiner's decision whether dissimilarities should be considered discrepancies is directly tied to the decision whether the comparison should be an exclusion.

Examiners can make exclusions based on differences in pattern classes or overall ridge flow (level 1 features), or minutiae and paths of individual ridges (level 2). Although exclusions can be based solely on differences in level-1 information, when there is significant distortion, differences in both level-1 and level-2 features are required [12]. Ridge edges and pores (level 3 details) cannot be the sole factor in exclusion determinations. After recent research studies reported a surprisingly high rate of erroneous exclusions [2,17,18], there has been more discussion of erroneous exclusions, often with examples of how distortion or other factors could make mated prints appear very different [e.g., 13]. Some agencies have begun to change the criteria for an exclusion. For example, three agencies in Arizona now require an anchor point (e.g., a core or delta) in both prints and discrepancies in both level-1 and level-2 details to render an exclusion: "Only after noting distinct differences in two or more target groups in their relation to the first-level anchor point does the examiner have sufficient disagreement to exclude." [16]

Errors and disagreements among examiners may be due in part to lack of guidance on the relative costs and benefits of each decision, or systematic pressures encouraging some decisions more than others. These pressures will vary by agency or among cases, and examiners' responses to these pressures will vary. For example, given a print of marginal suitability, an examiner must decide whether to compare or not. Approximately half of the Black Box survey respondents reported that they are either not permitted to make (32%) or discouraged from making (19%) an inconclusive determination if the latent and exemplar are both of value and include a large potentially corresponding area [2]. The rate of erroneous exclusions may be explained in part by environments in which some examiners felt discouraged from making inconclusive determinations and knew that exclusions would not be subjected to verification.

In response to the misidentification of a latent print in the 2004 Madrid bombing [19], a Federal Bureau of Investigation (FBI) Laboratory review committee evaluated the scientific basis of friction ridge examination. That committee recommended research, including the study described in this report: an evaluation of exclusions in the latent print examination process [20]. In light of the high erroneous exclusion rate reported on Black Box and other studies [17,18], and the recent interest in exclusion determinations [13,16], we have conducted additional analyses of data from the Black Box and White Box studies to understand the associations between a variety of factors and exclusion determinations, particularly those factors contributing to erroneous exclusions. The primary purpose of this research is to assist examiners in improving the examination process, and to provide information to the broader community regarding the accuracy, reliability, and implications of exclusion decisions.

2 Materials and methods

This report presents new analyses of data collected in the Black Box ("BB") studies [2,3] and White Box ("WB") studies [6,7,9]; the test procedure, participants, and fingerprint data are summarized in Appendix SI-1.

The Black Box study was designed to study the accuracy and reliability of examiners' conclusions (without insight into how they make those conclusions); it offers a much larger sample size. The White Box study was designed to study the bases for examiners' determinations; examiners provided detailed markup to reveal the information they relied upon to make decisions. In each study, practicing latent print examiners performed comparisons under test conditions designed to correspond to that part of casework in which a single latent is compared to a single exemplar print.

The prevailing latent print examination methodology is known as Analysis, Comparison, Evaluation, and Verification (ACE-V) [21,22]; the test workflow in both studies conformed to ACE-V, but did not include a Verification phase. During the analysis phase, only the latent was presented, and the examiner recorded a value determination of value for individualization (VID), value for exclusion only (VEO), or no value (NV). If VID or VEO, the examiner proceeded to the Comparison/Evaluation phase, in which the exemplar was presented for side-by-side comparison with the latent, and made an evaluation determination of individualization (the fingerprints came from the same finger), exclusion (the fingerprints did not come from the same finger), or inconclusive (neither individualization nor exclusion is possible). Examiners were required to rate the difficulty of each comparison. When an exclusion determination was made, the examiner was required to select a reason for the exclusion from a short list of options. Detailed descriptions of the materials and methods for these studies are reported in [2,3,6] and summarized in Appendix SI-1.

In both studies, latent-exemplar image pairs were selected to be challenging, similar to casework in which highly similar candidate exemplars are returned by an Automated Fingerprint Identification System (AFIS). However, there were important differences in how image pairs were selected that affect the overall rates measured in the two studies (details in Appendix SI-3.1). In Black Box, all image pairs were collected under controlled conditions so that they could be known definitively to be mated (from the same source) or nonmated (from different sources); the latents included a broad range of quality, including a greater proportion assessed by participants as NV. In White Box, because the objective was to investigate the bases for determinations (rather than their accuracy), a wider variety of attributes (such as substrate and processing methods) were included, and some of the image pairs were collected from operational data; selection of mated image pairs was designed to focus on the threshold between individualization and inconclusive. In surveys of participants, a large majority of BB and WB respondents agreed that the fingerprints were representative of (or similar to) casework, and that the overall difficulty of comparisons was similar to casework [2,6].

The Black Box study included a main test in which each examiner (n=169) was assigned 100 image pairs; in a subsequent repeatability test, 72 of those examiners were reassigned 25 of those image pairs. Together, these tests yielded responses to 17,121 distinct presentations of image pairs. In the White Box study, each examiner (n=170) was assigned 22 image pairs for a yield of 3730 valid responses. Additional details regarding test sizes are included in Appendix SI-2.2.

3 Overview of exclusion concepts

This section provides an overview of exclusion concepts and rates from BB and WB, to serve as a baseline for understanding the results presented in Section 4, which focus specifically on the factors associated with exclusions.

3.1 False negative and true negative rates

We refer to the exclusion of a mated pair as a false negative (FN) and the exclusion of a nonmated pair as a true negative (TN). We refer to false negatives as "erroneous" because those conclusions contradict ground truth, but we avoid referring to true negatives as "correct" because we have no absolute criteria to judge whether an inconclusive determination would have been more appropriate. True and false negative rates can be reported in two ways:

- For factors associated with latents (e.g., image quality, analysis minutiae counts), we report proportions of all mated or nonmated ***presentations*** (i.e., including NV determinations) that resulted in exclusions (indicated by TNR_{PRES} and FNR_{PRES}).
- For factors associated with comparisons (e.g., comparison difficulty, corresponding minutiae), we report proportions of mated and nonmated ***comparisons*** (i.e., omitting NV determinations) that resulted in exclusions (indicated by TNR_{CMP} and FNR_{CMP}).

	FNR		TNR	
	PRES	CMP	PRES	CMP
BB	5.3%	7.5%	71.2%	79.2%
WB	4.5%	5.5%	50.7%	73.9%

Table 1: Overall exclusion rates for BB (5543 presentations, 4985 comparisons) and WB (848 presentations, 582 comparisons). Detailed determination counts and rates in Appendix SI-2.2.

Table 1 summarizes exclusion rates for BB and WB. These rates are similar to those reported in other studies [18,23,24]. However, we know that exclusion rates can vary greatly by examiner and depending on the specific images being compared. Differences in mean exclusion rates between WB and BB can generally be explained by differences in participants, test procedures, how image pairs were selected — and in differing distributions of the factors we discuss in Results. BB results were published prior to WB, and in particular the high FNR was widely discussed; therefore WB participants may have changed their behavior in response. The lower FNR on White Box may also be attributable to differences in how examinations were performed as a consequence of WB requiring markup. On a common subset of the data, the higher FNR on BB was statistically significant, but the difference in TNR was not. See Appendix SI-3.1 for supporting information on the effects of data selection.

3.2 Value for exclusion only

Although exclusion is a determination made during comparison and evaluation of a latent with an exemplar, examiners first assess the potential for exclusion during the analysis of the latent by itself. Agencies differ in their handling of VEO latents (latents that are not suitable for individualization but could potentially be used for exclusion). In the Black Box survey of participants, 55% reported that their standard operating procedures did not differentiate between VEO and NV; 14% did not differentiate between VEO and VID; the remainder had a separate VEO category that they used in standard practice (17%) or only upon request (13%). In the BB survey, those agencies that did not differentiate between VEO and NV usually discouraged or did not permit use of inconclusive as a comparison determination (survey results in Appendix SI-2.4). The associated errors and error rates will differ depending upon which approach is taken: VEO latents are generally poor quality and are disproportionately likely to result in inconclusives. Differing practices in how VEO latents are normally handled may have contributed to inter-examiner variability in value assessments seen in these tests. Some examiners appear to have used VEO to mean "limited value," as evidenced by individualizations made on latents assessed as VEO. The concept of VEO may be appropriate to reconsider: VEO is based on the concept that latents suitable for exclusion are a superset of those suitable for individualization; however, not all latents suitable for identification are suitable for exclusion, and vice versa [16].

3.3 Support for exclusion vs. individualization

During comparison, an examiner assesses the amount of information supporting individualization and the amount of information supporting exclusion, then decides if there is sufficient support for either determination; if there is not sufficient support for either, the determination will be inconclusive. One indication we have for how much support there was for each determination is interexaminer agreement on the final determinations. Each image pair was examined by multiple examiners (average of 23 in BB; 12 in WB). Their determinations can be regarded as a measure of consensus, as shown in Figure 1: the x axis indicates the percentage of examiners who determined that there was a sufficient basis for individualization, and the y axis indicates the percentage of examiners who determined that there was a sufficient basis for exclusion. These "votes" can be thought of as describing points in a continuum in which each examiner must make decisions : for example, although no examiner is telling us that (for a specific comparison) there is 60% support for individualization and 5% support for exclusion, we can see that 60% of examiners felt that there was sufficient support for individualization and 5% felt there was sufficient support for exclusion. In WB, examiners marked corresponding minutiae so that we had insight into how each examiner evaluated the extent of support for ***individualization***. However, the markup often provided little or no insight into how each examiner evaluated the extent of support for ***exclusions***, and therefore, voted results provide the best information we have available as to the sufficiency for exclusion.

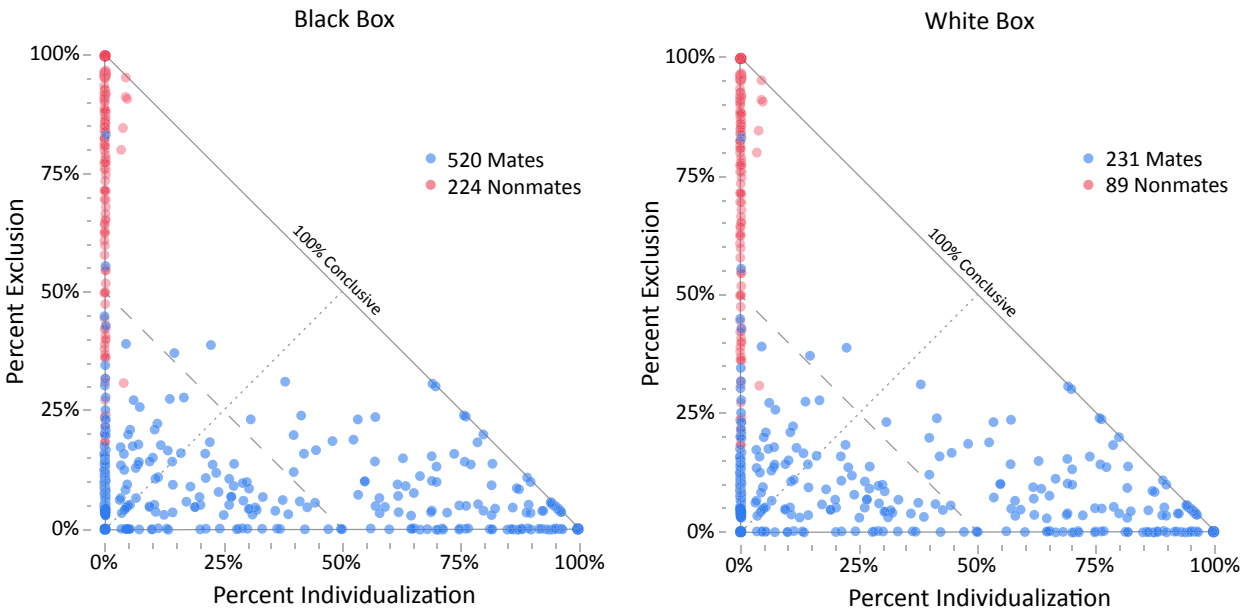

Figure 1: Determination rates on each image pair in BB (mean of 23 examiners per image pair) and WB (mean 12 examiners per image pair). Points at the origin represent image pairs that examiners agreed unanimously could neither be excluded nor individualized; points at the bottom right were unanimous individualizations; points at the top left were unanimous exclusions; BB and WB differed notably in the number of unanimous determinations. NV is treated as inconclusive. 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. Left graph is reproduced from [2].

Figure 1 shows that the distributions of determinations by image pair were similar on BB and WB. For many mated image pairs (blue), there was a great deal of disagreement among examiners regarding whether to individualize (true positive), exclude (false negative), or be inconclusive. For nonmated image pairs (red), there were few individualizations (false positives) and therefore almost all of the variation was regarding whether to exclude (true negative). What these charts do not reveal is that the proportions of unanimous determinations (superimposed data points in the three corners of each chart) were notably different on the two tests: the proportions of unanimous decisions are greatly influenced by data selection (details in Appendix SI-3.1).

Erroneous exclusions are sometimes confused with missed IDs, which we define as an exclusion, inconclusive, or NV determination on an image pair that the majority of examiners individualized. In BB, 4.7% of responses on mated pairs were missed IDs (WB, 9.4%); in BB, 27% of missed IDs were erroneous exclusions (WB, 20%); details in Appendix SI-3.

Prior to the Black Box study, we would have expected erroneous exclusions to be concentrated on a small subset of the mated image pairs. This expectation was shown to be incorrect. Erroneous exclusions were widely distributed across the image pairs tested — although they were more likely to occur on some image pairs than others, as we will explore in Results. To a first approximation, modeling erroneous exclusions as random events that are equally likely to occur on any mated comparison provides a good description of our data (Appendix SI-4). Erroneous exclusions were made by at least one examiner on 46% of BB mated image pairs and 35% of WB mated image pairs; a greater proportion of BB mated pairs were erroneously excluded by at least one examiner than WB pairs because each image pair was presented to more examiners on BB than on WB (mean of 22 examiners per image pair on BB vs. 12 on WB). Many of the mated image pairs that were not excluded by any examiner were unanimously NV (10% of BB, 0% of WB) or unanimously ID (10% of BB, 23% of WB).

The (inter-examiner) reproducibility of true negatives was much higher than that of false negatives: in BB 87% of true negatives were reproduced (71% in WB), but only 15% of false negatives (11% in WB). Most erroneous exclusions would not have been independently corroborated if they were blind verified: in BB, we estimated FNR after blind verification to be 0.85% [2]. However, blind verification (and even non-blind verification) of

exclusions is not standard practice in many organizations, and therefore the initial erroneous exclusions would remain undetected in most cases (details in Appendix SI-3.2). In BB we showed that the lack of reproducibility of determinations is related to the lack of (intra-examiner) repeatability of determinations: when examiners were retested after seven months, 91% of true negatives were repeated, but only 30% of false negatives [3].

3.4 Negative predictive value

Measuring true and false negative rates requires definitive knowledge of which image pairs are mated, which of course is not feasible in operational casework. In casework, we would like to know how often exclusions are correct and under which circumstances they are more or less likely to be correct. Negative predictive value (NPV) refers to the proportion of exclusions that are true negatives. This rate depends substantially on the prevalence of mated pairs among the examinations performed: as shown in Figure 2, as the proportion of mated pairs increases, NPV decreases because a larger proportion of the exclusion determinations will be made on mated pairs. It is therefore essential to account for differences in mating proportions when comparing NPV across datasets. As described in [2] and Appendix SI-15, we can extrapolate a measured NPV to any arbitrary proportion of mated vs. nonmated comparisons based on the separately measured true and false negative rates. In order to compare the effects of a given factor on NPV, we first normalize the results by projecting NPV to equal proportions of mates and nonmates (NPV_{50}). This projection requires knowing a priori for each level of each factor the proportion of comparisons that were mated: for example, we can normalize the NPV estimates for BB latent value assessments because we know that 68% of VEO latents were mated and 83% of VID latents were mated, and therefore we can project our estimates to what NPV would have been if each were 50% mated (using the method described in Appendix SI-15).

Figure 2 shows the results from both tests extrapolated over the full range of possible mating proportions.

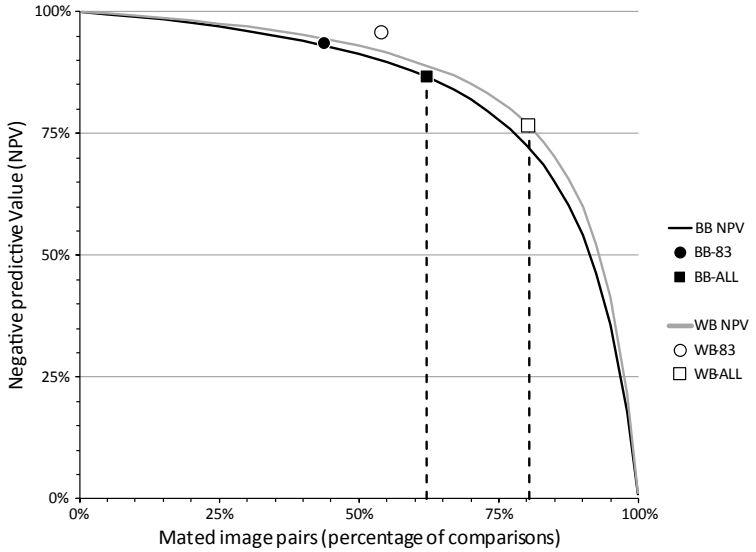

Figure 2: NPV as measured at actual test proportions of mate and nonmate comparisons (markers), and as extrapolated as a function of the mating proportion (curves). A subset of 83 image pairs included in both tests is also indicated (which allows comparing the tests while controlling for differences in data selection). The black curve extrapolates from BB where 62% of all comparisons were performed on mated pairs, $NPV_{62\%}$=86.6%. The gray curve extrapolates from WB where 80% of all comparisons were performed on mated pairs, $NPV_{80\%}$=76.6%.

4 Results

In this section, we discuss factors associated with exclusion rates in order to understand why examiners exclude and when they make erroneous exclusions. We first discuss several measures describing the information available in the latent alone (quality, value, and number of analysis minutiae). We then discuss measures of the comparison of the latent and exemplar (the reasons examiners gave for their exclusions,

discrepancies, corresponding minutiae, corresponding cores and deltas, comparison difficulty). Finally, we discuss the extent to which true and false negative rates can be attributed to individual examiner differences.

4.1 Latent quality and value

Latent quality metrics and examiner value determinations are both assessments of the quality and quantity of information in the latent itself, separate from the comparison. Any measure assessing the latent alone will be an imperfect predictor of exclusion rates because it does not account for the quality of the exemplar or the overlap between the latent and exemplar.

The FBI's Latent Quality Metric (LQMetric)‡ automatically assesses the quality of latent fingerprint images, based on a variety of factors such as clarity, continuity of ridge flow, and quality and quantity of minutiae. LQMetric estimates the probability that a latent would hit if searched in the FBI's Next Generation Identification (NGI) AFIS (specifically, the probability that an image-only (LFIS) search would return a mate as the rank 1 candidate if the subject were in the database). For example, an LQMetric value of 80 predicts that if the subject is present in the database, there is an 80% probability that a mate would be returned at rank 1. This ability to match on an automated system is similar to but not always the same as how an examiner would assess the quality or value of a latent.

Figure 3 shows the relations between LQMetric and examiner determinations (additional data in Appendix SI-7). As LQMetric increases, the proportion of NV latents decreases, as does the proportion of inconclusive comparisons. On nonmated image pairs, we see that TNR_{PRES} generally increases with LQMetric: as the available quantity and quality of information in the latents increased, examiners were more likely to make an exclusion determination. On mated image pairs, however, we see higher error rates (FNR_{PRES}) on intermediate quality latents: very poor-quality latents tend not to be compared or result in inconclusives; very high-quality latents tend to be individualized. NPV increases as LQMetric increases, as a result of the increasing true negative rates among comparisons (TNR_{CMP}) and relatively flat false negative rates (FNR_{CMP}).

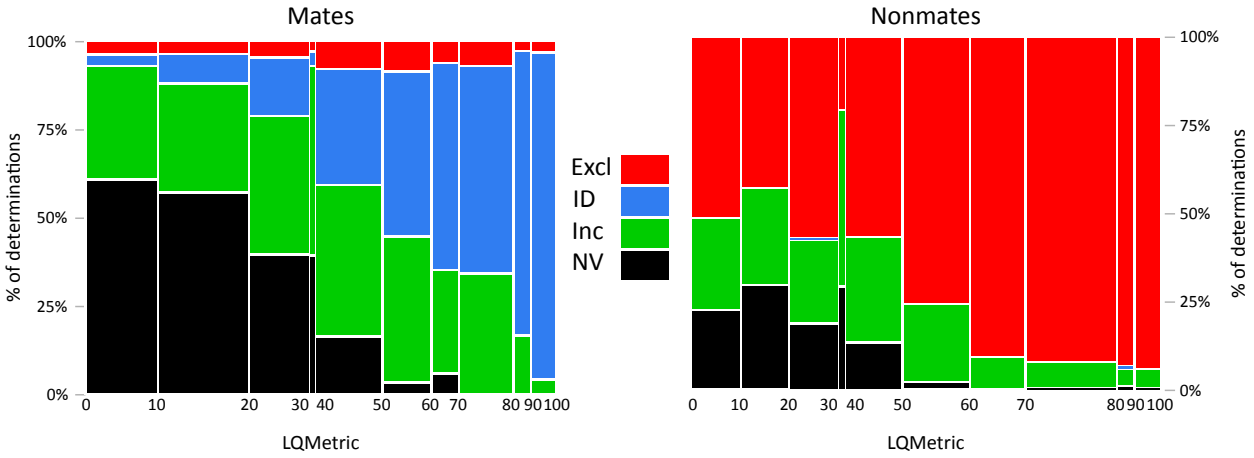

Figure 3: Mosaic plots of the distribution of determinations by LQMetric, for mates and nonmates (BB, n=11,578 mated and 5,543 unmated image pair presentations).

Inter-examiner reproducibility of true negatives increases with LQMetric; the reproducibility of false negatives is low regardless of quality, but is higher on intermediate quality latents (Appendix SI-8).

Examiner's value assessments provide information similar to LQMetric, because value and LQMetric are correlated: most VEO latents have an LQMetric below about 45, and most VID latents have an LQMetric above 45 (Appendix SI-7). On nonmated comparisons, we observe the expected result that TNR is much higher on latents assessed as VID than on latents assessed as VEO (BB: TNR_{VID}=89% vs. TNR_{VEO}=36%; WB: TNR_{VID}=82% vs. TNR_{VEO}=56%; details in Appendix SI-7). On mated comparisons, we did not observe a notable association

‡ *LQMetric is included in the FBI's Universal Latent Workstation (ULW) software [ULW], release 6.5 or later.*

between latent value assessments (VEO vs. VID) and FNR. However, because we included relatively few very high-quality latents, the difference in exclusion rates between VEO and VID latents was limited.

Among VID latents, LQMetric provides gradations that effectively predict which mated comparisons are more or less likely to result in individualizations; among VEO latents, exclusion rates did not vary notably with LQMetric; and at any LQMetric value, examiners were much more likely to make a conclusive comparison determination on latents rated VID than those rated VEO.

4.2 Minutiae marked during analysis

Figure 4 shows the association between exclusion rates and the number of minutiae marked on the latent during analysis ("analysis minutiae") in WB. For nonmates, TNR increases with the number of minutiae. When zero or very few analysis minutiae were marked, the latent determination was usually NV, and therefore there were few exclusions. True negatives occurred at low minutia counts: among latents with zero analysis minutiae (n=69) were five exclusions; among latents with 1-3 analysis minutiae (n=124) were 16 exclusions. The majority of nonmates with seven or more analysis minutiae were excluded, as was every nonmated latent with at least 20 analysis minutiae (n=33).

For mates, FNR was zero or near zero for low and very high minutia counts. No mated latent with more than 28 analysis minutiae (n=145) was excluded. Only one erroneous exclusion occurred with fewer than four analysis minutiae (n=288).

Broadly, these trends are very similar to those described for LQMetric: TNR and NPV increase with the quality of the latent, and FNR is lower for the best and worst quality latents. This finding is corroborated by Pacheco, et al.] who reported TNR increasing with "Strength of Value" and FNR peaking at the middle level of "Difficulty;" both of these measures were based largely on minutia counts.

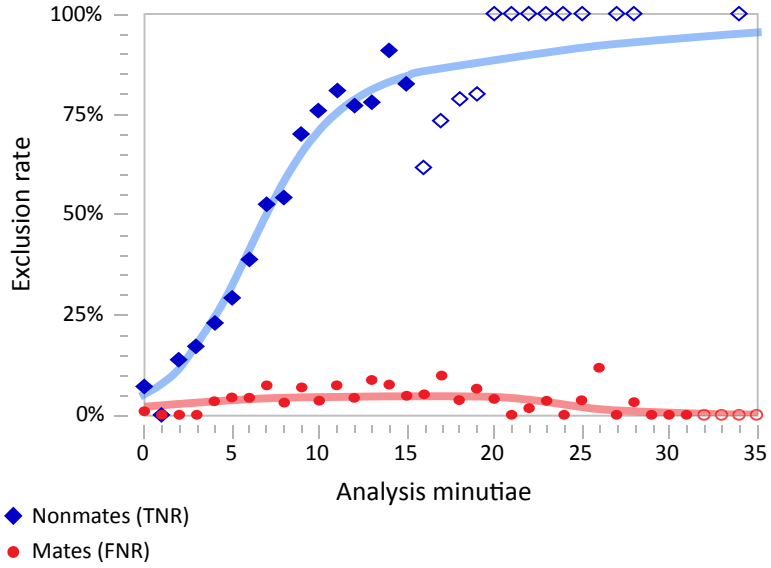

Figure 4: True and false negative rates by the number of minutiae marked on latents during analysis (WB). Each marker represents an exclusion rate (true or false) calculated over all nonmated or mated presentations for the specified minutia count. Open markers indicate rates measured on fewer than 20 presentations. Piecewise cubic polynomial splines were fit to 3730 minutia counts and determinations (logistic regression using the technique of knotted splines [25] as implemented in SAS JMP 11, using 3 knots). (n=848 for TNR_{PRES}; n=2882 for FNR_{PRES}). Data is shown truncated at 35 minutiae: all nonmated data is shown; 1% of mated data is not shown (with no false negatives).

