Forensic Science International 275 (2017) 65–75

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Factors associated with latent fingerprint exclusion determinations

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

[a] Noblis, Reston, VA, USA
[b] Latent Print Support Unit, Federal Bureau of Investigation Laboratory Division, Quantico, VA, USA
[c] Counterterrorism and Forensic Science Research Unit, Federal Bureau of Investigation Laboratory Division, 2501 Investigation Parkway, Quantico, VA 22135, USA

ARTICLE INFO

Article history:
Received 6 September 2016
Received in revised form 9 February 2017
Accepted 14 February 2017
Available online 22 February 2017

Keywords:
Forensic science
Biometrics
Decision
Exclusion
Fingerprints
Quality assurance

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. Much of the literature on latent print examination focuses on individualization, and much less attention has been paid to exclusion. This experimental 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 exclusion rates include the quality of the latent, value determinations, analysis minutia count, comparison difficulty, and the presence of cores or deltas. An understanding of these associations will help explain the circumstances under which errors are more likely to occur and when determinations are less likely to be reproduced by other examiners; the results should also lead to improved effectiveness and efficiency of training and casework quality assurance. 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.

Published by Elsevier Ireland Ltd.

1. Introduction

Historically, the latent print[1] [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

* Corresponding author. Fax: +1 703 632 7801.
E-mail address: joann.buscaglia@ic.fbi.gov (J. Buscaglia).

[1] 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.

http://dx.doi.org/10.1016/j.forsciint.2017.02.011
0379-0