4.3 Reasons for exclusions

The factors discussed above (quality, latent value, analysis minutiae) are all assessments of the latent alone. Here and in the following sections we assess factors associated with the comparison of each latent and exemplar.

Examiners were asked to indicate what observed differences in the prints led to each exclusion determination by selecting one of the options listed in Table 2; the options provided on White Box were designed to further partition those on Black Box. Interexaminer reproducibility of exclusion reasons was low (Appendix SI-6). Examiners usually attributed exclusions to minutia differences regardless of whether their exclusions were erroneous (mated) or not (nonmated).

Pattern class was cited as the reason for a greater proportion of false negatives than true negatives in BB. However, the proportion of exclusions based on pattern class differences may be influenced by data selection, which differed for mates and nonmates and between the two tests.

	Black Box				White Box			
	Mates		Nonmates		Mates		Nonmates	
Pattern class/ridge flow	174	28%	624	16%	-	-	-	-
Pattern classes differ	-	-	-	-	12	9%	37	9%
Core or delta differences	-	-	-	-	8	6%	42	10%
Minutiae and/or level 3	437	72%	3323	84%	-	-	-	-
One or more minutiae differ	-	-	-	-	104	80%	343	80%
Level 3 features differ	-	-	-	-	3	2%	3	1%
Other	-	-	-	-	3	2%	5	1%
Total exclusions	611		3947		130		430	

Table 2: Distribution of exclusion reasons. Categories are defined in Appendix SI-6.§

The repeatability of false negatives was higher when based on pattern class/ridge flow differences (41%) than when based on minutiae or level-3 features (26%) (details in Appendix SI-6).

In WB, examiners were given the opportunity to elaborate on the exclusion reason with a short text response, ten of which (among 49 provided) appear to justify an inconclusive determination rather than exclusion (examples in Appendix SI-6). We assume that these (and possibly other) erroneous exclusions were due to examiners confusing the concepts of exclusion and non-identification.

4.4 Discrepancies and corresponding minutiae

In WB, examiners were instructed to mark any discrepancies used to support an exclusion determination. Marking of discrepancies was not notably associated with whether the latent and exemplar were mated: examiners marked discrepancies on 31% of false negatives and 37% of true negatives. Even when the exclusion reason was that minutiae differed, examiners marked discrepancies on only 40% of exclusions. Reproducibility of discrepancies was not substantially greater than chance [9]. Discrepancies were marked in 6% of inconclusives. Therefore, marked discrepancies did not provide much insight into how examiners assess sufficiency for exclusion — unlike sufficiency for individualization (which is reasonably well-described by the number of corresponding minutiae) (details in Appendix SI-9).

Examiners were able to indicate definitive and debatable correspondences between the latent and exemplar — and for exclusions were instructed to to mark anchors (reference points) used to establish discrepancies as debatable correspondences. Debatable correspondences were marked on about 15% of exclusions (both true and false negatives). However, examiners marked definitive correspondences on 30% of false negatives, and 16% of true negatives; seven or more were marked on 12% of false negatives but on only two true negatives (0.5%). All of the exclusions with nine or more corresponding minutiae marked (n=8) were erroneous: three false negatives had 15-17 corresponding minutiae marked (details in Appendix SI-9).

We reviewed erroneous exclusions in order to understand the factors contributing to the errors. Among those responses where the reasons and markup were adequate to understand the basis, we found that erroneous exclusions were generally caused by one of the following:

- Misinterpreted pattern class due to distortion, inadequate overlap, or insufficient area (indicated by examiners citing pattern class differences, or core or delta differences);
- Incorrect anchoring ("corresponding" minutiae in the wrong regions, or incorrectly rotated images);
- Incorrect ridge counting or misinterpretation of distortion resulting in false "discrepancies" (only portions of the image have markup in agreement with other examiners); or

§ One WB exclusion is omitted because no exclusion reason was recorded.

- Inappropriate use of the "one discrepancy" rule (exclusions made despite high numbers of corresponding minutiae, e.g., nine or more).

After the initial analysis of a latent print, examiners sometimes revised their markup of the latent during comparison with the exemplar (previously reported in [7]). Examiners deleted and added a greater proportion of their marked minutiae on individualizations than inconclusives, and a greater proportion on inconclusives than exclusions. Among exclusions (and inconclusives), added minutiae were more common on mated pairs than nonmated pairs, in unclear areas than clear areas, and on difficult comparisons than easy comparisons. Overall, the rate at which examiners added minutiae was about twice as high on false negatives as on true negatives (8.5% vs. 4.6% increase in minutia count); the rate at which examiners deleted minutiae was similar for true and false negatives (3%) [7].

4.5 Cores and deltas

Making an exclusion is generally more straightforward if a core or delta is present in both the latent and exemplar. In WB, examiners often did not mark cores and deltas that were present on the latent; similarly, they usually did not mark those features as corresponding, especially when excluding or inconclusive (see discussion in Appendix SI-8). Therefore, in this analysis we used data from a pretest screening process that indicated whether a core or delta was present in both the latent and exemplar. We found that FNR was lower on those image pairs that had a core or delta than those that did not (FNR_{CMP} = 3.4% vs. 8.7%) and TNR was higher when a core or delta was present (TNR_{CMP} = 80.0% vs. 66.1%). Therefore, NPV was much higher when a core or delta was present in both the latent and exemplar: WB NPV_{50} was 96% when a core or delta was present vs. 88% when no core or delta was present.

Examiners did not often cite core or delta differences as the exclusion reason. Indicating core or delta differences as an exclusion reason was not significantly associated with errors (Table 1).

4.6 Difficulty

Examiners were asked to rate the difficulty of each comparison on a five-level scale from very easy to very difficult. The more difficult an examiner described a comparison, the more likely that that examiner's comparison determination was inconclusive. TNR dropped markedly with increasing difficulty (e.g., TNR_{CMP} dropped from 99% (very easy) to 51% (very difficult) on BB, and from 87% to 36% on WB). On mated pairs involving high-quality latents (high LQMetric), false negative errors were more common on difficult comparisons than on easy comparisons; however, on mated pairs involving low-quality latents, false negative errors were more common on easy comparisons than on difficult comparisons (details in Appendix SI-11).

Largely as a consequence of the relatively strong association between TNR and difficulty, both studies clearly show NPV decreasing with increasing difficulty of the comparison: difficult exclusions were more likely to be erroneous than were easy exclusions. However, we do not project NPV_{50} based on difficulty because we are concerned that difficulty may be assessed differently depending on the determination and therefore may be confounded with mating. (See Appendix SI-11 and Appendix SI-15 for additional data and discussion).

The processes by which image pairs are selected determines the range of difficulty of comparisons. We only included nonmated pairs that had highly similar pattern classes; if instead we selected nonmated image pairs at random from the general population, the vast majority would have unrelated pattern classes, resulting in a much greater proportion of very easy exclusions, and therefore TNR_{CMP} would be expected to be much higher.

4.7 Summary of factors associated with true and false negatives

Figure 5 and Figure 6 summarize the findings discussed above and compare the relative strength of association of each factor with TNR and FNR. Overall, we see clear trends in the TNR data whereas the trends in the FNR data are less clear. No single factor stands out as superior for explaining when examiners exclude (details in Appendix SI-13).

Figure 5 shows that TNR_{CMP} generally increases with increasing latent quality (as measured by LQMetric, value assessment, median analysis minutiae, and the presence of a core or delta), and ease of comparison.

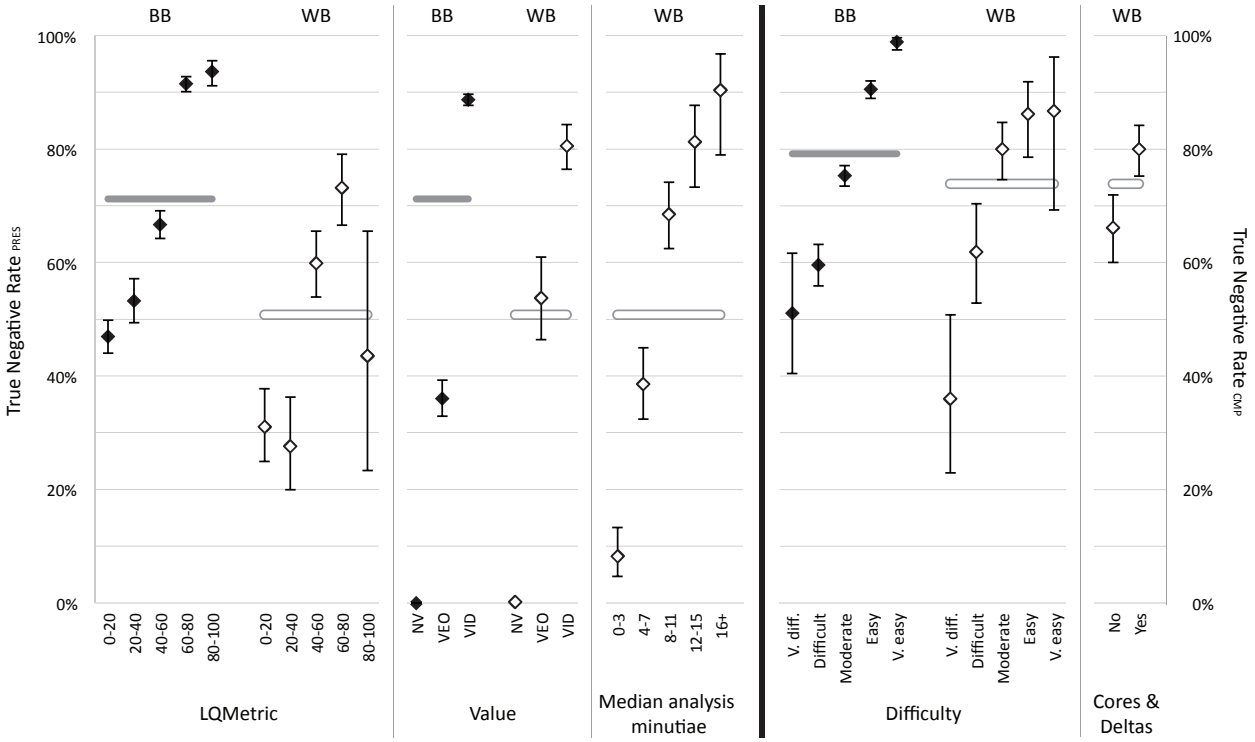

Figure 5: Associations between TNR and factors. Vertical bars indicate 95% binomial confidence intervals. Factors measured on latents (LQMetric, value, median analysis minutiae) are based on all presentations (TNR_{PRES}); difficulty and cores & deltas are based on comparisons (TNR_{CMP}). Horizontal lines indicate overall mean TNR: Black Box TNR_{PRES}=71.2% (5543 presentations), TNR_{CMP}=79.2% (4985 comparisons); White Box TNR_{PRES}=50.7% (848 presentations), TNR_{CMP}=73.9% (582 comparisons).

As seen in Figure 6, the associations are not as strong for FNR as we saw for TNR. One reason that the highest quality latents (high LQMetric, high minutia counts, and the presence of a core or delta) are associated with relatively low FNR is that these latents were usually individualized (e.g., see Figure 3). The lowest quality latents are also associated with relatively low FNR_{PRES}, because these latents usually resulted in NV or inconclusive determinations; FNR_{CMP} is not low on these latents (Appendix SI-13).

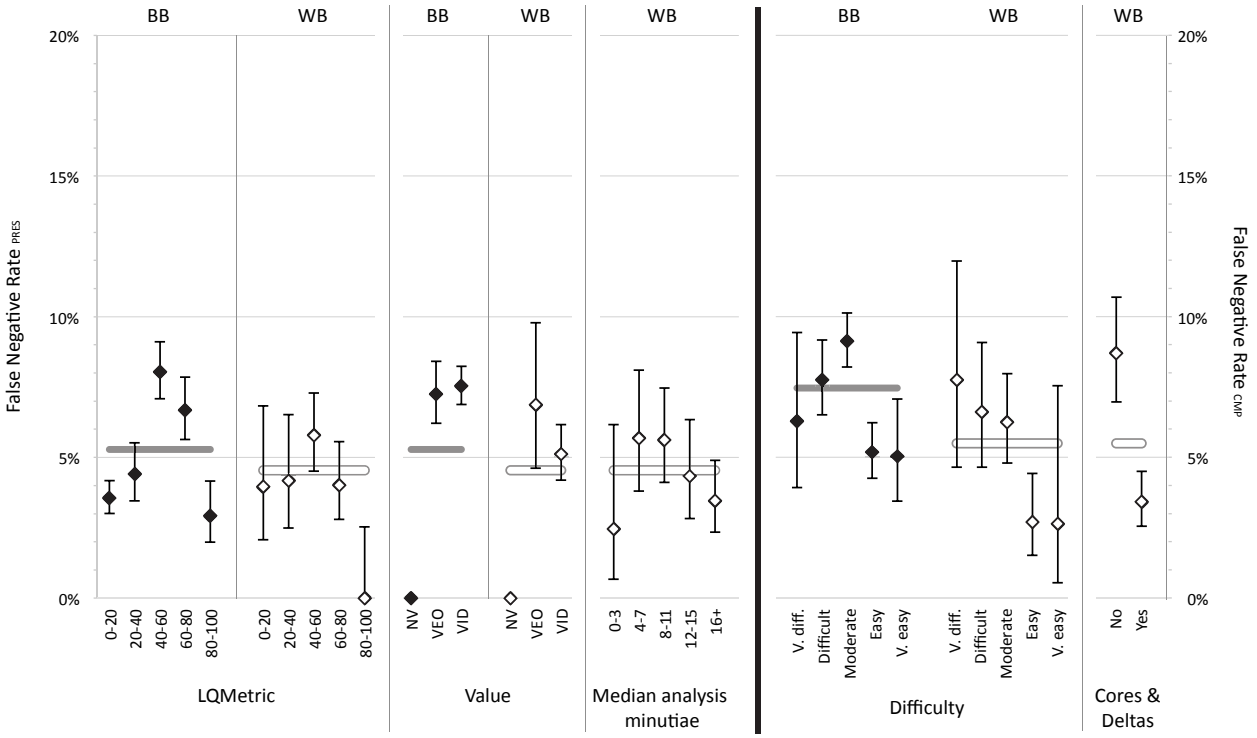

Figure 6: Associations between FNR and factors. Vertical bars indicate 95% binomial confidence intervals. Factors measured on latents (LQMetric, value, median analysis minutiae) are based on all presentations (FNR_{PRES}); difficulty and cores & deltas are based on comparisons (FNR_{CMP}). Horizontal lines indicate overall mean FNR: Black Box FNR_{PRES}=5.3% (11,578 presentations), FNR_{CMP}=7.5% (8189 comparisons); White Box FNR_{PRES}=4.5% (3730 presentations), FNR_{CMP}=5.5% (2966 comparisons).

Figure 7 summarizes the associations of these factors with NPV_{50}. Each measure of latent quality is a strong predictor of NPV: exclusion determinations are more likely to be correct when the latents are high quality. Similarly, exclusion determinations are more likely to be correct when a core or delta is present in both prints. We cannot normalize our estimates of NPV for difficulty because we do not know a priori the mated proportions for each difficulty level; nevertheless, we found that NPV decreased substantially with difficulty. Additional NPV data is presented in Appendix SI-13 and Appendix SI-15.

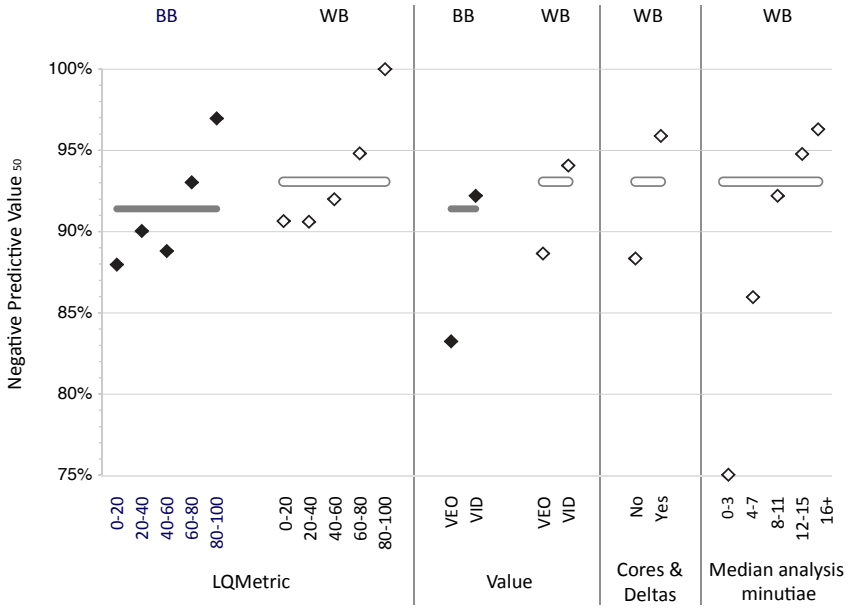

Figure 7: Associations between NPV_{50} and factors. Horizontal lines indicate overall NPV: Black Box NPV_{50} = 91.4% (n=4558 exclusions); White Box NPV_{50} = 93.1% (n=561 exclusions). Confidence intervals were not estimated for lack of a standard method and because of debatable modeling assumptions.

In addition to the factors presented here, we looked for an association between finger position and erroneous exclusions. With the possible exception of a higher FNR on left index fingers, no significant association was detected (Appendix SI-13).

4.8 Examiner effects

Image-based metrics cannot fully account for variability in exclusion rates because examiner determinations are not always unanimous. Examiners differ substantially in true and false negative rates: some examiners make erroneous exclusions at nearly double the average rate, while many others had FNRs that were substantially lower than the group mean (Appendix SI-6.1). Examiners' false negative rates were not strongly correlated with their true negative rates, and differences among examiners in FNR could not be accounted for as a consequence of differences in their overall conclusion rates (after omitting those comparisons resulting in erroneous exclusions).

The relatively high overall FNR on BB and WB was not due to just a few outlier examiners. In BB, 85% of examiners made at least one erroneous exclusion — although 65% of participants said that they were unaware of ever having made an erroneous exclusion after training. In WB, only 44% of examiners made any erroneous exclusions on the test, which is consistent with the fact that each examiner was assigned fewer image pairs than on BB and therefore had fewer opportunities to make errors.

In BB and WB, participants completed a background survey to assess their experience and the types of standard operating procedures they follow in casework. The BB survey included several questions germane to exclusions; responses are summarized in Appendix SI-2.4. No notable relations were found between erroneous exclusions and the survey responses related to exclusions. Certified examiners had higher TNR than non-certified examiners; otherwise years of experience and certification were not effective at discriminating exclusion performance among practicing latent print examiners (details in Appendix SI-6.2).

The results from our studies and others [e.g., 18] demonstrate that practical tests could be designed to compare the performance (including true and false negative rates) of individual examiners. By selecting image pairs on which examiners do not make unanimous determinations, it would be relatively straightforward to select test data that would efficiently differentiate among examiners.

5 Conclusions

We have shown how errors are related to specific attributes of latent prints in such a way as to enable more focused training and proficiency testing. Our findings will also help researchers interpret results that may appear to differ across studies and to sample fingerprints appropriately for use in future experimental designs.

By studying the performance of experts and focusing on their errors and disagreements, we increase our understanding of the factors that affect the accuracy and reproducibility of exclusions. Our findings suggest ways to improve training and thus the performance of individual examiners. However, after putting the examiner under the microscope, it is important to step back and consider the context in which the examiner works, and consider broader systems approaches to managing errors and disagreements. The concept and terminology of exclusion vary among agencies. The participants in these studies came from agencies with differing policies with respect to whether and how exclusions are used, whether exclusions are verified, whether examiners exclude because they are discouraged from making inconclusive decisions, and how latents of value for exclusion only should be treated. Some of the erroneous exclusions may be due to lack of familiarity with the concept of exclusion; some examiners apparently confused exclusions and non-identifications.

The lack of a standard method to precisely document the basis for exclusions contributed to examiners not consistently providing the detailed information needed to evaluate the extent of support for exclusions [18]. We assume that the same limitation exists operationally, which may have an impact on quality assurance, potentially making it difficult to detect questionable exclusions and impeding the verification of difficult decisions. Marking of discrepancies and other reference points is invaluable in trying to understand the basis for an exclusion. If the reason for exclusion is that the pattern classes differ, more detailed documentation of the rationale may not be necessary. However, if differences in minutiae or level-3 details are cited as the reason for exclusion, it becomes important to document these differences.

Requiring examiners to distinguish between inconclusive and exclusion determinations reduces ambiguity, but requires additional effort during examination. Given that in most operational casework the distinction is not important, is there truly a need to make this distinction in all cases as required by current guidelines? In some casework, such as in AFIS candidate review, there may be a reason to reconsider whether examiners should be given the option of non-identification when further differentiation is not needed.

6 Acknowledgments

We thank the latent print examiners who participated in these studies. This is publication number 16-24 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. The Universal Latent Workstation and LQMetric were developed by Noblis for the FBI CJIS Division. This work was funded in part under a contract award to Noblis, Inc. from the FBI Biometric Center of Excellence and in part by the FBI Laboratory Division. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the U.S. Government.

7 References

1 Hicklin RA, et al. (2011). Latent fingerprint quality: a survey of examiners. Journal of Forensic Identification, 61(4): 385-419.

2 Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2011). Accuracy and reliability of forensic latent fingerprint decisions. Proc Natl Acad Sci, 108(19): 7733-7738. http://dx.doi.org/10.1073/pnas.1018707108

3 Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2012). Repeatability and reproducibility of decisions by latent fingerprint examiners. PloS ONE, 7(3), e32800. http://dx.doi.org/10.1371/journal.pone.0032800

4 Hicklin RA, Buscaglia J, Roberts MA (2013). Assessing the clarity of friction ridge impressions. Forensic Sci Int, 226(1):106-117. http://dx.doi.org/10.1016/j.forsciint.2012.12.015

5 Ulery BT, Hicklin RA, Kiebuzinski GI , Roberts MA, Buscaglia J (2013). Understanding the sufficiency of information for latent fingerprint value determinations. Forensic Sci Int, 230(1): 99-106. http://dx.doi.org/10.1016/j.forsciint.2013.01.012

6 Ulery BT, Hicklin RA, Roberts MA, Buscaglia J (2014). Measuring what latent fingerprint examiners consider sufficient information for individualization determinations. PLoS ONE, 9(11), e110179. http://dx.doi.org/10.1371/journal.pone.0110179

7 Ulery BT, Hicklin RA, Roberts MA, Buscaglia J (2014). Changes in latent fingerprint examiners' markup between Analysis and Comparison. Forensic Sci Int, 247: 54-61. http://dx.doi.org/10.1016/j.forsciint.2014.11.021

8 Kalka ND, Hicklin RA (2014). On relative distortion in fingerprint comparison. Forensic Sci Int, 244: 78-84. http://dx.doi.org/10.1016/j.forsciint.2014.08.007

9 Ulery BT, Hicklin RA, Roberts MA, Buscaglia J (2016). Interexaminer variation of minutia markup on latent fingerprints. Forensic Sci Int. 264:89-99. http://dx.doi.org/10.1016/j.forsciint.2016.03.014

10 National Research Council (2009) Strengthening Forensic Science in the United States: a path forward (National Academis Press, Washington, DC). https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf

11 SWGFAST (2013). Standard Terminology of Friction Ridge Examination (Latent/Tenprint Document #19), Ver. 4.1. http://swgfast.org/documents/terminology/121124_Standard-Terminology_4.0.pdf

12 SWGFAST (2013). Standards for Examining Friction Ridge Impressions and Resulting Conclusions (Latent/Tenprint Document #10). http://www.swgfast.org/documents/examinations-conclusions/130427_Examinations-Conclusions_2.0.pdf.

13 Triplett M (2012) The Need to Validate Principles and the Value of Reproducible Results. IDentification News, 42-4. Aug. 2012

14 Thornton JI (1977) One-dissimilarity doctrine in fingerprint identification. International Criminal Police Review, 306: 89-95.

15 Expert Working Group on Human Factors in Latent Print Analysis (2012). Latent print examination and human factors: improving the practice through a systems approach. U.S. Department of Commerce, National Institute of Standards and Technology http://www.nist.gov/oles/upload/latent.pdf

16 Ray E, Dechant PJ (2013) Sufficiency and Standards for Exclusion Decisions, Journal of Forensic Identification, 63 (6): 675–697.

17 Langenburg G (2009) A Performance Study of the ACE-V Process: A Pilot Study to Measure the Accuracy, Precision, Reproducibility, Repeatability, and Biasability of Conclusions Resulting from the ACE-V Process, Journal of Forensic Identification, 59 (2): 219–257.

18 Neumann, C., Champod, C., Yoo, M., Genessay, T., & Langenburg, G. (2013). Improving the understanding and the reliability of the concept of "sufficiency" in friction ridge examination. National Institute of Justice, Washington DC. https://www.ncjrs.gov/pdffiles1/nij/grants/244231.pdf

19 Office of the Inspector General (2006) Review of the FBI's handling of the Brandon Mayfield case (US Department of Justice, Washington, DC). https://oig.justice.gov/special/s0601/final.pdf

20 Budowle B, Buscaglia J, Perlman RS (2006) Review of the scientific basis for friction ridge comparisons as a means of identification: committee findings and recommendations. Forensic Sci Commun 8:1. https://www2.fbi.gov/hq/lab/fsc/backissu/jan2006/research/2006_01_research02.htm

21 Huber R.A. (1959) Expert witness. Criminal Law Quarterly 2:276-296.

22 Ashbaugh D. (1999) Quantitative-qualitative friction ridge analysis: an introduction to basic and advanced ridgeology (CRC Press, New York).

23 Pacheco I, Cerchiai B, Stoiloff S (2014). Miami-Dade Research Study for the Reliability of the ACE-V Process: Accuracy & Precision in Latent Fingerprint Examinations. https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=270637

24 Langenburg, G. (2012). A critical analysis and study of the ACE-V process (unpublished doctoral dissertation). Université de Lausanne, Lausanne. http://www.unil.ch/files/live//sites/esc/files/shared/Langenburg_Thesis_Critical_Analysis_of_ACE-V_2012.pdf

25 Stone CJ, Koo CY (1985). Additive splines in statistics. *Proc Stat Comp Sect Am Statist Assoc*, *27*, 45-48.

Final accepted manuscript, subsequently published as
Ulery, B. T., Hicklin, R. A., Roberts, M. A., & Buscaglia, J. (2017). Factors associated with latent fingerprint exclusion determinations. *Forensic Science International, 275*, 65–75.
https://doi.org/10.1016/j.forsciint.2017.02.011

Supporting Information Appendices

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.
AFIS	Automated Fingerprint Identification System (generic term)
Analysis phase	The first phase of the ACE-V method. In this test, the examiner annotated the latent and made a value determination before seeing the exemplar print.
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. For brevity, in this report we use "Comparison" to refer to the Comparison/Evaluation phase.
Comparison determination	The determination of individualization, exclusion, or inconclusive reached in the Comparison phase of the test. SWGFAST [1] refers to this determination as the Evaluation Conclusion.
Corresponding minutia	Explicit annotation by an examiner associating a marked minutia in the latent with a marked minutia in the exemplar, as defined in ANSI/NIST-ITL [2]. Examiners were instructed to mark all such correspondences that they used to make their Comparison determinations.
Determination	See value determination, Comparison determination.
Difficulty	In this test, examiners assessed comparisons on a 5-level scale (Very Easy/Obvious, Easy, Moderate, Difficult, Very Difficult).
Discrepancy	An examiner's assessment that a dissimilarity between two friction ridge impressions originates in the skin itself and cannot be explained as an artifact or distortion. WB participants were instructed to mark discrepancies as needed to support an exclusion determination.

Dissimilarity	A difference in appearance between two friction ridge impressions. For example, a dissimilarity may arise as an artifact of distortion in the print or scarring in the skin. Some dissimilarities may be determined to be discrepancies by an examiner.
Exclusion	The comparison determination that the latent and exemplar fingerprints did not come from the same finger. For our purposes, this is *exclusion of source*, which means the two impressions originated from different sources of friction ridge skin, but the subject cannot be excluded, whereas *exclusion of subject* means the two impressions originated from different subjects.
Exemplar	A fingerprint from a known source, intentionally recorded.
FN	False negative: an (erroneous) exclusion of a mated image pair by an examiner.
FNR	False negative rate: percentage of determinations on mated image pairs that were (erroneous) exclusions.
FP	False positive: an (erroneous) individualization of a nonmated image pair by an examiner
FPR	False positive rate: percentage of determinations on nonmated image pairs that were (erroneous) individualizations.
Inconclusive	The comparison determination that neither individualization nor exclusion is possible.
Individualization	The comparison determination that the latent and exemplar fingerprints originated from the same source. In the United States, 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."[1,3]
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 [4]. 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	FBI's Latent Quality Metric (LQMetric) software automatically assesses the quality of latent fingerprint images. LQMetric is included in the FBI's Universal Latent Workstation (ULW) software [5], release 6.5 and later.
Mated	A pair of images (latent and exemplar) known *a priori* to derive from impressions of the same source (subject and finger). Compare with "individualization," which is an examiner's ***determination*** that the prints are from the same source.
Marked minutia	An annotation by an examiner on the print indicating the presence of a minutia at that location.
Minutia	An event along the path of a single friction ridge, either a bifurcation or ridge ending. Examiners were instructed to mark features such as scars, dots, incipient ridges, creases and linear discontinuities, ridge edge features, or pores as "other" features, not as minutiae. In this study, examiners did not differentiate between bifurcations and ending ridges.
Missed ID	Failure by an examiner to individualize a mated pair that was individualized by any (or most) other examiners (also known as a "missed individualization" or "missed identification").
NGI	The FBI's Next Generation Identification AFIS.
Nonmated	A pair of images (latent and exemplar) known *a priori* to derive from impressions of different sources (different fingers or different subjects).
NPV	Negative predictive value: the percentage of exclusion determinations that are true negatives (i.e., made on nonmated image pairs).
NV	No value: An examiner's determination that the latent image is not of value for individualization or exclusion. See also VEO and VID.
Repeatability	Intraexaminer agreement: when one examiner provides the same determination in response to an image or image pair, on multiple occasions.
Reproducibility	Interexaminer agreement: when multiple examiners provide the same determination in response to an image or image pair.
Source	An area of friction ridge skin used to create an impression. Two impressions are said to be from the "same source" when they have in common a region of overlapping friction ridge skin.

TN	True negative: the exclusion of a nonmated image pair by an examiner.
TNR	True negative rate: percentage of determinations on nonmated image pairs that were exclusions.
TP	True positive: the individualization of a mated image pair by an examiner.
TPR	True positive rate: the percentage of determinations on mated image pairs that were individualizations.
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 6].
VCMP	Of value for comparison (VEO or VID)
VEO	Value for exclusion only: An examiner's determination that the latent is not of value for individualization and contains some friction ridge information that may be appropriate for exclusion if an appropriate exemplar is available. See also NV and VID.
VID	Value for individualization: An examiner's determination that the latent is of value and is appropriate for potential individualization if an appropriate exemplar is available. See also VEO and NV.

Appendix SI-2 Materials and methods

Detailed descriptions of the experimental designs of both BB and WB have been published previously [6,7,8]. This section summarizes aspects of those designs specifically important to this paper.

Appendix SI-2.1 Test procedures

Table S1 summarizes some of the key differences between BB and WB.

	Black Box	White Box
Primary objective	Accuracy and reliability of examiner determinations (of all types)	Associations between markup and determinations, especially sufficiency for individualization (the threshold between individualization and inconclusive)
Fingerprints	Laboratory-collected prints, intended to be representative of difficult casework; mating known with certainty. 356 latents; 520 mated pairs; 224 nonmated pairs	Laboratory and casework prints, selected to vary broadly over a four-dimensional design space: number of corresponding minutiae, clarity, presence of cores and deltas, and complexity. 301 latents; 231 mated pairs; 89 nonmated pairs
Participants	169 practicing latent print examiners, 72 of whom participated in the follow-on BB Repeatability study; 1% international participants	170 practicing latent print examiners; 18% international participants
Comparisons	Each examiner was assigned 100 image pairs (mean 68% mated). 25 of these were reassigned in the follow-on BB Repeatability study.	Each examiner was assigned 22 image pairs (17 mated, 5 nonmated)
Determinations	Latent value, comparison determination	Latent value, exemplar value, comparison determination
Markup	None	Clarity, minutiae, cores, deltas, corresponding features
Ancillary questions	Comparison difficulty (5 levels), inconclusive reason (3 options), exclusion reason (2 options)	Comparison difficulty (5 levels), exclusion reason (5 options)

Table S1: Summary comparison of the two tests.

Appendix SI-2.2 Fingerprints

The fingerprints for these studies were collected at the FBI Laboratory and at Noblis under controlled conditions, and (White Box study only) from operational casework datasets collected by the FBI. All prints were impressions of distal segments of fingers, including some sides and tips. We sought diversity in fingerprint data, within a range typical of casework. In both studies nonmated pairs were based on difficult comparisons

resulting from searches of IAFIS[e] or selected for a comparable level of difficulty. In BB, mated pairs were randomly selected from the multiple latents and exemplars available for each finger position; in WB both mated and nonmated pairs were selected selected to vary broadly over a four-dimensional design space.

In support of the distinct study objectives, BB data selection emphasized prints representative of casework whose mating was known with certainty in order to study the accuracy and reliability of examiners' determinations; WB data selection emphasized a broad variety of quality characteristics in order to establish what constitutes sufficiency for individualization.

BB fingerprints included 356 latents from 165 distinct fingers (from 21 people), and 484 exemplars. These were combined to form 744 distinct latent-exemplar image pairs (520 mated, 224 nonmated). WB fingerprints included 301 latents from 247 distinct fingers (from 166 people), and 319 exemplars. These were combined to form 320 distinct latent-exemplar image pairs (231 mated, 89 nonmated).

The fingerprints and image pairs in BB and WB may or may not be representative of casework for any particular agency. In surveys of participants, a large majority of BB and WB respondents agreed that the fingerprints were representative of (or similar to) casework, and that the overall difficulty of comparisons was similar to casework [6,8].

[e] *IAFIS was the FBI's Integrated Automated Fingerprint Identification System. In 2013, IAFIS latent print services were replaced by the FBI's Next Generation Identification (NGI) system.*

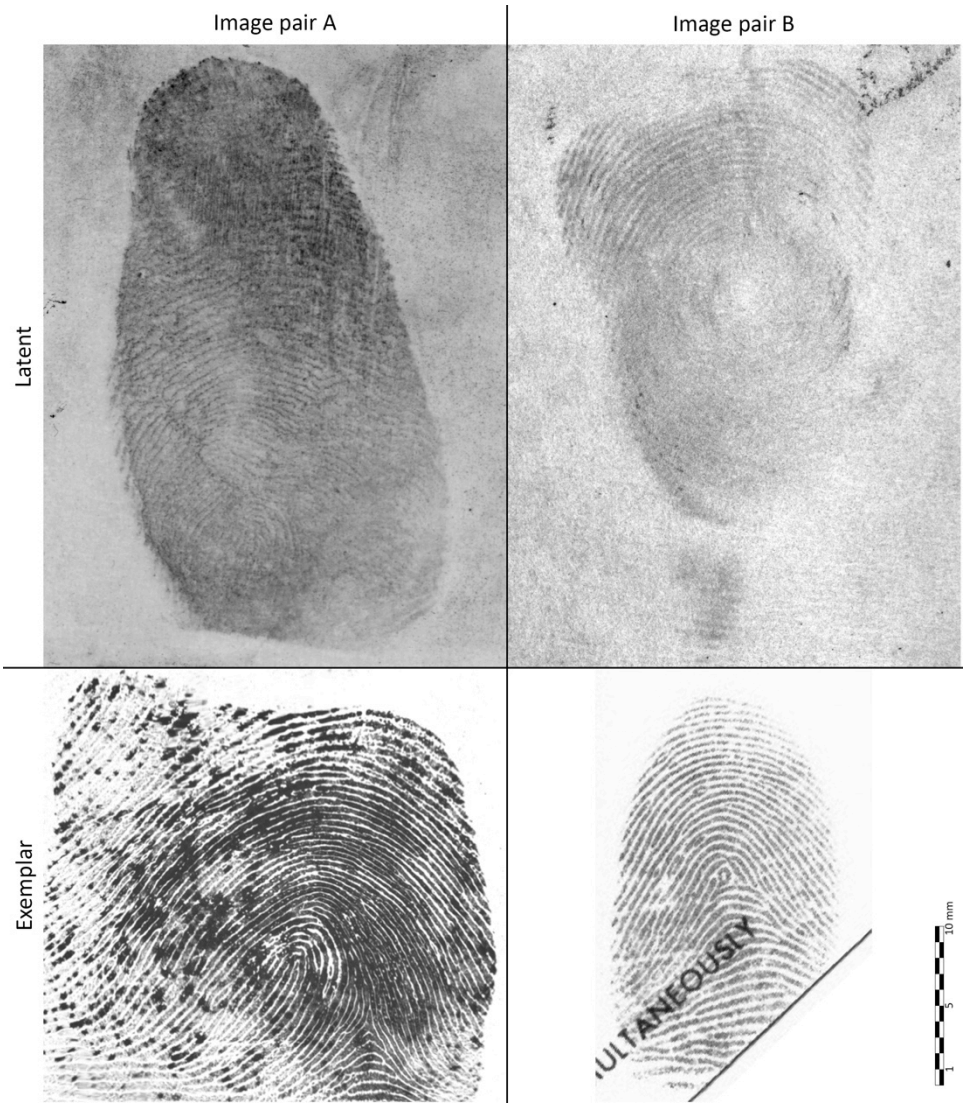

Fig. S1. Two examples of mated image pairs that resulted in erroneous exclusions. Determinations by WB examiners on image pair A: 2 exclusions, 2 NV, 8 inconclusives, 2 individualizations; up to 12 corresponding minutiae were marked; one examiner who erroneously excluded marked 10 corresponding minutiae. Determinations by BB examiners on image pair B: 5 exclusions, 4 NV. All images are shown at the same resolution.

Appendix SI-2.3 Participants

Participation was open to practicing latent print examiners from across the fingerprint community. Most of the participants were volunteers, but some were required or requested to participate by their employers. Participants were diverse with respect to organization, training history, and other factors.

- In BB, a total of 169 latent print examiners participated; most were volunteers, while the others were encouraged or required to participate by their employers. The latent print examiners were generally highly experienced: median experience was 10 years, and 83% were certified as latent print examiners. More detailed descriptions of participants are included in [6].
- In WB, a total of 170 latent print examiners participated: 90% were certified (or qualified by their employers) as latent print examiners; 82% were from the U.S. More detailed descriptions of participants are included in [8].

Appendix SI-2.4 Survey responses related to exclusions

Participants in each study were asked to complete a survey (included in full in [6] and [8]). Table S2 summarizes responses to questions asked in the Black Box study that were related to exclusions; one of these questions was also asked of White Box participants.

25. Are you aware of ever having made an erroneous exclusion (after training)? (Check all that apply - may add to over 100%)				
No response	2	1%		
No	103	65%		
Yes, on casework; detected after it was reported to contributor	10	6%		
Yes, on a proficiency test only	4	3%		
Yes, on casework; detected during verification	43	27%		
On question 25, one examiner indicated yes both on a proficiency test and on casework detected during verification. Two examiners indicated yes both on casework detected after reporting and on casework detected during verification.				
28. If the latent and exemplar are both of value, include a large potentially corresponding area, no other latent or exemplars images are available, and you already have all processing information related to the latent, are you permitted to make an inconclusive determination? (Given the standard operating procedures that you/your agency currently use)				
Inconclusive determinations are discouraged but possible in this case	31	19%		
Inconclusive determinations are freely accepted in this case	77	48%		
Inconclusive determinations are not permitted in this case	51	32%		
29. In determining the value/sufficiency of a latent impression, how do you define an impression that is not suitable for individualization but could potentially be used for exclusion? (Given the standard operating procedures that you/your agency currently use)			***White box***	
It has its own category used in standard practice, such as "Of value for exclusion only" or "Limited value"	27	17%	33	20%
It has its own category, such as "Of value for exclusion only" or "Limited value" – but only used upon request	21	13%	42	25%
No value	88	55%	81	48%
Of value	23	14%	13	8%
30. How often in casework do you make a conclusion that a latent and the exemplars provided definitively did not come from the same source? (Given the standard operating procedures that you/your agency currently use)				
Never	5	3%		
Used only on request	4	3%		
Rarely	16	10%		
Often	134	84%		
31. How do you use the term "exclusion" as a conclusion? (Given the standard operating procedures that you/your agency currently use)				
Any comparison that is not an individualization is an exclusion	7	4%		
Exclusion means that the latent did not come from any finger for that subject, but could have come from other friction ridge skin (e.g. palm) from that subject	16	10%		
Exclusion means that the latent did not come from any friction ridge skin for that subject	81	51%		
Exclusion means that the latent did not come from the source of the exemplar (e.g. a specific finger), but could have come from another finger from that subject	18	11%		
Not used	37	23%		

Table S2: Survey responses relevant to exclusions. BB survey responses were provided by 159 of the 169 examiners. (WB, 169 of 170).

The responses to questions #28 and #29 were correlated: Fig. S2 shows that inconclusive comparison determinations are more often discouraged or not permitted when VEO latents are not routinely compared.

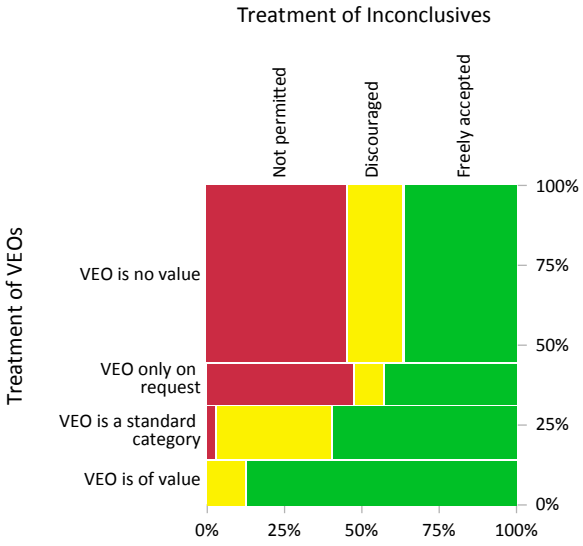

Fig. S2: Associations between examiners' survey responses regarding treatment of VEOs (latents of value for exclusion only, question #29) and inconclusives (question #28). (BB, n=159 survey responses).

Appendix SI-3 Summary of BB and WB test sizes and determination rates

In order to more fully understand the data on exclusions from the BB and WB studies, it is important to account for some of the notable similarities and differences in the determination rates measured in the two studies. Some of these differences can be attributed to differences in test procedures and data selection (discussed in Appendix SI-2). This section summarizes data from the BB and WB studies. In general, the BB data shown here has been previously published, but the WB data has not; the BB data is included in order to assist in comparing the studies. After briefly summarizing overall determination rates (Table S3 and Fig. S3), we present new data selected to account for important similarities and differences (Appendix SI-3.1 and Appendix SI-3.2). This data is valuable in demonstrating what factors must be considered when interpreting results from similar studies and designing future experiments.

The Black Box study yielded 17,121 valid analysis-phase responses from 169 examiners [6]. Each examiner was initially assigned 100 image pairs from a pool of 744 total pairs; 72 of these examiners participated in a follow-on repeatability study, which included some additional (not repeated) presentations [7]. Examiners made 3947 latent NV determinations, yielding a total of 13,174 comparisons.

The White Box study yielded 3730 valid analysis-phase responses from 170 examiners [8]. Each examiner was assigned 22 image pairs from a pool of 320 total pairs. Comparison-phase results are based on 2966 comparisons where neither the latent nor the exemplar was assessed to be NV; this count omits 762 NV determinations (713 analysis-phase latent NV, 43 Comparison-phase latent NV, and 6 Comparison-phase exemplar NV) and 2 invalid determinations (software issue).

Table S3 and Fig. S3 summarize the determination rates on BB and WB. Some of the striking differences in these distributions reflect differences in test procedures (discussed in Appendix SI-2.1) and data selection (discussed in Appendix SI-3.1). Additionally, BB results were published prior to WB, and in particular the high FNR was widely discussed, and therefore WB participants may have changed their behavior in response. BB and WB differed in participants, with a larger proportion of international participants in WB. The requirement in WB to mark features may also have had an effect on determination rates.

		Total	Mates			Nonmates		
		count	count	% PRES	% CMP	count	% PRES	% CMP
BB	NV (not compared)	3,947	3,389	29.3%	n/a	558	10.1%	n/a
	Exclusion	4,558	611	5.3%	7.5%	3,947	71.2%	79.2%
	Inconclusive	4,907	3,875	33.5%	47.3%	1,032	18.6%	20.7%
	Individualization	3,709	3,703	32.0%	45.2%	6	0.1%	0.1%
	Total comparisons	*13,174*	*8,189*	*70.7%*		*4,985*	*89.9%*	
	Total presentations	*17,121*	*11,578*			*5,543*		
WB	NV (not compared)	713	462	16.0%	n/a	251	29.6%	n/a
	NV (in comparison)	49	35	1.2%	n/a	14	1.7%	n/a
	Invalid data (No determination)	2	1			1		
	Exclusion	561	131	4.5%	5.5%	430	50.7%	73.9%
	Inconclusive	705	554	19.2%	23.2%	151	17.8%	25.9%
	Individualization	1,700	1,699	59.0%	71.3%	1	0.1%	0.2%
	Total comparisons	*2,966*	*2,384*	*82.7%*		*582*	*68.6%*	
	Total presentations	*3,730*	*2,882*			*848*		

Table S3: Summary of sample sizes and determination rates in BB and WB [6,8]. False negative rates (FNR_{PRES} and FNR_{CMP}) are highlighted in yellow; true negative rates (TNR_{PRES} and TNR_{CMP}) in blue. % PRES (% CMP) describes how the determination types were distributed over all presentations (comparisons) of either mates or nonmates.

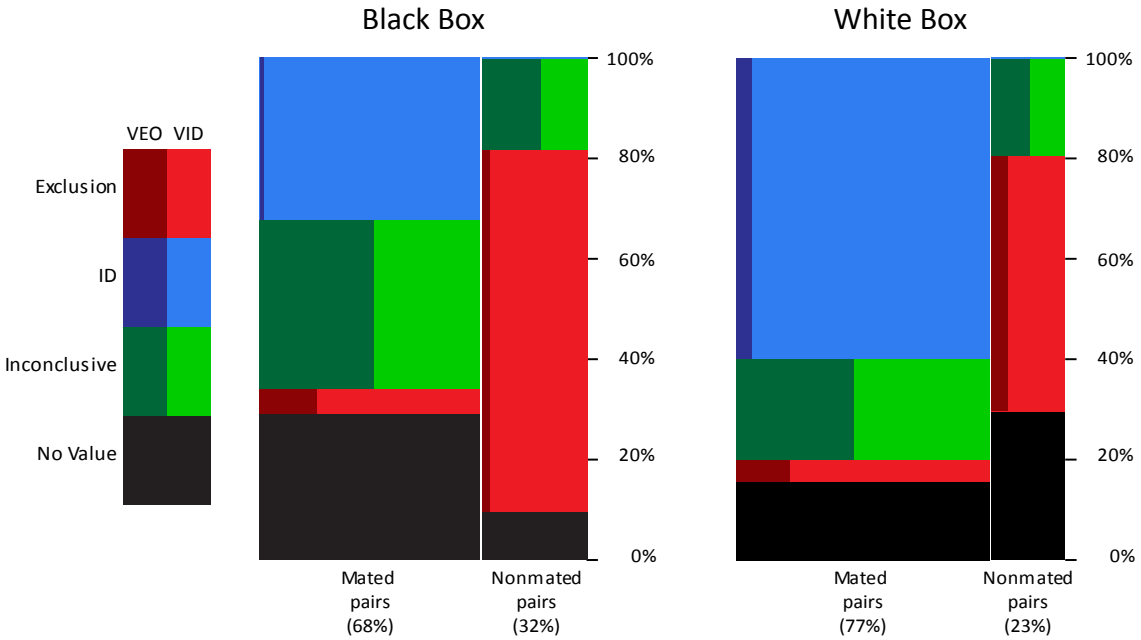

Fig. S3: Distributions of determinations in BB (n=17,121 determinations) and WB (n=3730 determinations). BB data was previously published [6], included here for ease of comparison.

Appendix SI-3.1 Effects of data selection

Data selection differed between the two tests, and also differed for mated and nonmated image pairs within each test. In BB, although the latents for mated and nonmated image pairs were selected from a single pool of available prints, the process of selecting challenging nonmated image pairs omitted many NV latents; this resulted in a much greater proportion of NV and inconclusive determinations for mated than nonmated pairs. In WB, all image pairs were selected to provide coverage of a multi-dimensional design space [8]; we

deliberately limited the proportion of image pairs on which we expected unanimous determinations in order to focus on the boundaries of sufficiency for individualization. As a result of these design decisions, the distributions of latent quality differed among the four subpopulations of latents (Fig. S4 and Fig. S5), as did the proportions of image pairs on which examiners could reach conclusions (Fig. S6).

The differences shown in Fig. S4-Fig. S6 and Table S4 are important when interpreting differences in determination rates between the two tests. The proportions of mated and nonmated pairs also varied within each test as a function of latent quality (Fig. S4); this is important when interpreting differences in NPV associated with factors such as LQMetric or minutia counts.

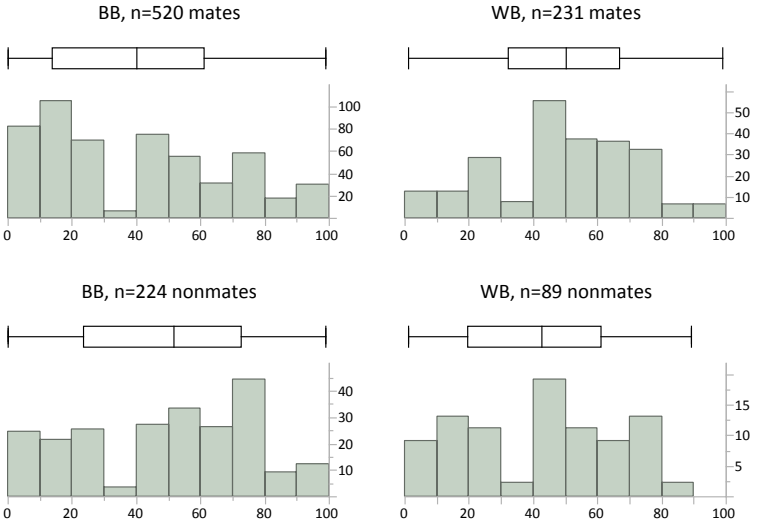

Fig. S4: LQMetric on latents selected for mated and nonmated image pairs in BB and WB. The LQMetric algorithm tends not to produce estimates in the range 30-40; this does not reflect a gap in the actual quality distribution of latents selected for BB and WB. See Appendix SI-7 for associations of LQMetric with value assessments and determinations.

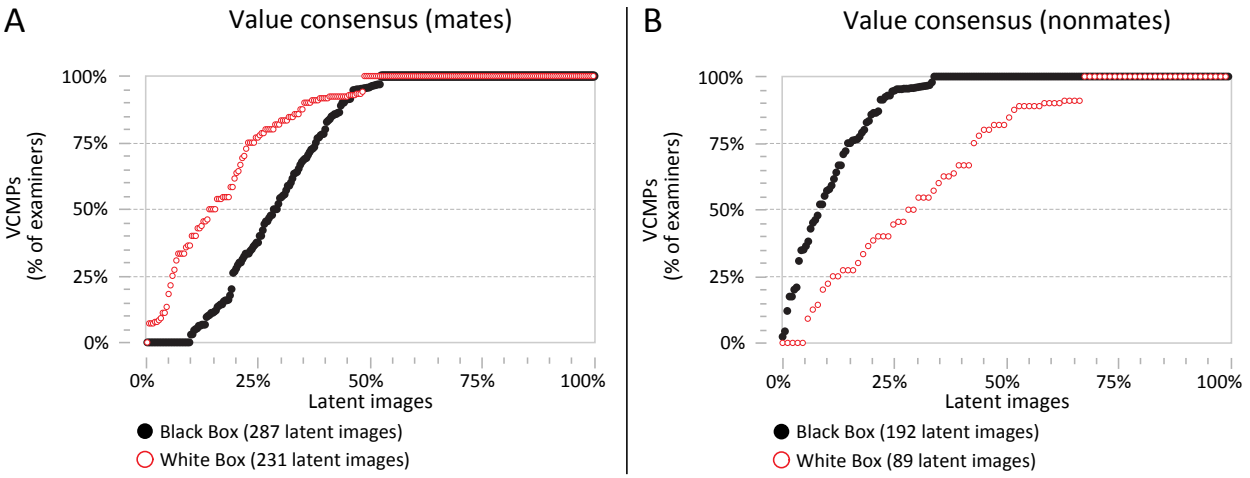

Fig. S5: Consensus on value for comparison on the latents selected for (A) mated and (B) nonmated image pairs in BB and WB.

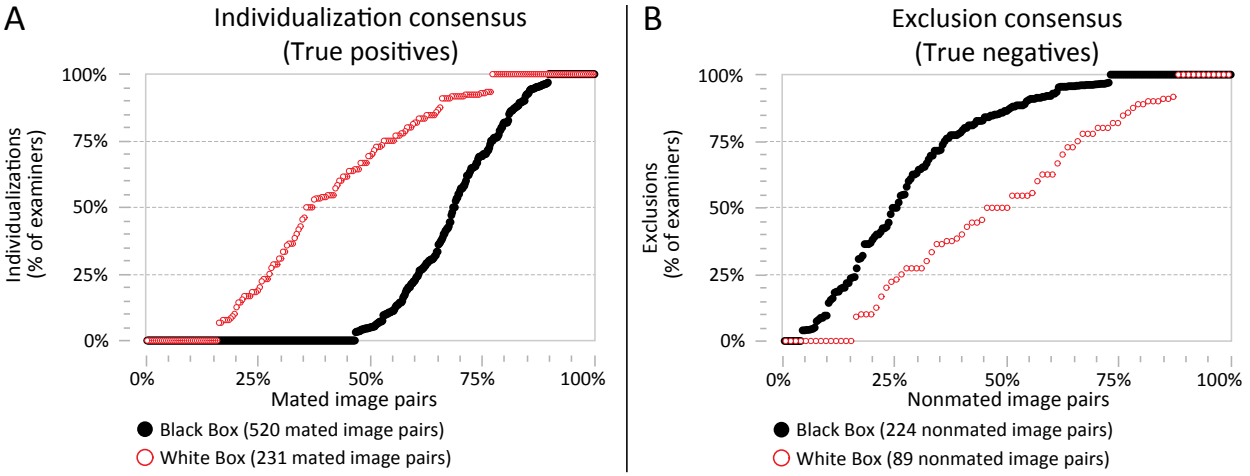

Fig. S6: Examiner consensus on (A) individualization determinations on mated pairs; (B) exclusion determinations on nonmated pairs. BB: n=11,578 determinations on 520 mates, 5543 determinations on 224 nonmates; WB: n=2882 determinations on 231 mates, 848 determinations on 89 nonmates. Examiner consensus on false negatives is shown in Fig. S8A.

		Image pairs	Unanimous NV or inconclusive	Unanimous exclusion	Unanimous ID
Mates	BB	520	38%	0%	10%
	WB	231	7%	0%	23%
Non-mates	BB	224	9%	25%	0%
	WB	89	16%	12%	0%

Table S4: Proportions of image pairs with unanimous determinations in BB and WB.

There were 83 image pairs (33 mates, 50 nonmates) that were presented on both BB and WB. This common subset provides one means of comparing determination rates across the tests while controlling for differences in data selection. Table S5 summarizes determinations on those 83 image pairs. Although some rates differed notably between the two tests as shown in Table S3 (e.g., individualization and inconclusive rates on mated pairs; exclusion rates on nonmated pairs), these large differences are not present on this common subset, indicating that the differences in rates were indeed due to data selection.

		Total	**Mates**			**Nonmates**			**% Mates**
		count	**count**	**% total**	**% comp**	**count**	**% total**	**% comp**	
BB	NV (not compared)	329	36	5%	n/a	293	23%	n/a	11%
	Exclusion	691	45	6%	6%	646	52%	67%	7%
	Inconclusive	519	206	26%	28%	313	25%	33%	40%
	Individualization	496	496	63%	66%	0	0%	0%	100%
	Total comparisons	*1706*	*747*	*95%*		*959*	*7%*		*44%*
	Total	*2035*	*783*			*1252*			*38%*
WB	NV (not compared)	197	38	9%	n/a	157	33%	n/a	19%
	NV (in comparison)	8	2	0%	n/a	6	1%	n/a	25%
	Exclusion	235	10	2%	3%	225	47%	71%	4%
	Inconclusive	178	88	21%	23%	90	19%	28%	49%
	Individualization	280	279	67%	74%	1	0%	0%	100%
	Total comparisons (VCMP)	*693*	*377*	*90%*		*316*	*66%*		*54%*
	Invalid data (No determination)	1	0			1			
	Total	*899*	*417*			*481*			*46%*

Table S5: Sample sizes and determination rates on 83 image pairs that were common to both tests (33 mated, 50 nonmated image pairs). False negative rates (FNR_{CMP}) are highlighted in yellow; true negative rates (TNR_{CMP}) in blue.

On this common subset of image pairs, the difference in FNR (6% vs. 3%) is significant at alpha=0.05 based on a chi-squared test, but not the difference in TNR (67% vs. 71%); however, the observations were not

independent (violating a chi-square test assumption) and the 83 image pairs were not assigned equally often on the two tests. The differences in how often each image pair was assigned appears to account for much of the measured difference in TNR and some of the difference in FNR; after controlling for how often each image pair was assigned (by modeling exclusions as a response to test ID and image pair ID using logistic regression), FNR remains significantly higher on BB than WB (chi-square test, alpha=0.05).

As an alternative method of comparing test results on this common subset, we can tally overall results by image pair. FNR_{PRES} was higher on BB than on WB on 18 of the 22 mated pairs that were excluded by at least one examiner. TNR_{PRES} was higher on BB than on WB on 25 of the 45 nonmated pairs that were excluded by at least one examiner (17 were higher on WB, 3 were equal).

In BB the true negative rate was much greater than the true positive rate (TNR_{CMP}=79.2% >> TPR_{CMP}=45.2%; TNR_{PRES}=71.2% >> TPR_{PRES}=32.0%), but this was not true in WB (TNR_{CMP}=72.0% ≈ TPR_{CMP}=70.2%; TNR_{PRES}=50.7% < TPR_{PRES}=59.0%). The relative differences observed between TNR and TPR can be attributed to data selection: if the mate and nonmate datasets are selected in different ways, we should not have expectations regarding the relative differences between TNR and TPR.

Appendix SI-3.2 Reproducibility of determinations

Table S6 summarizes the reproducibility of each type of determination in BB and WB, conditioned on the type of determination made. As we have just discussed, data selection has a strong effect on reproducibility and therefore rates can be expected to differ between tests. Nevertheless, our measures provide a rough understanding of reproducibility for the types of data included in our tests.

				Examiner B			
			#	NV	Inconc	Excl	ID
BB Mates	Examiner A	NV	3389	76.9%	18.9%	2.7%	1.5%
		Excl	611	15.0%	43.9%	15.2%	26.0%
		Inconc	3875	16.6%	61.8%	6.9%	14.7%
		ID	3703	1.3%	15.4%	4.3%	79.0%
BB Nonmates	Examiner A	NV	558	54.0%	29.2%	16.7%	0.0%
		Excl	3947	2.4%	11.0%	86.5%	0.1%
		Inconc	1032	15.8%	42.0%	42.1%	0.1%
		ID	6	0.0%	17.8%	82.2%	0.0%
WB Mates	Examiner A	NV	498	54.7%	23.9%	5.3%	16.1%
		Excl	131	20.2%	27.7%	11.2%	40.8%
		Inconc	554	21.5%	39.9%	6.6%	32.0%
		ID	1699	4.7%	10.4%	3.1%	81.7%
WB Nonmates[f]	Examiner A	NV	265	60.9%	16.8%	22.3%	0.0%
		Excl	430	13.8%	15.5%	70.6%	0.2%
		Inconc	151	29.4%	26.3%	44.1%	0.1%
		ID	1	0.0%	20.0%	80.0%	0.0%

Table S6: Reproducibility of determinations, showing the probability of an independent determination by a second examiner conditioned on the decision of a first examiner. Reproducibility rates of exclusions are highlighted (yellow: false negatives; blue: true negatives). Percentages sum to 100% on each row and were calculated by considering all pairwise combinations of responses and weighting each examiner A determination equally.[g]

[f] *One WB nonmate omitted due to missing comparison determination.*

[g] *Table 7 in [7] reported false negative reproducibility of 19.2%. However, that was limited to a subset of participants (those who had participated in the retest).*

Appendix SI-4 False negatives vs. missed IDs

When an examiner fails to individualize a mated pair that can be individualized by another examiner (or, alternatively, by a majority of examiners), it is considered in some agencies a "missed ID." We have found that missed IDs and erroneous exclusions are often confused, and therefore included this section to clarify the distinction. Here we define a missed ID as an exclusion, inconclusive, or NV determination on a mated pair that the majority of examiners individualized: in BB, 4.7% (WB, 9.4%) of responses on mated pairs were missed IDs, as shown in Fig. S7.

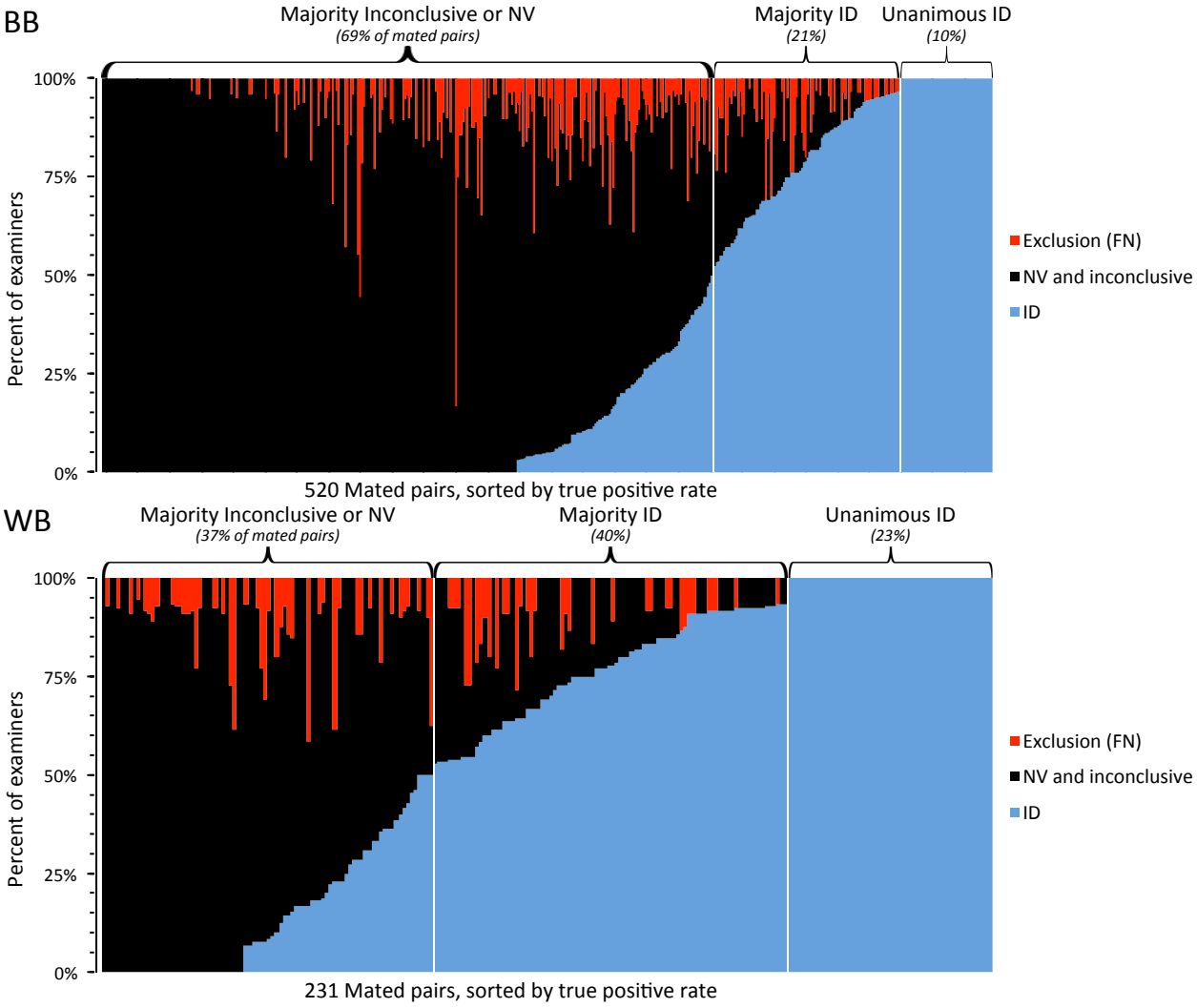

Fig. S7: Examiner determinations on mated pairs, illustrating false negatives (red) and missed IDs (black and red in the majority ID area). Charts are equivalent to the individualization consensus curves in Fig. S6 (left chart), but further differentiate between exclusions and other non-individualization determinations. (BB: n=11,578 determinations on mated pairs, 544 of which were missed IDs; WB: n=2882, 270 of which were missed IDs).

As shown in Fig. S7 and Table S7A, on mated image pairs that the majority of examiners did not individualize, erroneous exclusions accounted for a small minority of non-individualizations (BB 6%, WB 8%). On those image pairs that were individualized by the majority of examiners, erroneous exclusions accounted for a greater proportion of non-individualizations: in BB 27% (WB 20%) of missed IDs were erroneous exclusions. However, the proportion of erroneous exclusions that were missed IDs differed significantly between the tests:

24% in BB, but 42% in WB (Table S7B), because a greater proportion of image pairs in WB resulted in majority IDs (Fig. S7).

Non-ID determinations on mated pairs

A	BB			WB		
	Count	Excl (FN)	Inconc/NV	Count	Excl (FN)	Inconc
Majority ID (missed IDs)	544	27%	73%	270	20%	80%
Minority ID	7331	6%	94%	913	8%	92%
Total	7875	8%	92%	1183	11%	89%

B	BB			WB		
	Count	Majority ID (missed IDs)	Minority ID	Count	Majority ID (missed IDs)	Minority ID
Excl (FN)	611	24%	76%	131	42%	58%
Inconc/NV	7264	5%	95%	1052	20%	80%
Total	7875	7%	93%	1183	23%	77%

Table S7: Associations between missed IDs and erroneous exclusions, among non-individualization determinations on mated pairs. A) highlighted cells are the proportions of missed IDs that are false negatives; B) highlighted cells are the proportions of false negatives that are missed IDs. (BB, n=7875 non-individualization determinations on mated pairs; WB, n=1183).

Appendix SI-5 Image effects on erroneous exclusions

The factors we discuss in the Results only partially explain the false negatives we observed. To a first approximation, modeling erroneous exclusions as random events that are equally likely to occur on any mated comparison provides a good description of much of our data (Fig. S8). The chart on the left shows the actual distribution of FNR_{PRES} across image pairs on each test. The chart on the right shows the results of simulations in which each comparison was equally likely to result in an erroneous exclusion: under this assumption, we would expect no erroneous exclusions on some image pairs and multiple erroneous exclusions on others as described by a binomial distribution. The similarity of the actual data (left) to the simulated data (right) demonstrates that much of the observed variation in FNR_{PRES} by image pair can be attributed to chance. A small number of image pairs were much more likely to be erroneously excluded than could be explained by chance under the equal probability assumption (rightmost outliers in left chart), and more image pairs were never erroneously excluded than predicted by equal probabilities (left tails). The disproportionate false negative rates for some image pairs can be attributed at least in part to the factors we discuss.

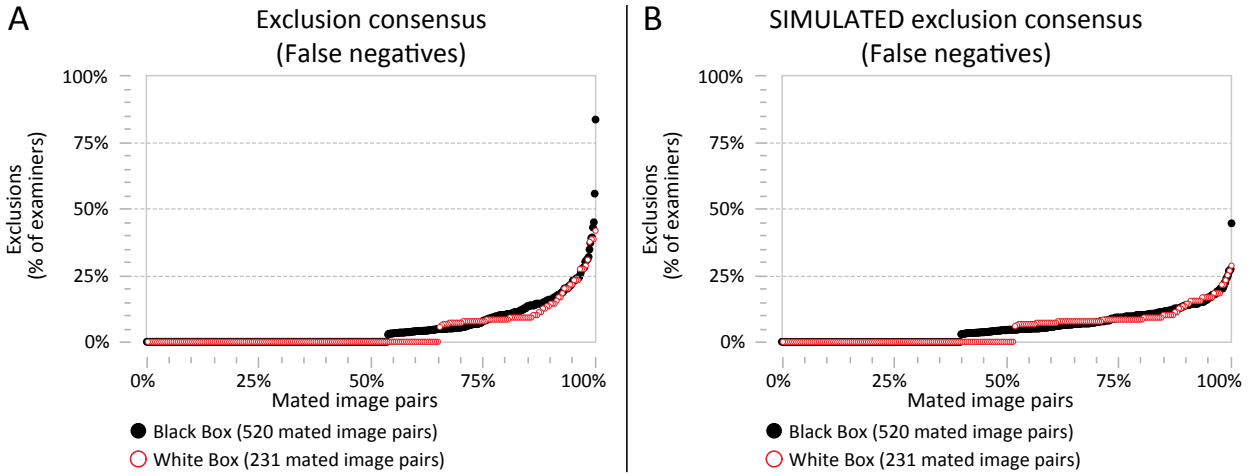

Fig. S8: Variation in erroneous exclusion rates by image pair. (A) Actual FNR_{PRES} for each image pair; (B) simulated FNR_{PRES} for each image pair assuming a constant overall FNR_{CMP} (BB FNR_{CMP} = 7.5%; WB FNR_{CMP} = 5.5%). BB data is shown in black (520 mates, mean of 22 examiners per image pair) and WB in red (231 mates, mean of 12 examiners per image pair).

Fig. S9 shows a more detailed view of this data, taking into account the number of examiners who actually compared each image pair. Each chart in Fig. S9 plots an exclusion rate for each mated BB image pair against the number of examiners who compared that image pair (i.e., did not rate the latent NV). The chart on the left shows the actual exclusion rates; the chart on the right shows simulated rates. The relative overdispersion in FNR_{PRES} in the actual data represents the extent to which erroneous exclusions were more or less likely to occur on some image pairs.

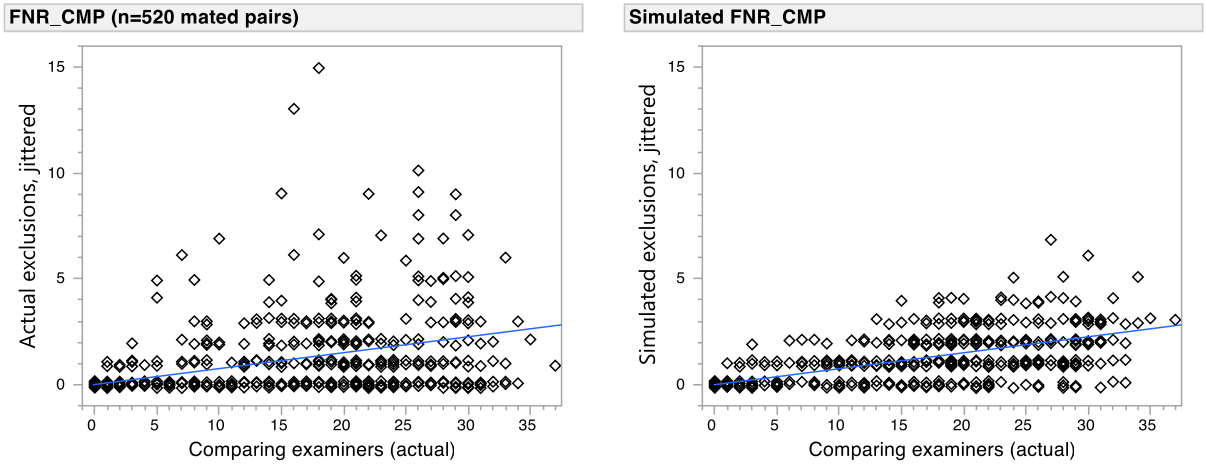

Fig. S9: Variation in erroneous exclusion rates by image pair. (A) Black Box test results; (B) simulated results assuming no image effect beyond which latents are compared. The simulation preserved the exact test structure (assignments of image pairs to examiners) and the actual latent value determinations, but replaced examiner comparison responses by random values with Prob(exclusion) = 7.5%; this mean rate (expected value) is shown by the blue reference lines. Vertical dispersion in the simulation (right) relative to the expected value reflects measurement imprecision due to small sample sizes; the relative increase in vertical dispersion in the actual data (left) reflects real differences in false negative rates from one image pair to another. (BB, n=520 mated image pairs).

Appendix SI-6 Examiner effects on exclusions

Erroneous exclusions were widely distributed among examiners (as they were for image pairs). In BB, 85% of examiners made at least one erroneous exclusion — although 65% of participants said that they were unaware of ever having made an erroneous exclusion after training. In WB, only 44% of examiners made any erroneous exclusions on the test, which is consistent with the fact that each examiner was assigned fewer mated pairs (17 on WB vs. a mean of 69 on BB) and therefore had fewer opportunities to make errors.

Appendix SI-6.1 Variation in FNR by examiner

The sample sizes were small for estimating the FNR of each participant. For example, if an examiner who was assigned 69 mated image pairs (the mean for BB) made 4 erroneous exclusions, then the 95% binomial confidence interval for that examiner's FNR_{PRES} is 1.6% to 14.2%. We can, however, determine from the overall distribution of these individual examiner estimates that some examiners make erroneous exclusions at nearly double the average rate, while many others had FNRs that were substantially lower than the group mean. Fig. S10 compares the actual dispersion in FNR_{CMP} by examiner on Black Box to the amount of dispersion that could be expected if there were no interexaminer differences in FNR_{CMP}. The simulation simply replaces each examiner's actual number of exclusions (chart A, y-axis) by a random value from a binomial distribution (chart B, y-axis) where the probability of a simulated exclusion by each examiner is the overall test mean and the number of trials is the actual number of mated comparisons performed by that examiner. The results for White Box were similar.

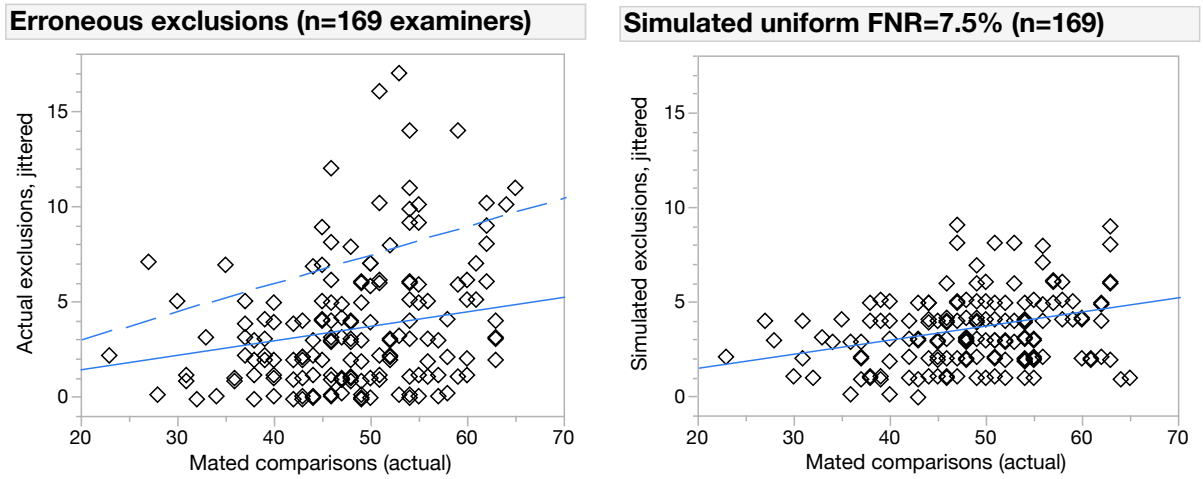

Fig. S10: Examiner effects: variation in erroneous exclusion rates. (A) Black Box test results; (B) simulated results assuming no examiner effect. The simulation models the expected amount of examiner variation using the actual comparison rates (latent value determinations) and assuming a constant FNR_{CMP} = 7.5% (indicated by solid reference line). Relative overdispersion in chart A reflects examiner differences (reference lines indicate FNR_{CMP} = 7.5% and 15%). (BB, n=169 examiners).

As shown by the correlation coefficients in Table S8, differences among examiners in FNR cannot be accounted for simply as a consequence of differences in their overall conclusion rates.

		Correlation	
		BB	WB
FNR_{PRES}	TNR_{PRES}	0.4305	0.0196
FNR_{CMP}	TNR_{CMP}	0.3671	-0.0108
FNR_{PRES}	CR_{PRES} excluding FNs	0.2633	0.2914
TNR_{PRES}	CR_{PRES} excluding TNs	0.7361	0.5615

Table S8: Pearson correlation coefficients for examiner rates (BB, n=169; WB n=170). CR_{PRES} represents the conclusion rate: the percentage of all assigned image pairs (mated and nonmated) on which an examiner determined either exclusion or individualization.

Appendix SI-6.2 Certification and experience

Thompson, et al. [9] found that experts were much better than novices at excluding highly similar prints. Langenburg [10] reported that on nonmated pairs "analysts with over ten years of experience were more likely to report 'exclusion' decisions. Less experienced analysts were more likely to report inconclusive decisions." Among practicing examiners, we also observed higher true negative rates associated with more years of experience. However, we are not observing changes in individual examiners over time, but are reporting associations with factors that are highly confounded in our sample of participants. For example, most of the examiners who lacked certification had fewer than eight years of experience; most examiners with more than 15 years of experience were IAI-latent certified (Fig. S11). At alpha = 0.05, the association between years of experience and TNR was statistically significant (p=0.0002 for TNR_{PRES}; p=0.0053 for TNR_{CMP})

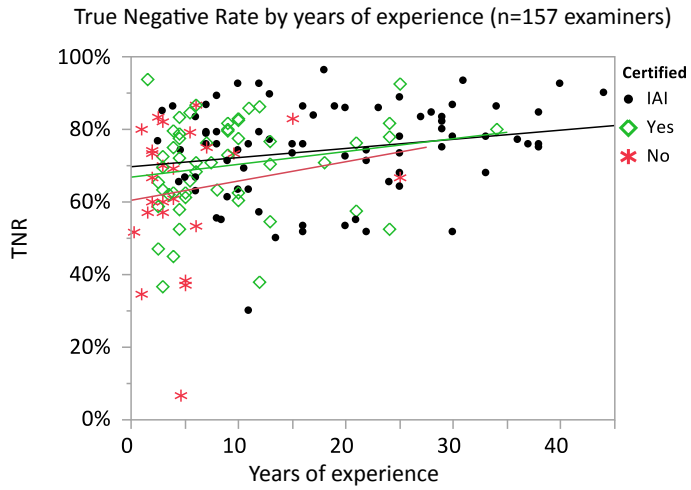

Fig. S11: True negative rates (TNR_{PRES}) by years of experience (BB, n=157 examiners). Data is limited to examiners who reported years of experience (5239 responses on 224 nonmated image pairs). Separate fits are shown for those examiners with IAI latent certification (n=78, black), other certification (n=53, green), and no certification (n=27, red); however, the association between years and TNR was statistically significant only when a single fit was performed on all data.

Langenburg [10] reported: "false negative error rates are higher in the least experienced group of experts (2 years of experience or less) and are reduced in the most experienced group. Simultaneously, the specificity is increasing. Thus we can see that experts are becoming more efficient at excluding with more experience (i.e. they are attempting more 'exclusion' decisions while simultaneously making fewer erroneous 'exclusions')." We found a weak association between years of experience and FNR. Contrary to Langenburg's findings, linear regression showed an increase in FNR with additional years of experience (Fig. S12). At alpha = 0.05, the association was statistically significant for FNR_{PRES} (p=0.013) and not statistically significant for FNR_{CMP} (p=0.06).

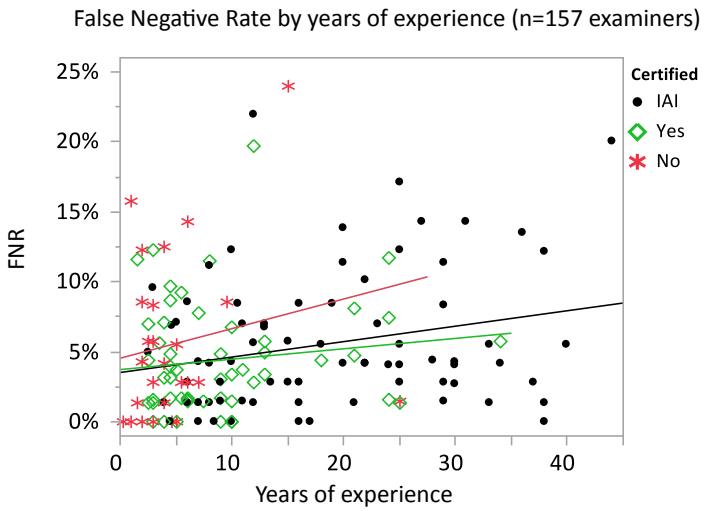

Fig. S12: False negative rates (FNR_{PRES}) by years of experience (BB, n=157 examiners). Data is limited to examiners who reported years of experience (10,767 responses on 520 mated image pairs). Separate fits are shown for those examiners with IAI latent certification (n=78, black), other certification (n=53, green), and no certification (n=27, red). The fits (slope parameters) for other certification and no certification are not statistically significant.

It should not be surprising that measures such as years of experience and certification correlate only weakly with performance measures, especially in a population comprised entirely of practicing latent print examiners [11]. Similar weak associations have been reported elsewhere [6,10].

Appendix SI-7 Reasons for exclusions

Examiners were asked to indicate what observed differences led to each exclusion determination by selecting one of the options listed in Table S9. Examiners were also given the opportunity to provide a short text response to elaborate on the exclusion reason (49 responses) and were specifically requested to comment when the reason was "other" (8 of the 49 responses). On review, 10 of the reasons appear to justify an inconclusive determination — not an exclusion determination. Examples:

- "not enough points for id, similar ridge flow"
- "unable to orientate image appropriately"
- "...or I haven't found my anchor points in the exemplar..."
- "Unable to locate any target groups in common between latent and known"

These responses suggest that some examiners may be confusing exclusion and "non-identification" determinations.

Test	Option	Description
BB	Pattern class/ridge flow alone	The exclusion could be made based on pattern class/ridge flow/level-1 information alone. The exclusion did not require review of minutiae and/or Level-3 information.
	Minutiae and/or level 3	The exclusion determination required comparison of Level-2 and/or Level-3 information.
WB	Pattern classes differ	The exclusion could be made based on pattern class alone.
	Core or delta differences	The exclusion could be made based on one or more of the following: differing ridge flow in the cores or deltas; differing core-delta ridge counts; or differing relations among the deltas.
	One or more minutiae differ	The exclusion determination could be made based on a comparison of Level-2 information.
	Level 3 features differ	The exclusion determination required comparison of Level-3 information.
	Other	None of the above categories satisfactorily explains the basis for the exclusion. Please briefly indicate the basis for the exclusion.

Table S9: Instructions for exclusion reasons. Black Box examiners selected one of two options; White Box examiners were instructed to select the first option that applied.

Table S10 (BB) and Table S11 (WB) describe the reproducibility of exclusions and exclusion reasons. Each table includes one row for each possible exclusion reason provided by an examiner ("Examiner A") on a mated or nonmated comparison; each cell value indicates the conditional probability that a second examiner ("Examiner B") would make a given response on the same image pair. Reproducibility of exclusions and exclusion reasons was generally low; when two examiners both excluded the same image pair, the reason given by the first examiner was not highly predictive of the reason given by the second examiner. The predominant reason given for exclusions was that minutiae differed. In the BB repeatability study, when pattern class/ridge flow was given as the initial reason, examiners often gave minutiae and/or level-3 features as the reason on the retest.

BB			Count	Examiner B				
				NV	Inconc	ID	Excl	
							Pattern	Minutiae
Examiner A Exclusions	Mates (FN)	Pattern class	174	31.8%	42.2%	7.7%	11.0%	7.3%
		Minutiae	437	8.3%	44.5%	33.2%	2.9%	11.1%
	Nonmates (TN)	Pattern class	624	5.6%	8.5%	0.1%	46.4%	39.4%
		Minutiae	3323	1.7%	11.5%	0.1%	7.4%	79.2%

Table S10: Reproducibility of exclusions and exclusion reasons in BB. When examiner A excluded, what examiner B did. Percentages are calculated as weighted sums over all other examiners assigned the same image pair, such that each exclusion by examiner A is weighted equally. (BB, n=4558 examiner A exclusions).

WB			**Count**	**Examiner B**							
				NV	**Inconc**	**ID**	**Excl**				
							Pat	**CD**	**Min**	**L3**	**Other**
Examiner A Exclusions	Mates (FN)	Pattern class	12	45%	30%	18%	1%	1%	4%	0%	0%
		Core or delta	8	17%	31%	34%	2%	0%	15%	0%	0%
		Minutiae	104	18%	27%	43%	0%	1%	9%	0%	0%
		Level 3	3	5%	15%	77%	0%	0%	3%	0%	0%
		Other	3	17%	30%	39%	0%	0%	13%	0%	0%
	Nonmates (TN)	Pattern class	37	24%	7%	0%	40%	8%	18%	0%	2%
		Core or delta	42	12%	12%	0%	7%	17%	49%	2%	1%
		Minutiae	343	13%	17%	0%	2%	6%	61%	0%	1%
		Level 3	3	14%	10%	0%	5%	24%	48%	0%	0%
		Other	5	22%	12%	0%	12%	10%	44%	0%	0%

Table S11: Reproducibility of exclusions and exclusion reasons in WB. When examiner A excluded, what examiner B did. Percentages are calculated as weighted sums over all other examiners assigned the same image pair, such that each exclusion by examiner A is weighted equally. (WB, n=560 examiner A exclusions).

An important factor contributing to the low reproducibility of exclusion reasons was individual examiner tendencies. Most exclusions were attributed to minutiae differences, but a few examiners attributed most of their exclusions to pattern class differences (Fig. S13). It is not known to what extent these explanations reflect substantive differences in how the decisions were made vs. how the examiners chose to describe their reasoning. Much of the dispersion shown in Fig. S13 may be due to random effects, including the random assignments of image pairs to examiners.

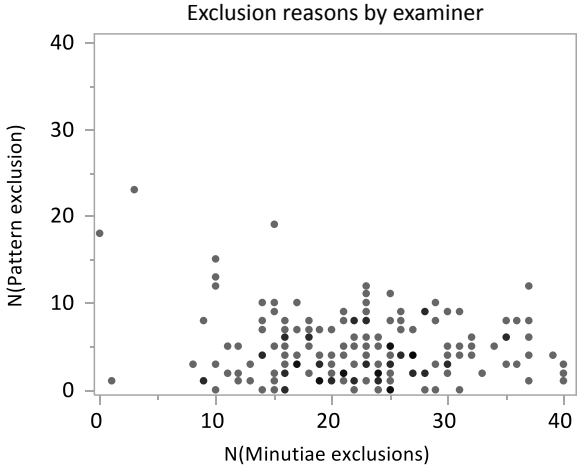

Fig. S13: Distribution of exclusion reasons given by examiners. Number of exclusions based on pattern class differences by number of exclusions based on minutiae differences (BB, n=169 examiners).

Table S12 describes the repeatability of exclusion reasons when examiners were retested after seven months (procedural details in [7]). At alpha=0.05, mated exclusions were significantly more likely to be repeated when the initial reason was Pattern Class than when it was Minutiae (41% vs. 26%); for nonmates, the difference was not statistically significant (86% vs. 92%).

	Initial Reason	Retest			Total	Exclusion repeated	Reason repeated
		Excluded		Not excluded			
		Pattern	Minutia				
Mates (FN)	Pattern	19	7	37	63	41%	30%
	Minutiae	4	38	121	163	26%	23%
	Not excluded	7	17	768	792		
	Total	30	62	926	1018		
Nonmates (TN)	Pattern	49	21	11	81	86%	60%
	Minutiae	25	331	33	389	92%	85%
	Not excluded	5	42	128	175		
	Total	79	394	172	645		

Table S12: Repeatability of exclusions, by reason. BB paired responses (test and retest) by 72 examiners after 7 months.

Appendix SI-8 LQMetric and latent value

The FBI's Latent Quality Metric (LQMetric) software automatically assesses the quality of latent fingerprint images. LQMetric was developed to predict whether a latent would match on an automated system; this ability to match is similar to but not always the same as how an examiner would assess the quality or value of a latent. LQMetric was calibrated to estimate the probability that an NGI image-only (LFIS) search would hit at rank 1, assuming the mate is in NGI. LQMetric is correlated with examiner value assessments and analysis minutia count (Fig. S14 and Fig. S15). Differences between BB and WB may be explained largely by differences in data selection (e.g., the NV latents in BB tended to be lower quality than those in WB).

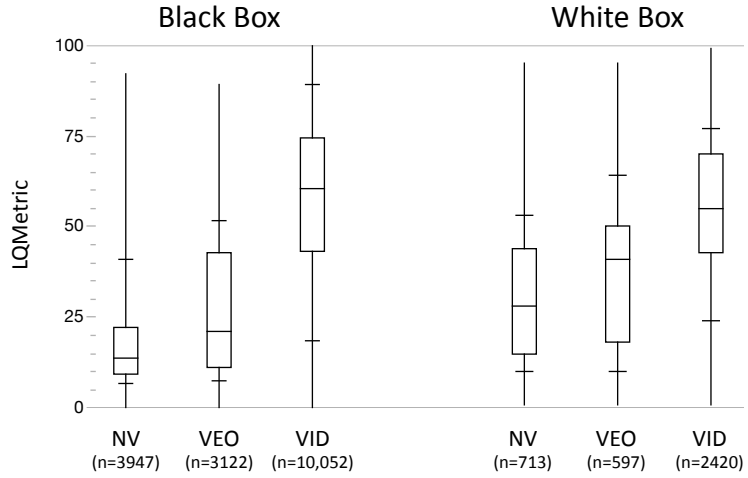

Fig. S14: Relation of examiner value assessments and LQMetric. BB: n=17,121 value determinations; WB: n=3730 analysis value determinations. Quartile box plots with crossbars indicating deciles (10%, 90%).

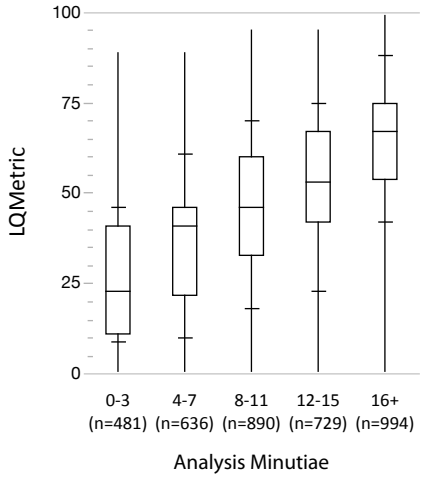

Fig. S15: Relation of analysis minutiae and LQMetric. (WB, 3730 analysis minutia markups). Quartile box plots with crossbars indicating deciles (10%, 90%).

Latent prints with high LQMetric values were associated with a greater proportion of exclusions being correct (high NPV). Fig. S16 shows how the various conclusion rates contribute to this result: as latent quality increases, fewer latents are assessed NV and more comparisons result in true conclusions (correct individualizations and exclusions); false negatives appear to be associated with moderate-quality latents; the false negative rate on low-quality latents was relatively low, in part because many of these prints were not compared. Table S13 through Table S16 present this data and the corresponding WB data in tabular form. The relation of LQMetric to NPV is also shown directly in Fig. S30.

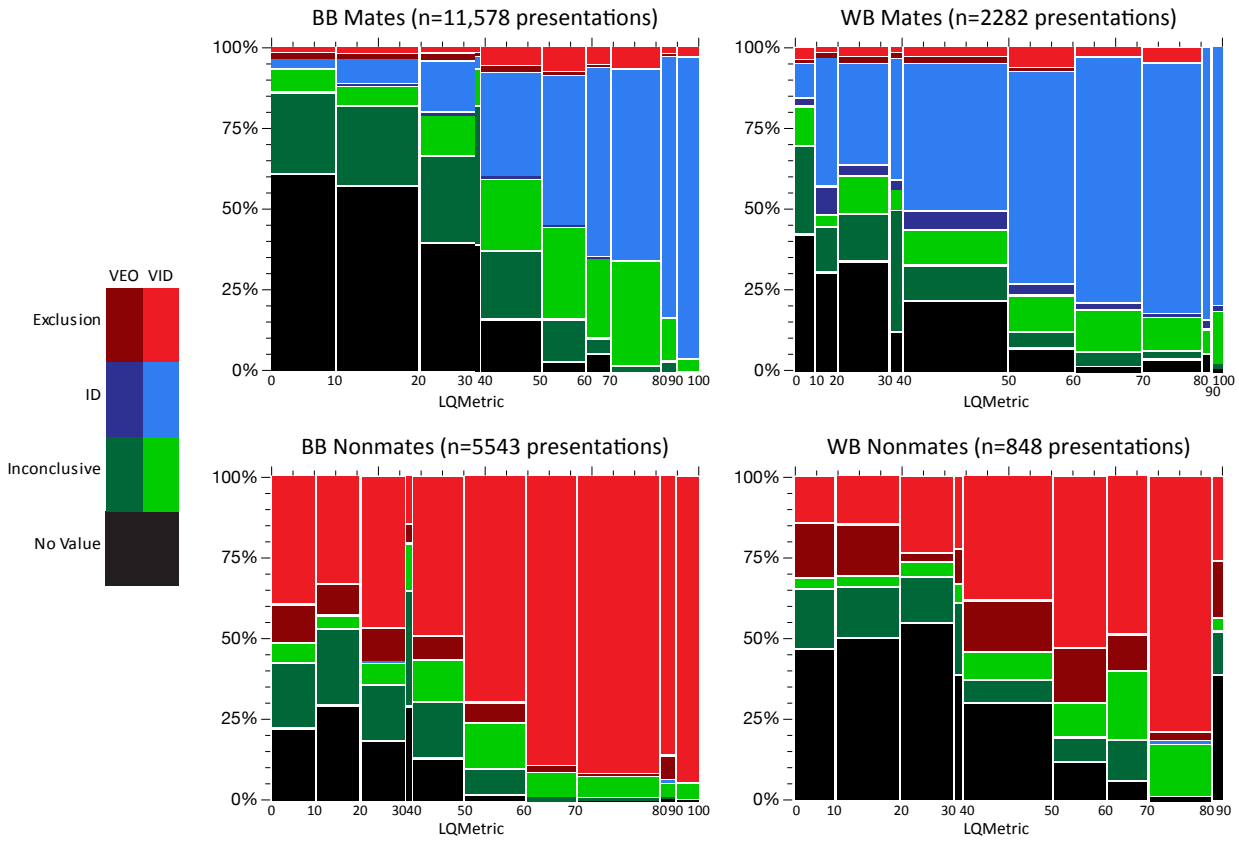

Fig. S16: Determinations by LQMetric, by mating.[h]

[h] *In the Black Box study, the process of selecting nonmated pairs filtered out a disproportionate number of NV latents. Thus for any given LQMetric value, a greater proportion of mated latents were assessed NV than nonmated latents. The reverse was true for White Box.*

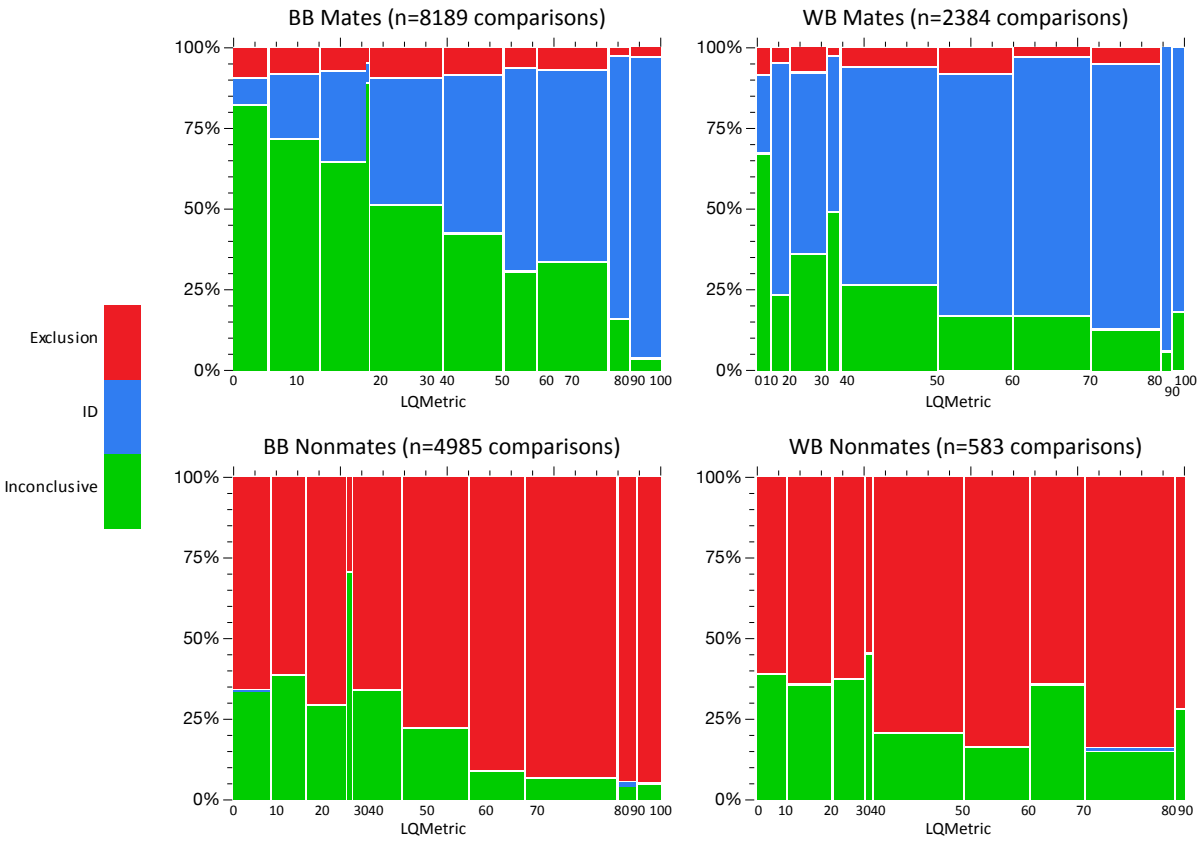

Fig. S17: Comparison determinations by LQMetric, by mating. Compare with Fig. S16, which describes determinations on all presentations; this describes comparisons only, omitting NV determinations, and does not differentiate VID and VEO.

LQMetric	NV	VEO			VID			FNR_{CMP}
		Excl	Inconc	Indiv	Excl	Inconc	Indiv	
0-10	1082	32	444	2	33	129	55	9%
10-20	1303	48	556	8	31	140	186	8%
20-30	601	38	402	14	30	185	238	7%
30-40	41	1	44	0	2	12	4	5%
40-50	274	28	347	8	101	368	534	9%
50-60	40	9	155	5	92	341	554	9%
60-70	43	4	29	1	38	172	404	6%
70-80	4	0	29	2	94	428	789	7%
80-90	1	1	13	0	10	55	330	3%
90-100	0	0	0	0	19	26	569	3%

Table S13: BB mated determinations by LQMetric (n=11,578 responses on mated pairs).

LQMetric	NV	VEO			VID			TNR_{CMP}
		Excl	Inconc	Indiv	Excl	Inconc	Indiv	
0-10	135	68	118	0	234	37	2	66%
10-20	172	56	132	0	190	24	0	61%
20-30	111	62	103	0	275	39	1	70%
30-40	20	4	24	0	10	10	0	29%
40-50	93	51	116	0	333	87	0	65%
50-60	17	49	64	0	545	110	1	77%
60-70	0	12	9	0	587	53	0	91%
70-80	6	6	10	0	981	69	0	93%
80-90	2	17	1	0	190	10	2	94%
90-100	2	0	0	0	277	16	0	95%

Table S14: BB nonmated determinations by LQMetric (n=5543 responses on nonmated pairs).

LQMetric	NV	VEO			VID			FNR_{CMP}
		Excl	Inconc	Indiv	Excl	Inconc	Indiv	
0-10	65	1	39	4	6	17	16	8%
10-20	50	2	20	13	3	5	61	5%
20-30	131	5	44	12	11	33	107	8%
30-40	13	1	31	3	1	6	33	3%
40-50	167	12	69	40	21	74	320	6%
50-60	37	5	21	16	29	51	296	8%
60-70	10	1	21	10	13	54	333	3%
70-80	19	1	10	4	19	41	311	5%
80-90	5	0	0	2	0	4	56	0%
90-100	1	0	1	1	0	13	61	0%

Table S15: WB mated determinations by LQMetric (n=2282 responses on mated pairs).

LQMetric	NV	VEO			VID			TNR_{CMP}
		Excl	Inconc	Indiv	Excl	Inconc	Indiv	
0-10	40	14	14	0	12	3	0	60%
10-20	66	20	20	0	19	2	0	64%
20-30	61	3	12	0	25	5	0	62%
30-40	7	2	4	0	4	1	0	55%
40-50	57	28	11	0	69	15	0	79%
50-60	18	18	5	0	56	9	0	84%
60-70	5	9	10	0	39	17	0	64%
70-80	3	4	0	0	98	19	1	84%
80-90	9	4	3	0	6	1	0	71%
90-100	0	0	0	0	0	0	0	60%

Table S16: WB nonmated determinations by LQMetric (n=848 responses on nonmated pairs).

Fig. S18 shows that inter-examiner reproducibility of true negatives increases with LQMetric; the reproducibility of false negatives is low regardless of quality.

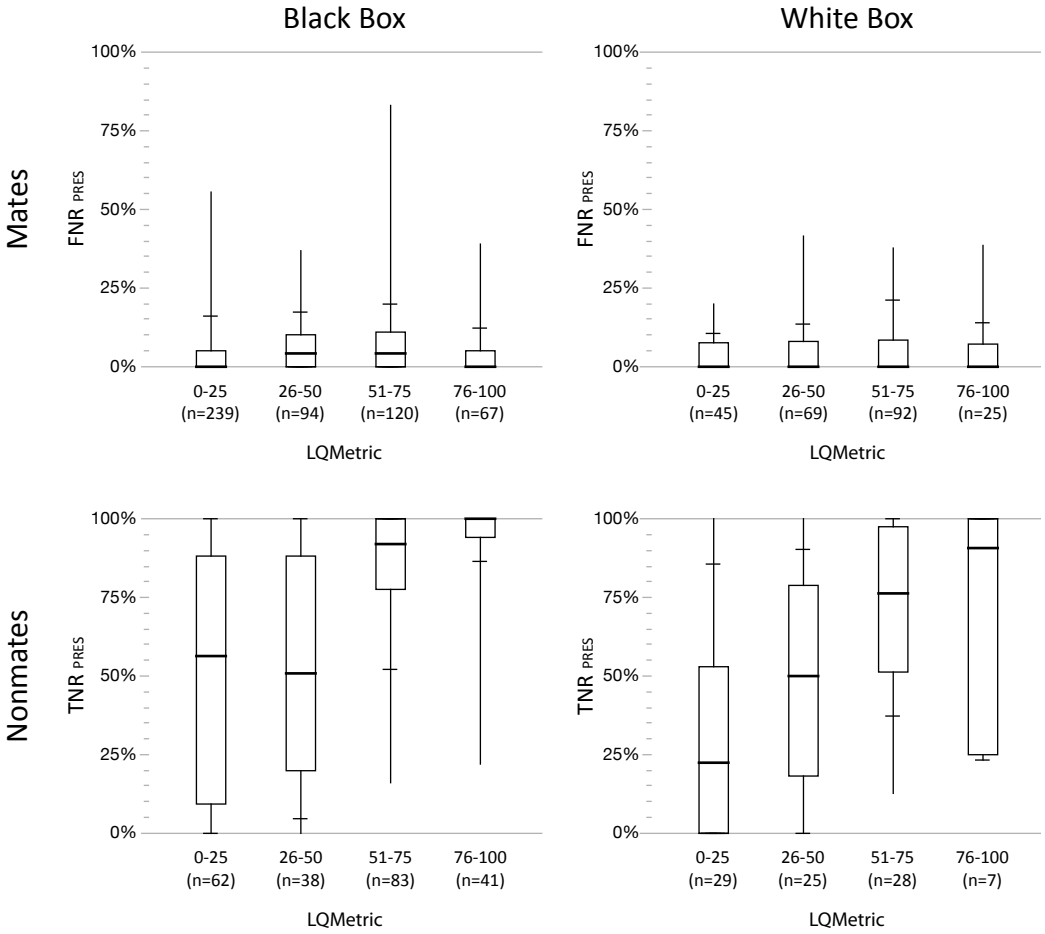

Fig. S18: Reproducibility of exclusions on image pairs by LQMetric. Quartile box plots with crossbars indicating deciles (10%, 90%).

Appendix SI-9 Analysis minutiae

Fig. S19 describes the distribution of comparison determinations as a function of analysis-phase minutia counts (see Figure 4 in the main document).

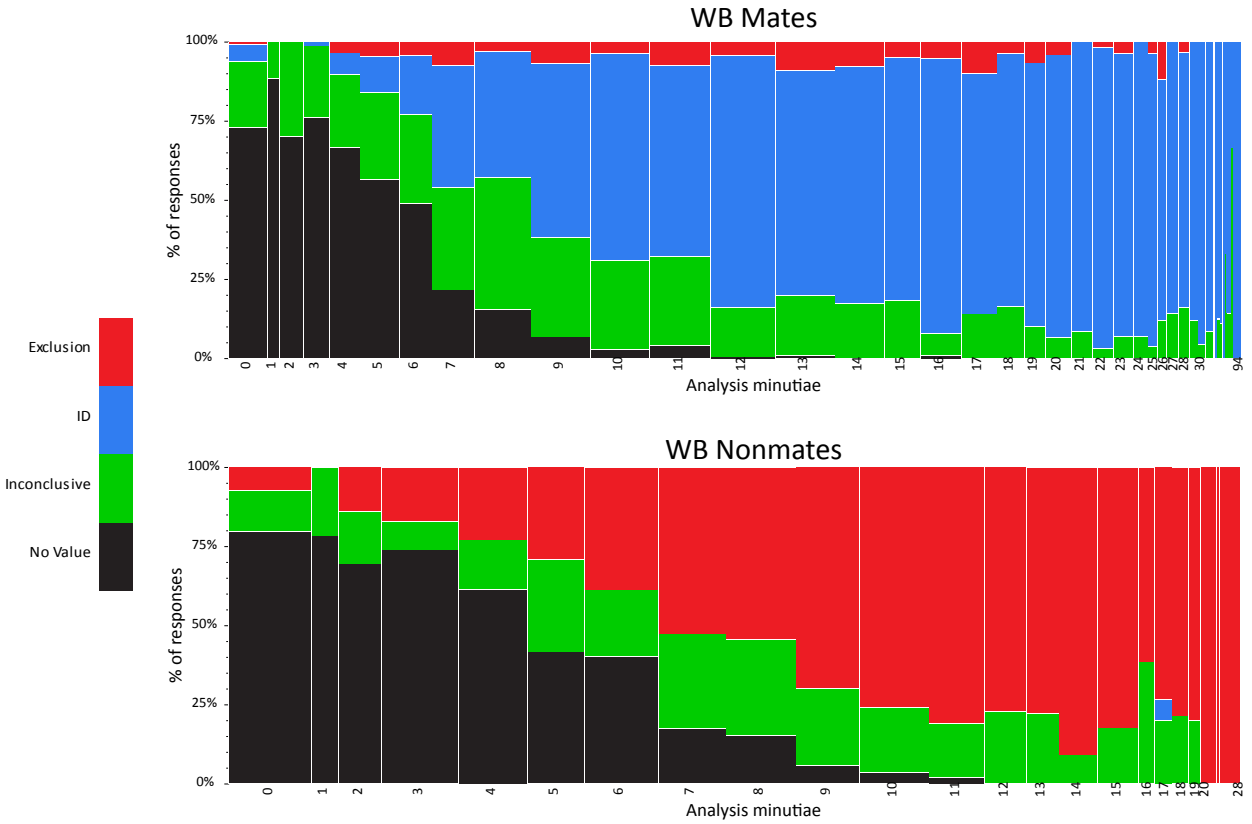

Fig. S19: Associations of examiner determinations with analysis-phase minutia counts. (WB, n=848 mates and 2882 nonmates). Note the limited sample sizes associated with high minutia counts.

Appendix SI-10 Discrepancies and corresponding minutiae

White Box examiners were instructed to mark any discrepancies used to support an exclusion determination, where a discrepancy was defined as "a feature that exists in one print and is definitely not present in the other print." Table S17 summarizes the distribution of comparisons on which discrepancies were marked, by mating, comparison determination, and exclusion reason. Examiners marked discrepancies on 31% of false negatives and 37% of true negatives. Examiners usually did not mark discrepancies on exclusions even when the reason was that minutiae differed. Discrepancies were sometimes marked on non-exclusions (43 inconclusives and 9 individualizations).

	Determination / Exclusion reason	Comparisons	Discrepancy marked: Total	%	Latent only	Exemplar only	Both
Mates	Exclusion (FN)	131	41	31%	17	6	18
	Pattern class	*12*	*1*	*8%*	*0*	*1*	*0*
	Core or delta	*8*	*2*	*25%*	*1*	*0*	*1*
	Minutiae	*104*	*35*	*34%*	*14*	*4*	*17*
	Level 3	*3*	*1*	*33%*	*0*	*1*	*0*
	Other	*3*	*1*	*33%*	*1*	*0*	*0*
	(missing reason)	*1*	*1*	*100%*	*1*	*0*	*0*
	Inconclusive	554	32	6%	20	3	9
	Individualization	1699	9	1%	8	1	0
	Total (mates)	**2384**	**82**	**3%**	**45**	**10**	**27**
Nonmates	Exclusion (TN)	430	159	37%	58	32	69
	Pattern class	*37*	*3*	*8%*	*3*	*0*	*0*
	Core or delta	*42*	*7*	*17%*	*3*	*2*	*2*
	Minutiae	*343*	*145*	*42%*	*52*	*30*	*63*
	Level 3	*3*	*1*	*33%*	*0*	*0*	*1*
	Other	*5*	*3*	*60%*	*0*	*0*	*3*
	Inconclusive	151	11	7%	8	0	3
	Individualization	1	0	0%	0	0	0
	Total (nonmates)	**582**	**170**	**29%**	**66**	**32**	**72**

Table S17. Proportions of comparisons where at least one discrepancy was marked, by mating and exclusion reason (WB n=2966 comparisons). 25 markups included nonminutia discrepancies (17 cores, 5 deltas, 3 others).

Examiners usually marked no corresponding minutiae when excluding (81% of exclusions). Fig. S20 and Table S18 summarize the counts of corresponding minutiae marked on those exclusions having at least one corresponding minutia marked (no corresponding minutiae were marked on 453 exclusions). As the number of corresponding minutiae increased, exclusions were more likely to be erroneous.

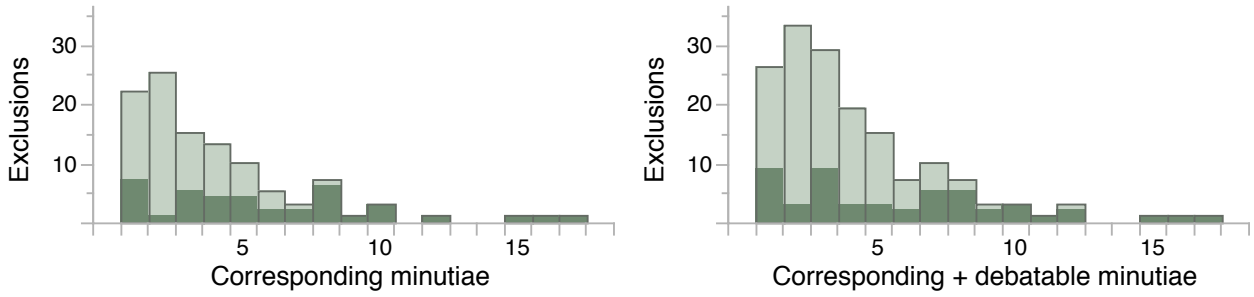

Fig. S20: Number of corresponding minutiae marked when examiners excluded in WB. Erroneous exclusions (mates) are shaded. (Left) definitive correspondences only (n=108 exclusions: 39 FN, 69 TN). (Right) definitive and debatable correspondences (n=159: 50 FN, 109 TN). Both charts omit markups on which no corresponding minutiae were marked.

		Definitive					
	Total	0		1-6		7+	
Mates (FN)	131	92	(70%)	23	(18%)	16	(12%)
Nonmates (TN)	430	361	(84%)	67	(16%)	2	(0.5%)
		Debatable					
	Total	0		1-6		7+	
Mates (FN)	131	111	(85%)	18	(14%)	2	(2%)
Nonmates (TN)	430	368	(86%)	60	(14%)	2	(0.5%)
		Definitive + debatable					
	Total	0		1-6		7+	
Mates (FN)	131	81	(62%)	29	(22%)	21	(16%)
Nonmates (TN)	430	321	(75%)	100	(23%)	9	(2.1%)

Table S18: Number of definitive and debatable corresponding minutiae marked when examiners excluded in WB.

		No corresponding minutiae		Corresponding minutiae	
	Count	No Discrepancies	Discrepancies	No Discrepancies	Discrepancies
FN	131	73 (56%)	8 (6%)	17 (13%)	33 (25%)
TN	430	246 (57%)	75 (17%)	25 (6%)	84 (20%)

Table S19. Proportions of false negatives and true negatives with and without corresponding minutiae (definitive or debatable) and discrepancies. (WB) Data includes 25 nonminutia discrepancies (17 cores, 5 deltas, 3 others).

Appendix SI-11 Corresponding cores and deltas

During data selection for WB, a pretest screening process determined whether a core or delta was present in both the latent and exemplar in each image pair. A corresponding core or delta was present on 126/231 mated pairs and an "apparently corresponding" (generally consistent shape and flow) core or delta on 46/89 nonmated pairs (54% of all image pair presentations). Examiners were instructed to mark during analysis all cores and deltas in the latent that could be accurately located to within approximately three ridge intervals. During Comparison, examiners were instructed: "For each feature marked in the latent, mark the corresponding feature if present in the exemplar." Despite these instructions, examiners often did not mark cores and deltas. Although multiple examiners marked a core or delta on 95% of the latents from image pairs selected as having a corresponding core or delta (lending support for the original classification), only 51% of analysis-phase latent markups indicated the presence of a core or delta on these latents. Among image pairs originally classified as having a corresponding core or delta, such correspondences were marked on only 8% of exclusions, 19% of inconclusives and 44% of individualizations (and never when NV).

Table S20 shows associations of determinations with the presence of a corresponding core or delta. The presence of a corresponding core or delta was associated with a higher rate of true negatives and a lower rate of false negatives. In WB, data selection controlled for corresponding minutia count, clarity, and complexity to avoid confounding these factors with the presence of cores and deltas.

A	Mates		Nonmates	
	CD	No CD	CD	No CD
NV	122	376	117	149
Exclusion	50	81	260	170
Inconclusive	298	256	65	86
Individualization	1106	593	0	1
Comparisons	1454	930	325	257
Exclusion rate$_{CMP}$	3.4%	8.7%	80.0%	66.1%

B	Mates		Nonmates	
	Marked CD	No marked CD	Marked CD	No marked CD
NV	0	498	0	266
Exclusion	8	123	22	408
Inconclusive	83	471	2	149
Individualization	548	1151	0	1
Comparisons	639	1745	24	558
Exclusion rate$_{CMP}$	1.3%	7.0%	91.7%	73.1%

Table S20: Association of determinations with the presence of a corresponding core or delta ("CD"). CD was determined (A) once for each image pair in a preliminary screening process; (B) by the examiner who made the determination marking the corresponding core or delta. (WB, n=3730).

Appendix SI-12 Difficulty

Examiners were asked to rate the difficulty of each comparison on a five-level scale from very easy to very difficult. Fig. S21 shows associations between examiners' difficulty ratings and their determinations: the more difficult an examiner described a comparison, the more likely that that examiner's comparison determination was inconclusive. Table S21 through Table S22 show interaction effects between difficulty and LQMetric. On BB difficult comparisons of mated pairs with low-LQMetric latents were ***less*** likely to be inconclusive; this reversal was not observed on WB (presumably due to smaller sample size, or differences in data selection and test procedures). Fig. S22 and Fig. S23 similarly describe associations between latent value assessments, difficulty, and comparison determinations.

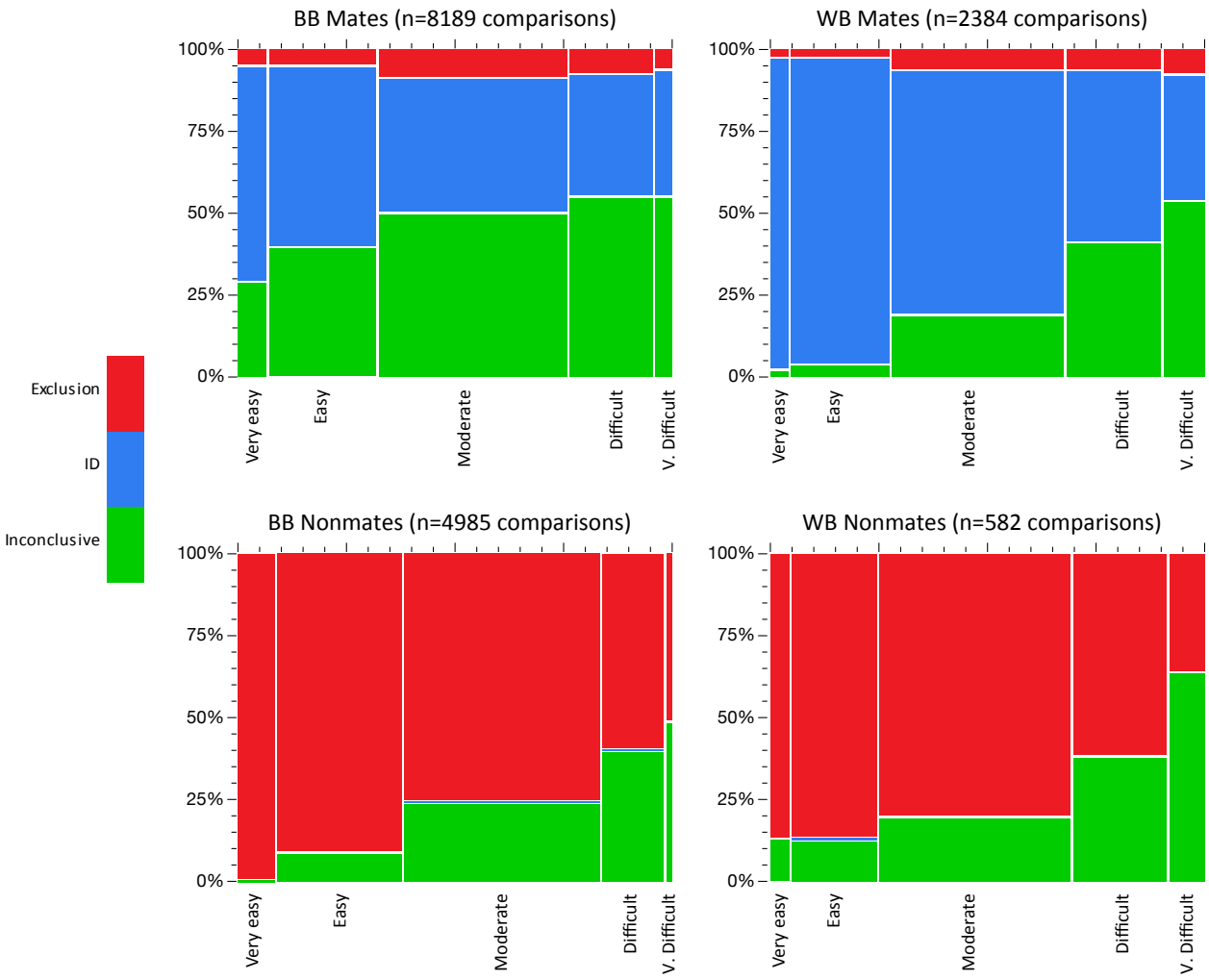

Fig. S21: Comparison determinations by difficulty and mating (BB, n=13,174; WB, n=2966).

		Mates					Nonmates				
LQMetric	**Difficulty**	**Count**	**NV**	**Inconc**	**Indiv**	**Excl (FNR_{CMP})**	**Count**	**NV**	**Inconc**	**Indiv**	**Excl (TNR_{CMP})**
	Very easy	391		19%	80%	1%	288		0%	0%	100%
	Easy	917		20%	77%	3%	687		1%	0%	99%
	Moderate	1004		27%	64%	9%	654		10%	0%	90%
High (67-100)	Difficult	279		38%	55%	8%	129		32%	0%	68%
	Very diff.	43		44%	56%	0%	14		50%	0%	50%
	(NV)	6					10				
	Total	2640	0%	25%	70%	5%	1782	1%	7%	0%	93%
	Very easy	121		30%	64%	6%	77		0%	0%	100%
	Easy	563		34%	61%	6%	439		10%	0%	90%
	Moderate	1412		48%	42%	11%	972		25%	0%	75%
Med (34-66)	Difficult	696		50%	41%	8%	310		42%	0%	57%
	Very diff.	141		52%	43%	6%	39		49%	0%	51%
	(NV)	365					122				
	Total	3298	11%	40%	41%	8%	1959	6%	23%	0%	71%
	Very easy	102		70%	12%	19%	94		4%	0%	96%
	Easy	562		78%	14%	8%	316		27%	0%	72%
	Moderate	1149		74%	18%	8%	646		39%	0%	61%
Low (0-33)	Difficult	659		68%	25%	7%	281		42%	0%	58%
	Very diff.	150		61%	30%	9%	39		49%	0%	51%
	(NV)	3018					426				
	Total	5640	54%	34%	9%	4%	1802	24%	26%	0%	50%

Table S21: Determinations by LQMetric and comparison difficulty (BB, n=11,578 responses on mates; 5543 responses on nonmated pairs).

		Mates					Nonmates				
LQMetric	**Difficulty**	**Count**	**NV**	**Inconc**	**Indiv**	**Excl (FNR_{CMP})**	**Count**	**NV**	**Inconc**	**Indiv**	**Excl (TNR_{CMP})**
	Very easy	70		1%	97%	1%	10		10%	0%	90%
	Easy	249		1%	97%	2%	49		8%	2%	90%
	Moderate	244		15%	80%	5%	70		16%	0%	84%
High (67-100)	Difficult	96		40%	53%	6%	27		48%	0%	52%
	Very diff.	39		56%	36%	8%	10		60%	0%	30%
	(NV)	24					15		0%	0%	0%
	Total	722	3%	14%	79%	4%	181	9%	19%	1%	71%
	Very easy	43		5%	91%	5%	6		0%	0%	100%
	Easy	253		7%	90%	4%	43		9%	0%	91%
	Moderate	563		18%	75%	6%	139		22%	0%	78%
Med (34-66)	Difficult	279		35%	57%	7%	63		32%	0%	68%
	Very diff.	107		52%	39%	7%	21		57%	0%	38%
	(NV)	195					69		0%	0%	0%
	Total	1440	14%	19%	62%	5%	341	21%	19%	0%	60%
	Very easy	0					14		21%	0%	79%
	Easy	53		11%	87%	2%	25		32%	0%	68%
	Moderate	160		33%	59%	8%	56		27%	0%	71%
Low (0-33)	Difficult	170		57%	37%	5%	40		48%	0%	53%
	Very diff.	98		59%	34%	7%	29		69%	0%	24%
	(NV)	239					162		0%	0%	0%
	Total	720	33%	30%	33%	4%	326	51%	20%	0%	29%

Table S22: Determinations by LQMetric and comparison difficulty (WB, n=2,882 responses on mated pairs; 848 responses on nonmates).

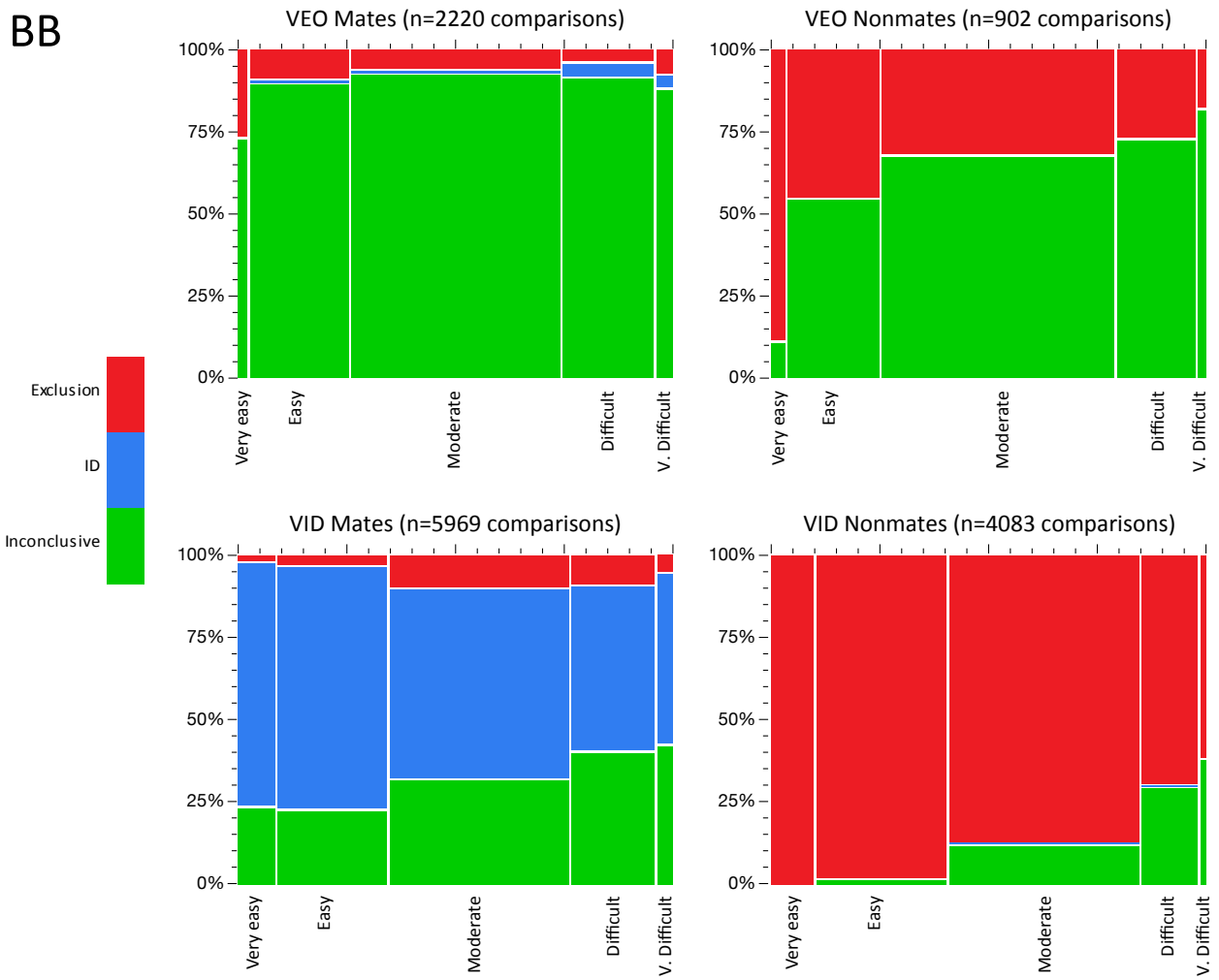

Fig. S22: Comparison determinations by difficulty, mating and latent value (BB, n=13,174 comparisons).

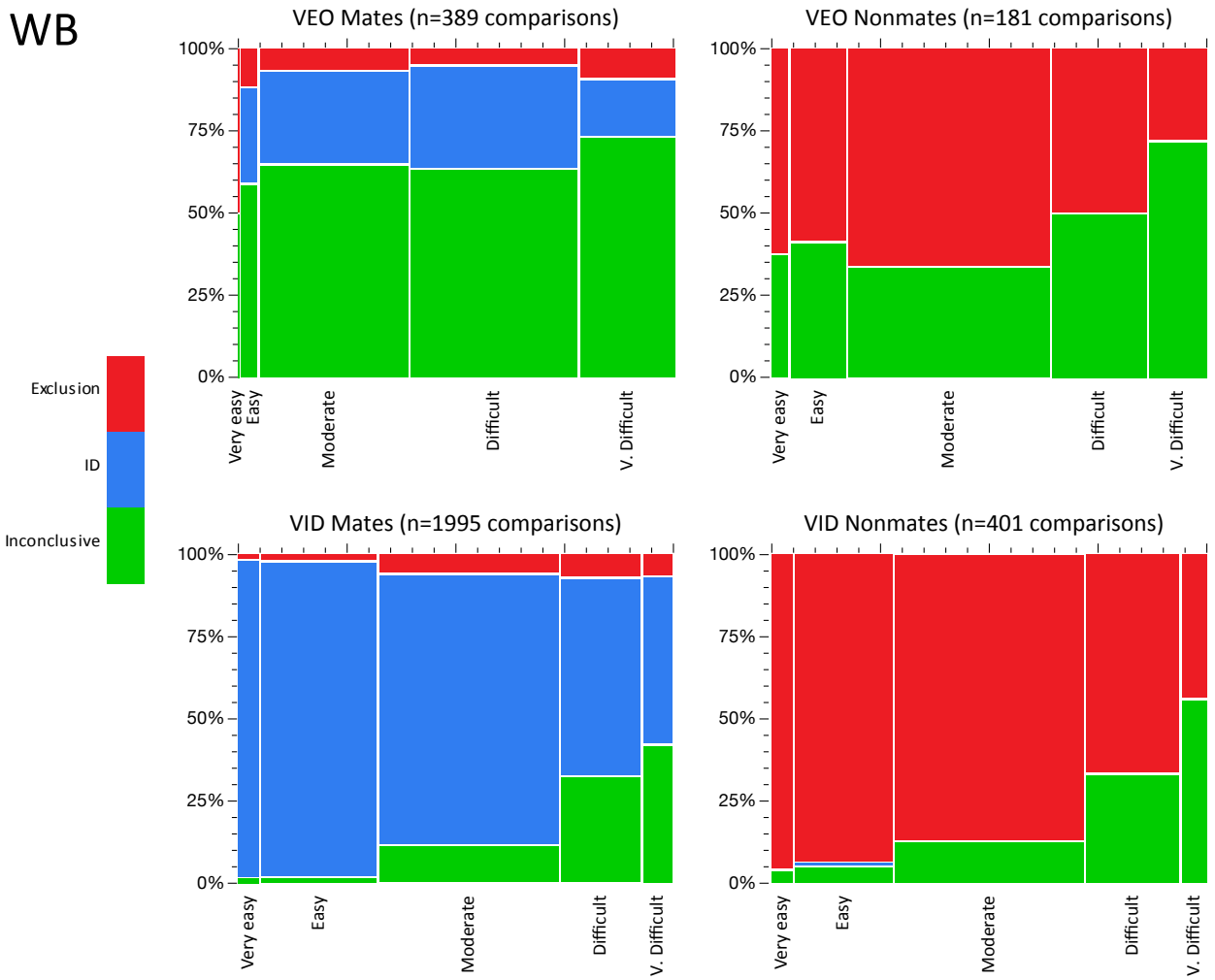

Fig. S23: Determinations by difficulty, mating and latent value (WB, n=2966 comparisons).

Appendix SI-13 Finger position

Table S23 presents data on the association between FNR_{CMP} and finger position. With the possible exception of left index fingers, on which we observe a high FNR, none of the other differences by finger position is statistically significant at alpha=0.05. Fig. S24shows associations between finger position and erroneous exclusions, restricted to mated pairs that were not unanimously inconclusive (or NV): this data also does not suggest a higher FNR on little fingers. Fig. S25 shows associations between finger position and VNP: again there was no notable association. It is possible that our process of selecting challenging image pairs may have been confounded with finger position

Ray and Dechant reported:

> *"Another trend at AZ DPS was erroneous exclusions on comparisons that eventually resulted in identifications to little fingers. Three of the errors (one third of the errors from fingers) were discovered on latent prints from little fingers …. Experience has shown that these fingers are the least likely to be identified. There seems to be some unconscious bias that leads examiners to spend less time on little fingers." [12]*

Our data does not confirm an association between little fingers and erroneous exclusions. However, Ray and Dechant's observation may have been due to the examiners' responding differently based on finger position; finger position was not indicated to the participants in this study.

Finger position	Compared mates	Excluded mates	FNR_{CMP}
01 R thumb	1142	81	7.1%
02 R index	855	64	7.5%
03 R middle	1009	68	6.7%
04 R ring	962	66	6.9%
05 R little	592	41	6.9%
06 L thumb	984	59	6.0%
07 L index	796	90	11.3%
08 L middle	707	50	7.1%
09 L ring	650	51	7.8%
10 L little	492	41	8.3%

Table S23: Counts of comparisons and exclusions by finger position. 95% binomial confidence intervals for the FNR_{CMP} values are approximately ±2% on this data, without accounting for the lack of independence (i.e., 2% is optimistically narrow). (BB, n=8189 mate comparisons).

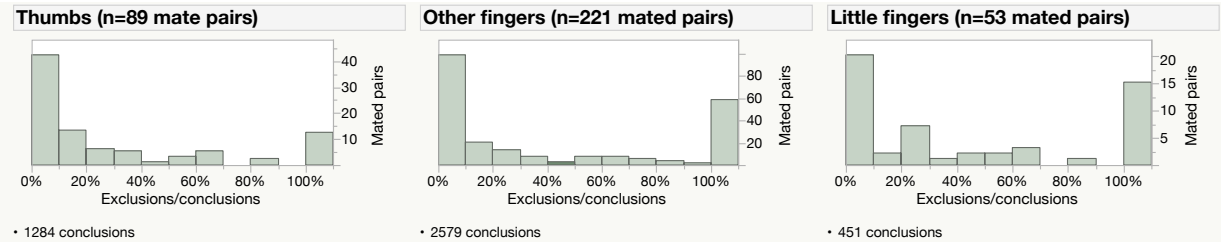

Fig. S24: Proportion of conclusions that were exclusions by finger position (conclusions = exclusions + individualizations). (BB, n=363 mated pairs on which at least one examiner concluded; 4314 conclusions).

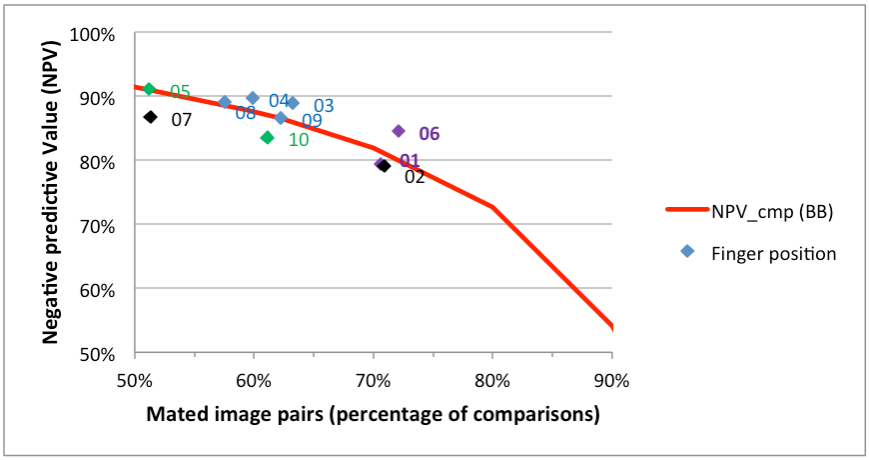

Fig. S25: Latent finger position as a predictor of NPV. Data points are colored by finger positions: thumbs (purple: 01, 06); index (black: 02, 07); little fingers (green: 05, 10). n=13,174 comparisons; the smallest subsample is finger 10 (left little) which included 492 comparisons of 49 mated pairs and 313 comparisons of 15 nonmated pairs (and resulted in 41 false negatives on 17 mated pairs, 208 true negatives on 14 nonmated pairs).

Appendix SI-14 Summary of factors associated with exclusions

Table S24 and Table S26 summarize associations between various factors and measures of accuracy for BB and WB, respectively. Table S25 and Table S27 report binomial confidence intervals for those measures. Fig. S26 through Fig. S29 depict much of this information in a graphical summary.

The measured value of NPV (NPV_{RAW}) depends substantially on the prevalence of mated pairs among the examinations performed (main paper, section 3.4). Therefore, in order to compare these measured values meaningfully, we project the measurements to a standard mating proportion: NPV_{50} is an estimate of what NPV would be if the mating proportion for each level of the factor were 50% mates. This projection (Appendix SI-15) requires knowing for each level of a factor what proportion of the comparisons were mated. For example, for VID, 59% of comparisons are mated, and 89% of the exclusions were on mated pairs (NPVraw=89%); after projecting to a 50:50 mix using the method described in in Appendix SI-15, NPV_{50} is 92%.

For difficulty, we have calculated %Mates and NPV_{50} as was done for the other rows, but show the measures in gray to indicate our strong reservations regarding the implicit modeling assumptions, namely that the prior %Mates compared for each difficulty level can be estimated as the proportion measured posterior to the comparisons. We are dubious of this approach because we expect some degree of confounding with the examiners' comparison determinations. For example, "easy" may have different meanings when referring to exclusions, inconclusives, and individualizations.

BLACK BOX		Presentations			Comparisons			Exclusions		PRES exclusion rates		CMP exclusion rates		NPV	
Factor	**Level**	**Mates**	**Nonmates**	**% Mates**	**Mates**	**Nonmates**	**% Mates**	**Mates**	**Nonmates**	**FNR_{PRES}**	**TNR_{PRES}**	**FNR_{CMP}**	**TNR_{CMP}**	**NPV_{RAW}**	**NPV_{50}**
	0-20	4049	1168	78%	1664	861	66%	144	548	3.6%	46.9%	8.7%	63.6%	79%	88%
	20-40	1612	659	71%	970	528	65%	71	351	4.4%	53.3%	7.3%	66.5%	83%	90%
LQMetric	40-60	2856	1466	66%	2542	1356	65%	230	978	8.1%	66.7%	9.0%	72.1%	81%	89%
	60-80	2037	1733	54%	1990	1727	54%	136	1586	6.7%	91.5%	6.8%	91.8%	92%	93%
	80-100	1024	517	66%	1023	513	67%	30	484	2.9%	93.6%	2.9%	94.3%	94%	97%
	NV	3389	558	86%	N/A	N/A	N/A	N/A	N/A	0.0%	0.0%	N/A	N/A	N/A	N/A
Value	VEO	2220	902	71%	2220	902	71%	161	325	7.3%	36.0%	7.3%	36.0%	67%	83%
	VID	5969	4083	59%	5969	4083	59%	450	3622	7.5%	88.7%	7.5%	88.7%	89%	92%
	V. diff.	N/A	N/A	N/A	334	92	78%	21	47	N/A	N/A	6.3%	51.1%	69%	N/A
	Difficult	N/A	N/A	N/A	1634	720	69%	127	429	N/A	N/A	7.8%	59.6%	77%	N/A
Difficulty	Moderate	N/A	N/A	N/A	3565	2272	61%	326	1711	N/A	N/A	9.1%	75.3%	84%	N/A
	Easy	N/A	N/A	N/A	2042	1442	59%	106	1306	N/A	N/A	5.2%	90.6%	92%	N/A
	V. easy	N/A	N/A	N/A	614	459	57%	31	454	N/A	N/A	5.0%	98.9%	94%	N/A
Excl reason	Pattern	N/A	N/A	N/A	N/A	N/A	N/A	174	624	N/A	N/A	N/A	N/A	78%	N/A
	Minutiae	N/A	N/A	N/A	N/A	N/A	N/A	437	3323	N/A	N/A	N/A	N/A	88%	N/A
	Not certified	1903	802	70%	1368	722	65%	108	516	5.7%	64.3%	7.9%	71.5%	83%	90%
Certification	IAI CLPE	5547	2268	71%	3988	2082	66%	309	1684	5.6%	74.3%	7.7%	80.9%	84%	91%
	Other certification	3434	2169	61%	2363	1912	55%	152	1534	4.4%	70.7%	6.4%	80.2%	91%	93%
Overall		11578	5543	68%	8189	4985	62%	611	3947	5.3%	71.2%	7.5%	79.2%	87%	91%

Table S24: Summary of factors affecting exclusions in BB. NPV_{RAW} = (nonmate exclusions)/(total exclusions); NPV_{50} is projected to 50% mates (% compared) as described in Appendix SI-15. Confidence intervals for FNR and TNR are in Table S25. For difficulty, the % mates compared is shown in gray because this response is confounded with mating. Certification was not reported by all BB examiners.

BLACK BOX		PRES exclusion rates						CMP exclusion rates					
Factor	**Level**	**FNR_{PRES}**	**$FNR_{PRES\text{-}Low}$**	**$FNR_{PRES\text{-}High}$**	**TNR_{PRES}**	**$TNR_{PRES\text{-}Low}$**	**$TNR_{PRES\text{-}High}$**	**FNR_{CMP}**	**$FNR_{CMP\text{-}Low}$**	**$FNR_{CMP\text{-}High}$**	**TNR_{CMP}**	**$TNR_{CMP\text{-}Low}$**	**$TNR_{CMP\text{-}High}$**
LQMetric	0-20	3.6%	3.0%	4.2%	46.9%	44.0%	49.8%	8.7%	7.3%	10.1%	63.6%	60.3%	66.9%
	20-40	4.4%	3.5%	5.5%	53.3%	49.4%	57.1%	7.3%	5.8%	9.1%	66.5%	62.3%	70.5%
	40-60	8.1%	7.1%	9.1%	66.7%	64.2%	69.1%	9.0%	8.0%	10.2%	72.1%	69.7%	74.5%
	60-80	6.7%	5.6%	7.8%	91.5%	90.1%	92.8%	6.8%	5.8%	8.0%	91.8%	90.4%	93.1%
	80-100	2.9%	2.0%	4.2%	93.6%	91.2%	95.6%	2.9%	2.0%	4.2%	94.3%	92.0%	96.2%
Value	NV	0.0%	N/A	N/A	0.0%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	VEO	7.3%	6.2%	8.4%	36.0%	32.9%	39.3%	7.3%	6.2%	8.4%	36.0%	32.9%	39.3%
	VID	7.5%	6.9%	8.2%	88.7%	87.7%	89.7%	7.5%	6.9%	8.2%	88.7%	87.7%	89.7%
Difficulty	V. diff.	N/A	N/A	N/A	N/A	N/A	N/A	6.3%	3.9%	9.5%	51.1%	40.4%	61.7%
	Difficult	N/A	N/A	N/A	N/A	N/A	N/A	7.8%	6.5%	9.2%	59.6%	55.9%	63.2%
	Moderate	N/A	N/A	N/A	N/A	N/A	N/A	9.1%	8.2%	10.1%	75.3%	73.5%	77.1%
	Easy	N/A	N/A	N/A	N/A	N/A	N/A	5.2%	4.3%	6.2%	90.6%	88.9%	92.0%
	V. easy	N/A	N/A	N/A	N/A	N/A	N/A	5.0%	3.5%	7.1%	98.9%	97.5%	99.6%
Excl reason	Pattern	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Minutiae	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Certification	Not certified	5.7%	4.7%	6.8%	64.3%	60.9%	67.7%	7.9%	6.5%	9.5%	71.5%	68.0%	74.7%
	IAI CLPE	5.6%	5.0%	6.2%	74.3%	72.4%	76.0%	7.7%	6.9%	8.6%	80.9%	79.1%	82.6%
	Other certification	4.4%	3.8%	5.2%	70.7%	68.8%	72.6%	6.4%	5.5%	7.5%	80.2%	78.4%	82.0%
Overall		5.3%	4.9%	5.7%	71.2%	70.0%	72.4%	7.5%	6.9%	8.1%	79.2%	78.0%	80.3%

Table S25: BB 95% binomial confidence intervals for FNR and TNR for the factors shown in Table S24.

WHITE BOX		Presentations			Comparisons			Exclusions		PRES exclusion rates		CMP exclusion rates		NPV	
Factor	**Level**	**Mates**	**Nonmates**	**% Mates**	**Mates**	**Nonmates**	**% Mates**	**Mates**	**Nonmates**	**FNR_{PRES}**	**TNR_{PRES}**	**FNR_{CMP}**	**TNR_{CMP}**	**NPV_{RAW}**	**NPV_{50}**
LQMetric	0-20	302	210	59%	187	104	64%	12	65	4.0%	31.0%	6.4%	62.5%	84%	91%
	20-40	431	124	78%	287	56	84%	18	34	4.2%	27.4%	6.3%	60.7%	65%	91%
	40-60	1158	286	80%	954	211	82%	67	171	5.8%	59.8%	7.0%	81.0%	72%	92%
	60-80	847	205	81%	818	197	81%	34	150	4.0%	73.2%	4.2%	76.1%	82%	95%
	80-100	144	23	86%	138	14	91%	0	10	0.0%	43.5%	0.0%	71.4%	100%	100%
Value	NV	462	251	65%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	VEO	407	190	68%	389	181	68%	28	102	6.9%	53.7%	7.2%	56.4%	78%	89%
	VID	2013	407	83%	1995	401	83%	103	328	5.1%	80.6%	5.2%	81.8%	76%	94%
Difficulty	V. diff.	N/A	N/A	N/A	232	50	82%	18	18	N/A	N/A	7.8%	36.0%	50%	N/A
	Difficult	N/A	N/A	N/A	528	126	81%	35	78	N/A	N/A	6.6%	61.9%	69%	N/A
	Moderate	N/A	N/A	N/A	959	260	79%	60	208	N/A	N/A	6.3%	80.0%	78%	N/A
	Easy	N/A	N/A	N/A	552	116	83%	15	100	N/A	N/A	2.7%	86.2%	87%	N/A
	V. easy	N/A	N/A	N/A	113	30	79%	3	26	N/A	N/A	2.7%	86.7%	90%	N/A
Excl reason	Pattern	N/A	N/A	N/A	N/A	N/A	N/A	20	79	N/A	N/A	N/A	N/A	80%	N/A
	Minutiae	N/A	N/A	N/A	N/A	N/A	N/A	110	351	N/A	N/A	N/A	N/A	76%	N/A
Certification	Not certified	322	94	77%	258	58	82%	13	40	4.0%	42.6%	5.0%	69.0%	75%	93%
	IAI CLPE	951	280	77%	803	206	80%	55	148	5.8%	52.9%	6.8%	71.8%	73%	91%
	Other certification	1592	469	77%	1311	313	81%	63	237	4.0%	50.5%	4.8%	75.7%	79%	94%
Core-Delta	No	1306	406	76%	930	257	78%	81	170	6.2%	41.9%	8.7%	66.1%	68%	88%
	Yes	1576	442	78%	1454	325	82%	50	260	3.2%	58.8%	3.4%	80.0%	84%	96%
Analysis minutiae	0-3	288	193	60%	67	46	59%	1	21	0.3%	10.9%	1.5%	45.7%	95%	97%
	4-7	412	224	65%	210	129	62%	21	81	5.1%	36.2%	10.0%	62.8%	79%	86%
	8-11	673	217	76%	609	195	76%	36	151	5.3%	69.6%	5.9%	77.4%	81%	93%
	12-15	600	129	82%	591	127	82%	39	106	6.5%	82.2%	6.6%	83.5%	73%	93%
	16+	909	85	91%	907	85	91%	34	71	3.7%	83.5%	3.7%	83.5%	68%	96%
Median analysis minutiae	0-3	163	174	48%	36	42	46%	4	14	2.5%	8.0%	11.1%	33.3%	78%	75%
	4-7	493	242	67%	258	140	65%	28	93	5.7%	38.4%	10.9%	66.4%	77%	86%
	8-11	783	257	75%	671	226	75%	44	176	5.6%	68.5%	6.6%	77.9%	80%	92%
	12-15	576	123	82%	555	122	82%	25	100	4.3%	81.3%	4.5%	82.0%	80%	95%
	16+	867	52	94%	864	52	94%	30	47	3.5%	90.4%	3.5%	90.4%	61%	96%
Overall		2882	848	77%	2384	582	80%	131	430	4.5%	50.7%	5.5%	73.9%	77%	93%

Table S26: Summary of factors affecting exclusions in WB. Confidence intervals for FNR and TNR are in Table S27. Certification was not reported by one WB examiner. In the Comparison phase, mates and nonmates were categorized as VEO or VID based on that examiner's analysis-phase assessment; assessments for which we do not have comparison determinations are omitted (including latent reassessed as NV, exemplar assessed as NV, and one missing determination). For difficulty, the % mates compared is shown in gray because this response is confounded with mating.

WHITE BOX		PRES exclusion rates						CMP exclusion rates					
Factor	**Level**	**FNR_{PRES}**	**$FNR_{PRES\text{-}Low}$**	**$FNR_{PRES\text{-}High}$**	**TNR_{PRES}**	**$TNR_{PRES\text{-}Low}$**	**$TNR_{PRES\text{-}High}$**	**FNR_{CMP}**	**$FNR_{CMP\text{-}Low}$**	**$FNR_{CMP\text{-}High}$**	**TNR_{CMP}**	**$TNR_{CMP\text{-}Low}$**	**$TNR_{CMP\text{-}High}$**
LQMetric	0-20	4.0%	2.1%	6.8%	31.0%	24.8%	37.7%	6.4%	3.4%	10.9%	62.5%	52.5%	71.8%
	20-40	4.2%	2.5%	6.5%	27.4%	19.8%	36.2%	6.3%	3.8%	9.7%	60.7%	46.8%	73.5%
	40-60	5.8%	4.5%	7.3%	59.8%	53.9%	65.5%	7.0%	5.5%	8.8%	81.0%	75.1%	86.1%
	60-80	4.0%	2.8%	5.6%	73.2%	66.6%	79.1%	4.2%	2.9%	5.8%	76.1%	69.6%	81.9%
	80-100	0.0%	0.0%	2.5%	43.5%	23.2%	65.5%	0.0%	0.0%	2.6%	71.4%	41.9%	91.6%
Value	NV	0.0%	N/A	N/A	0.0%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	VEO	6.9%	4.6%	9.8%	53.7%	46.3%	60.9%	7.2%	4.8%	10.2%	56.4%	48.8%	63.7%
	VID	5.1%	4.2%	6.2%	80.6%	76.4%	84.3%	5.2%	4.2%	6.2%	81.8%	77.7%	85.5%
Difficulty	V. diff.	N/A	N/A	N/A	N/A	N/A	N/A	7.8%	4.7%	12.0%	36.0%	22.9%	50.8%
	Difficult	N/A	N/A	N/A	N/A	N/A	N/A	6.6%	4.7%	9.1%	61.9%	52.8%	70.4%
	Moderate	N/A	N/A	N/A	N/A	N/A	N/A	6.3%	4.8%	8.0%	80.0%	74.6%	84.7%
	Easy	N/A	N/A	N/A	N/A	N/A	N/A	2.7%	1.5%	4.4%	86.2%	78.6%	91.9%
	V. easy	N/A	N/A	N/A	N/A	N/A	N/A	2.7%	0.6%	7.6%	86.7%	69.3%	96.2%
Excl reason	Pattern	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Minutiae	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Certification	Not certified	4.0%	2.2%	6.8%	42.6%	32.4%	53.2%	5.0%	2.7%	8.5%	69.0%	55.5%	72.4%
	IAI CLPE	5.8%	4.4%	7.5%	52.9%	46.8%	58.8%	6.8%	5.2%	8.8%	71.8%	65.2%	73.8%
	Other certification	4.0%	3.1%	5.0%	50.5%	45.9%	55.1%	4.8%	3.7%	6.1%	75.7%	70.6%	77.0%
Core-Delta	No	6.2%	5.0%	7.7%	41.9%	37.0%	46.8%	8.7%	7.0%	10.7%	66.1%	60.0%	71.9%
	Yes	3.2%	2.4%	4.2%	58.8%	54.1%	63.5%	3.4%	2.6%	4.5%	80.0%	75.2%	84.2%
Analysis minutiae	0-3	0.3%	0.0%	1.9%	10.9%	6.9%	16.2%	1.5%	0.0%	8.0%	45.7%	30.9%	61.0%
	4-7	5.1%	3.2%	7.7%	36.2%	29.9%	42.8%	10.0%	6.3%	14.9%	62.8%	53.8%	71.1%
	8-11	5.3%	3.8%	7.3%	69.6%	63.0%	75.6%	5.9%	4.2%	8.1%	77.4%	70.9%	83.1%
	12-15	6.5%	4.7%	8.8%	82.2%	74.5%	88.3%	6.6%	4.7%	8.9%	83.5%	75.8%	89.5%
	16+	3.7%	2.6%	5.2%	83.5%	73.9%	90.7%	3.7%	2.6%	5.2%	83.5%	73.9%	90.7%
Median analysis minutiae	0-3	2.5%	0.7%	6.2%	8.0%	4.5%	13.1%	11.1%	3.1%	26.1%	33.3%	19.6%	49.5%
	4-7	5.7%	3.8%	8.1%	38.4%	32.3%	44.9%	10.9%	7.3%	15.3%	66.4%	58.0%	74.2%
	8-11	5.6%	4.1%	7.5%	68.5%	62.4%	74.1%	6.6%	4.8%	8.7%	77.9%	71.9%	83.1%
	12-15	4.3%	2.8%	6.3%	81.3%	73.3%	87.8%	4.5%	2.9%	6.6%	82.0%	74.0%	88.3%
	16+	3.5%	2.3%	4.9%	90.4%	79.0%	96.8%	3.5%	2.4%	4.9%	90.4%	79.0%	96.8%
Overall		4.5%	3.8%	5.4%	50.7%	47.3%	54.1%	5.5%	4.6%	6.5%	73.9%	70.1%	77.4%

Table S27: WB 95% binomial confidence intervals for FNR and TNR for the factors shown in Table S26.

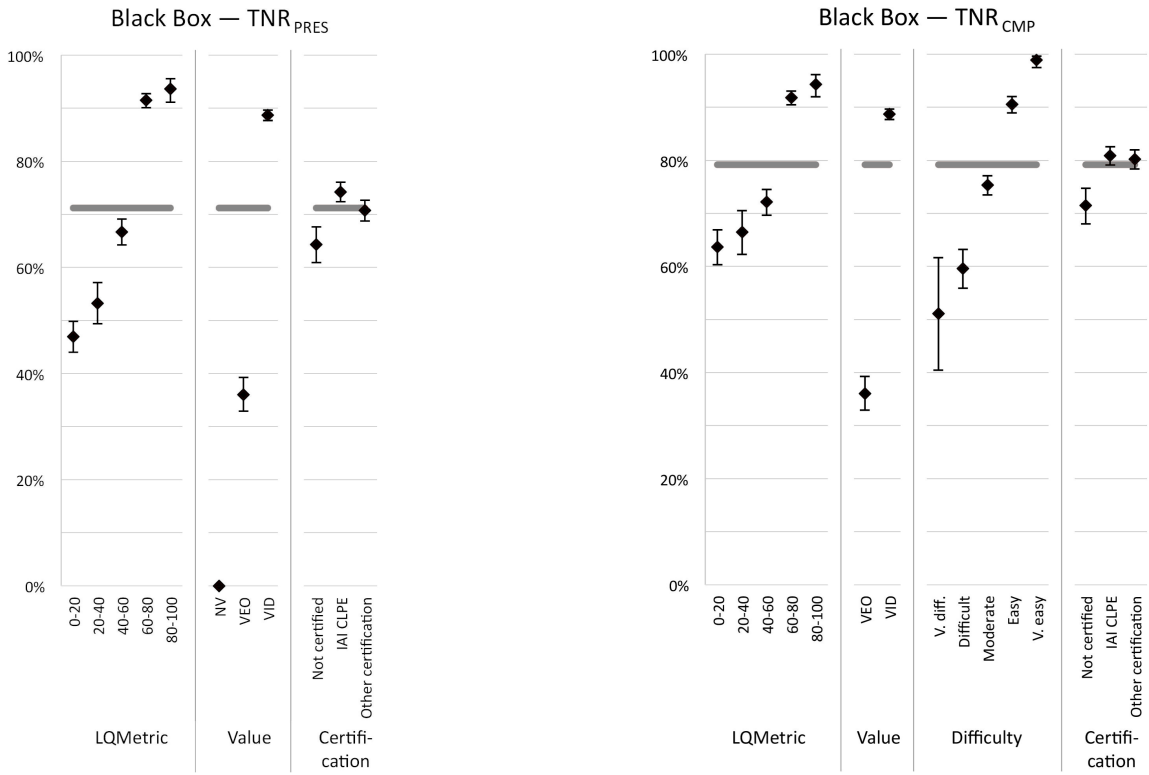

Fig. S26: Comparison of BB TNR_{PRES} and TNR_{CMP} for factors. Horizontal lines indicate overall mean rates: TNR_{PRES}=71.2% (5543 presentations), TNR_{CMP}=79.2% (4985 comparisons).

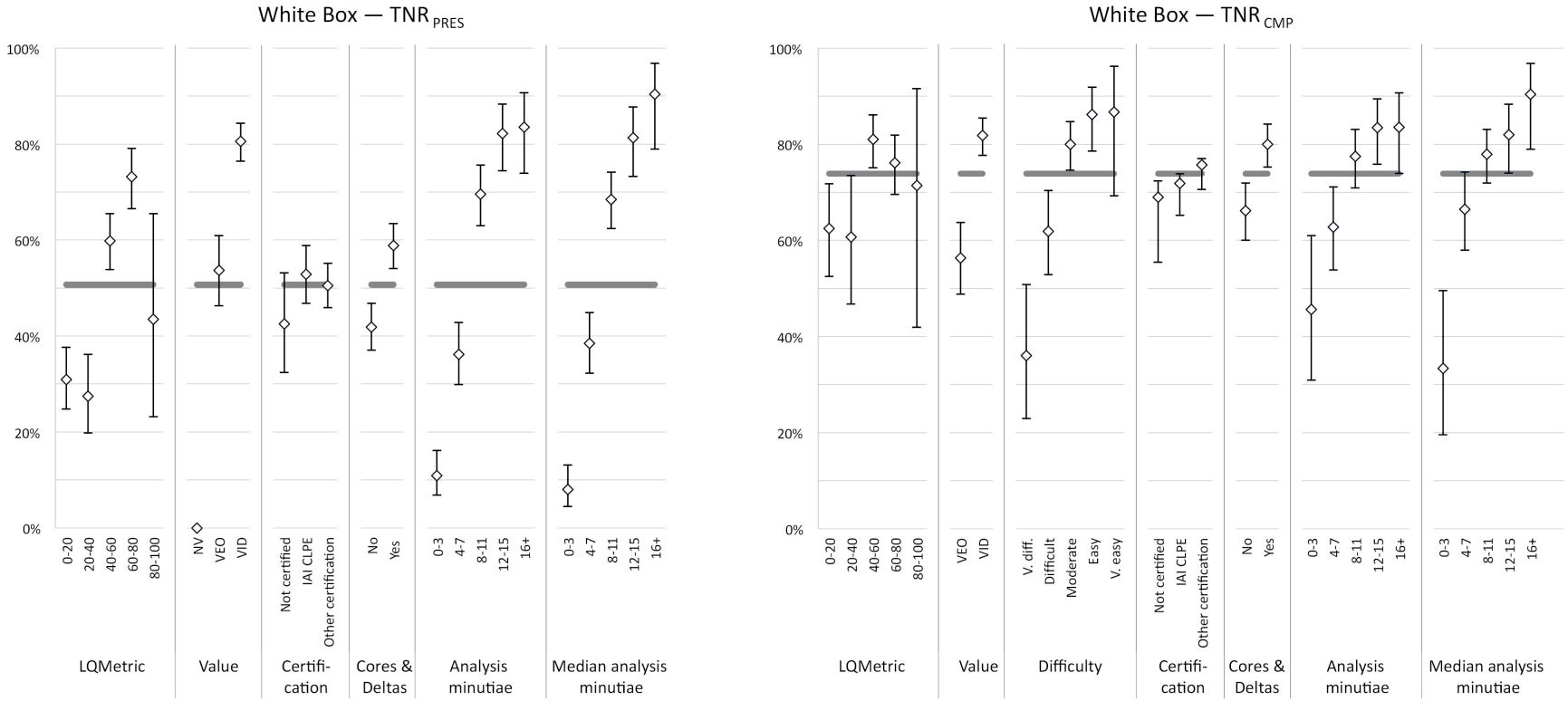

Fig. S27: Comparison of WB TNR_{PRES} and TNR_{CMP} for factors. Horizontal lines indicate overall mean rates: TNR_{PRES}=50.7% (848 presentations), TNR_{CMP}=73.9% (582 comparisons).

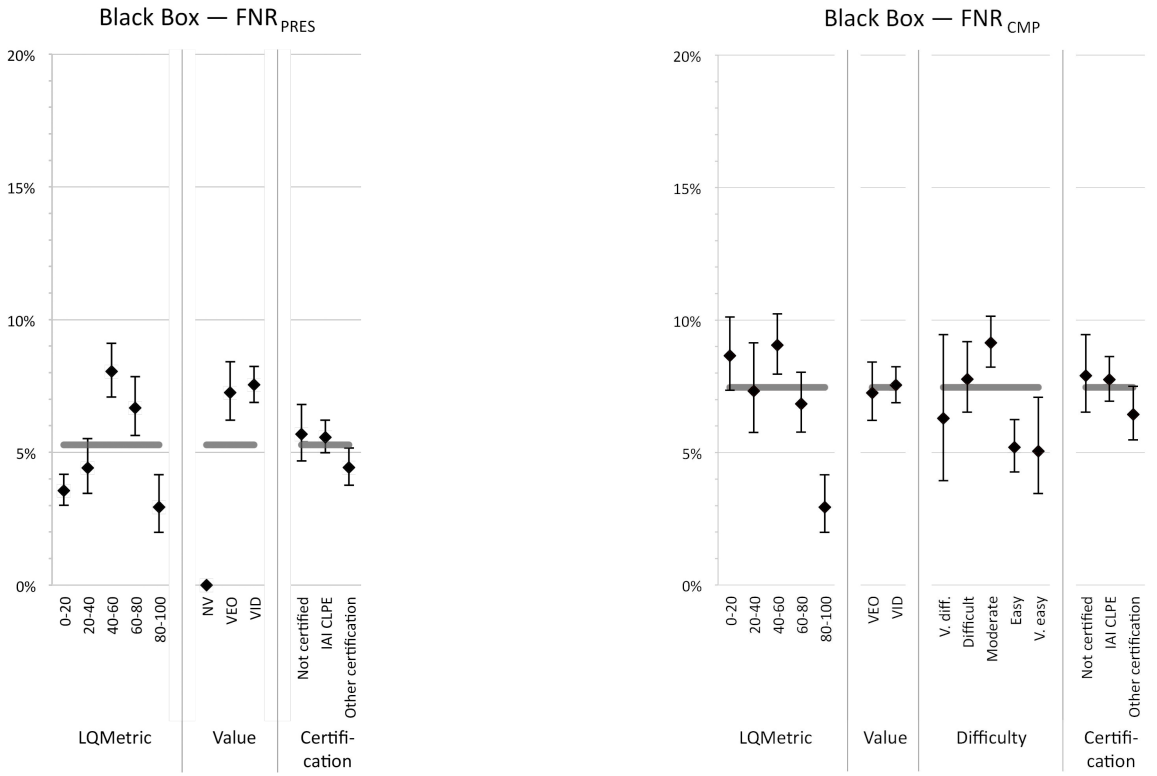

Fig. S28: Comparison of BB FNR_{PRES} and FNR_{CMP} for factors. Horizontal lines indicate overall mean rates: FNR_{PRES}=5.3% (11,578 presentations), FNR_{CMP}=7.5% (8189 comparisons).

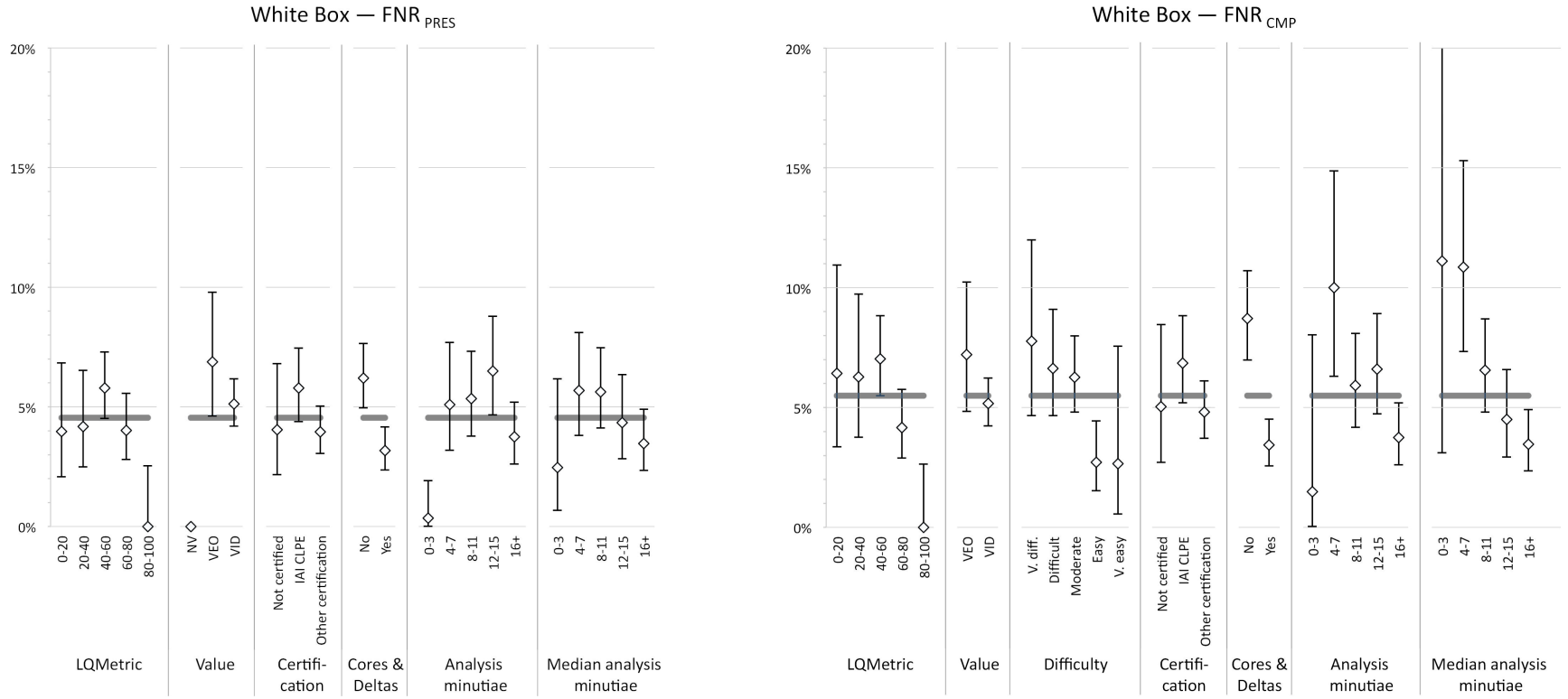

Fig. S29: Comparison of WB FNR_{PRES} and FNR_{CMP} for factors. Horizontal lines indicate overall mean rates: FNR_{PRES}=4.5% (3730 presentations), FNR_{CMP}=5.5% (2966 comparisons).

Appendix SI-15 Negative predictive value

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

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

where

NonmatePrevalence is the percentage of all comparisons that were performed on nonmated pairs,

MatePrevalence is the percentage of all comparisons that were performed on mated pairs,

TNR = Count of excluded nonmates / Count of nonmate comparisons, and

FNR = Count of excluded mates / Count of mate comparisons.

If comparisons are performed in a context where nonmated pairs are more common, true negatives will be relatively more common and NPV will higher. Conversely, if mated pairs are more common, erroneous exclusions will be relatively more common and NPV will lower.

When reporting NPV as a response to an independent variable, such as LQMetric, the mating prevalence and exclusion rates can be calculated for each level of the independent variable. However, mating prevalence is a confounder of relations between NPV and other response variables. For example, we have found that examiners tend to rate inconclusive comparisons as more difficult than individualizations and exclusions, so if data selection resulted in a greater proportion of mated pairs being inconclusive than nonmated pairs, then we would expect the proportion of difficult comparisons that are mated to be greater than the proportion of easy comparisons that are mated. The implication of this confounding is that we do not necessarily have a suitable measured value for mating prevalence for use in the above formula without introducing simplifying assumptions. This section presents unadjusted NPV results plotted against mating prevalence, where mating prevalence may be based on simplifying assumptions. The same choice of mating prevalence shown in these plots was used to project NPV to 50% mates. The choice of 50% mates is arbitrary, but has the advantages of being close to the actual test proportions (resulting in limited distortion of the actual measurements) and a suitable choice for cross-study comparisons.

Fig. S30 through Fig. S33 report unadjusted NPV measures for each level of the factor (LQMetric, latent value, core or delta, difficulty) in the context of the actual mating proportions for that factor level and the projected overall NPV for the test. These charts show that the accuracy of examiners' exclusions improves with latent image quality as measured by LQMetric and with comparison difficulty as rated by the examiner. The adjusted NPV_{50} measures are shown in Table S24 and Table S26.

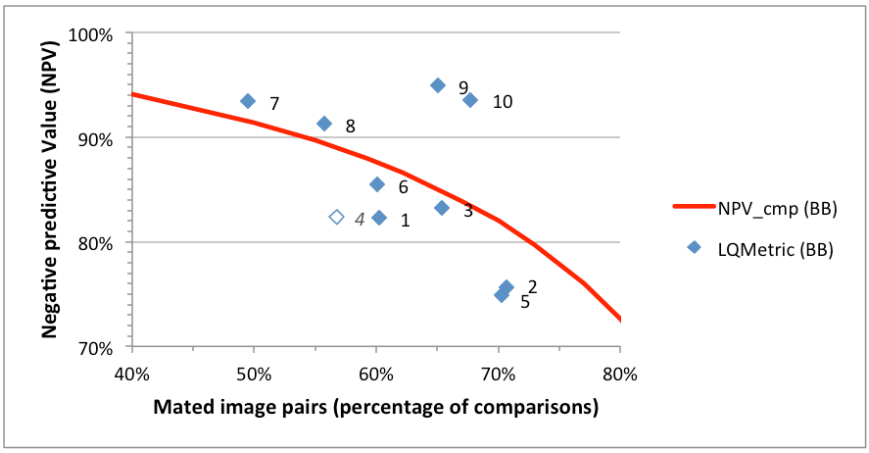

Fig. S30: Latent image quality as measured by LQMetric as a predictor of NPV (BB, n=4558 exclusions). LQMetric values reported by score intervals, labeled from "1" [0-10] to "10" [90-100]; the fourth interval (open marker) is based on only 1% of the data (111 comparisons of 8 latents resulting in 3 mate exclusions and 14 nonmate exclusions).

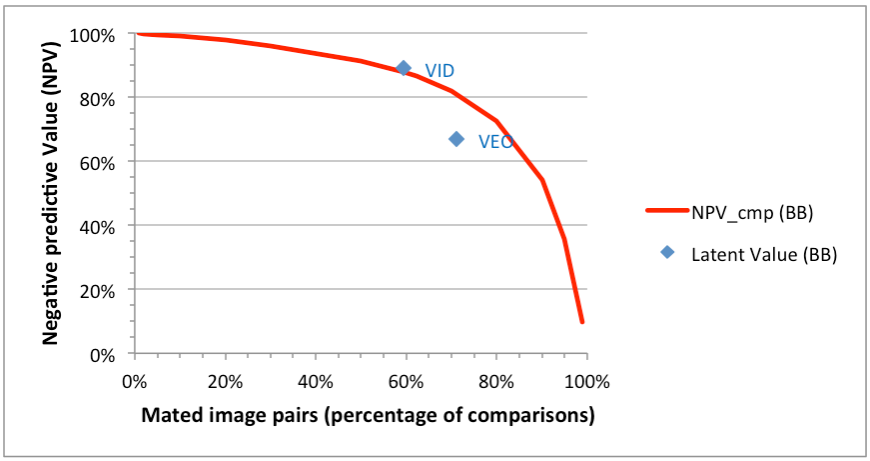

Fig. S31: Examiners' latent value determinations as a predictor of NPV (BB, n=486 exclusions on VEOs; 4072 on VIDs). The percentage of image pairs that are mated is calculated as the proportion of responses at each value level that were made on mated pairs: each image pair can contribute to multiple levels.

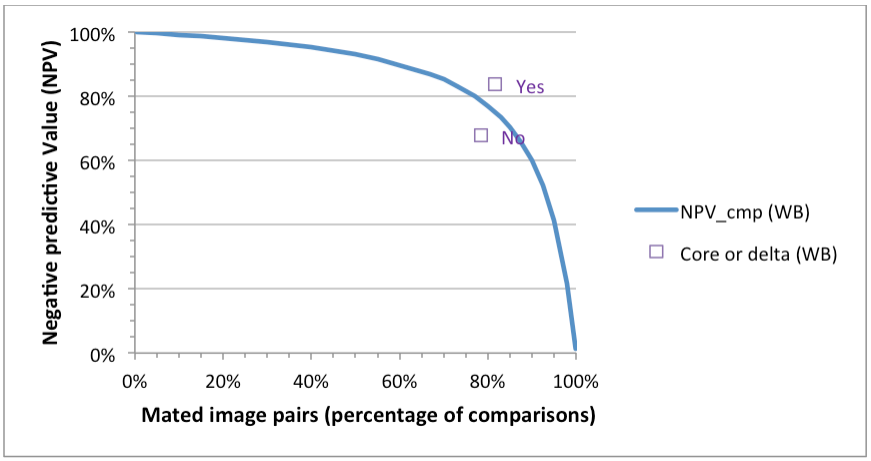

Fig. S32: Presence of a corresponding core or delta as a predictor of NPV (WB, n=561 exclusions). Image pairs were classified according to whether corresponding cores or deltas were determined to be present during a preliminary screening process. A corresponding core or delta was present on 126/231 mated pairs and 46/89 nonmated pairs.

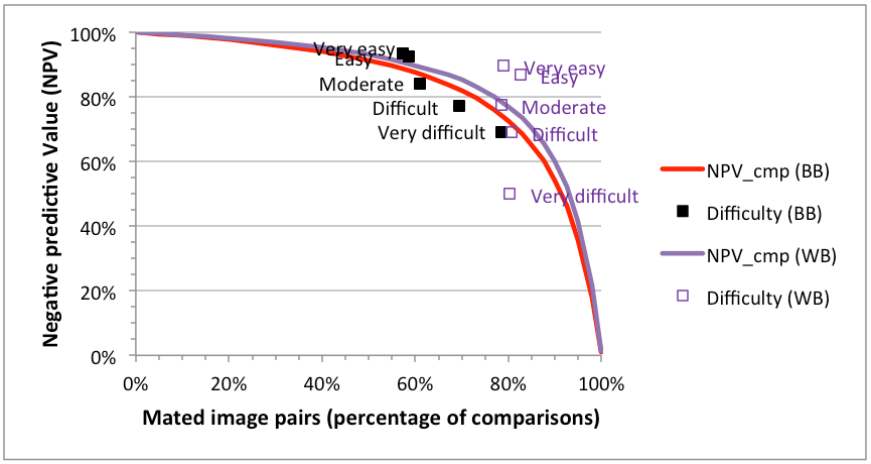

Fig. S33: Comparison difficulty as a predictor of NPV (BB, n=4558 exclusions; WB, n=561 exclusions). The percentage of image pairs that are mated is calculated as the proportion of responses at each difficulty level that were made on mated pairs: each image pair can (and typically does) contribute to multiple levels.

Fig. S34 shows interactions between latent value assessments and LQMetric, as factors predictive of NPV. NPV was much lower among VEO comparisons than VID comparisons (Fig. S31), except among those latents with high LQMetric values (Fig. S30).

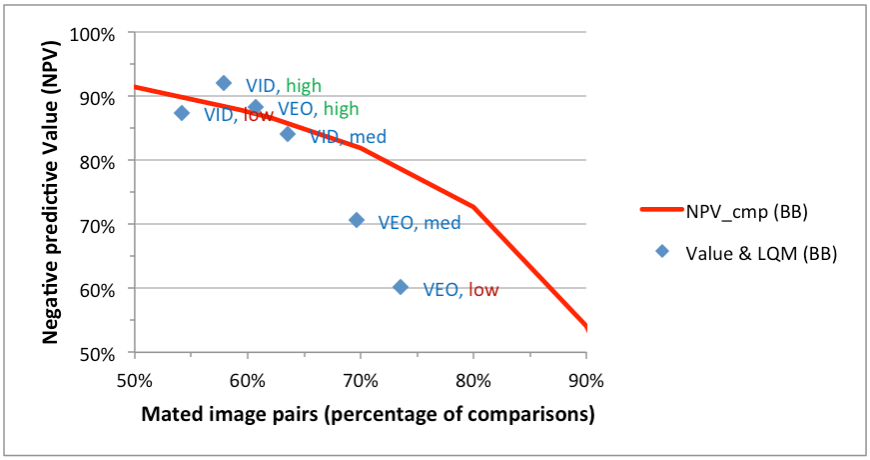

Fig. S34: Latent value determination and LQMetric as combined predictor of NPV (BB, n=4558 exclusions). LQMetric is summarized by tertile (low, medium, high).

Appendix SI-16 Supplemental Information References

1. SWGFAST (2013) Standards for Examining Friction Ridge Impressions and Resulting Conclusions, Version 2.0. http://www.swgfast.org/documents/examinations-conclusions/130427_Examinations-Conclusions_2.0.pdf
2. National Institute of Standards (2011) American National Standard for Information Systems: Data format for the interchange of fingerprint, facial & other biometric information. ANSI/NIST-ITL 1-2011. http://fingerprint.nist.gov/standard
3. SWGFAST (2012) Individualization / Identification Position Statement, Version 1.0. http://swgfast.org/Comments-Positions/120306_Individualization-Identification.pdf
4. SWGFAST (2011) Standard terminology of friction ridge examination, Version 3.0. http://swgfast.org/documents/terminology/110323_Standard-Terminology_3.0.pdf
5. Federal Bureau of Investigation; Universal Latent Workstation (ULW) Software. https://www.fbibiospecs.org/Latent/LatentPrintServices.aspx
6. Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2011) Accuracy and reliability of forensic latent fingerprint decisions. Proc Natl Acad Sci USA 108(19): 7733-7738. http://www.pnas.org/content/108/19/7733.full.pdf
7. Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2012), Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners. PLoS ONE 7:3. http://www.plosone.org/article/info:doi/10.1371/journal.pone.0032800
8. Ulery BT, Hicklin RA, Roberts MA, Buscaglia J (2014). Measuring what latent fingerprint examiners consider sufficient information for individualization determinations. PLoS ONE, 9(11), e110179. http://dx.doi.org/10.1371/journal.pone.0110179
9. Thompson MB, Tangen JM, McCarthy DJ (2013). Human matching performance of genuine crime scene latent fingerprints. Law and Human Behavior, 38(1), 84–93. http://dx.doi.org/10.1037/lhb0000051
10. Langenburg G (2012). A critical analysis and study of the ACE-V process (unpublished doctoral dissertation). Université de Lausanne, Lausanne. http://www.unil.ch/files/live//sites/esc/files/shared/Langenburg_Thesis_Critical_Analysis_of_ACE-V_2012.pdf
11. Burgman MA, et al. (2011). Expert status and performance. PLoS One, 6(7), e22998. http://dx.plos.org/10.1371/journal.pone.0022998
12. Ray E, Dechant PJ (2013) Sufficiency and Standards for Exclusion Decisions, Journal of Forensic Identification, 63 (6): 675–697.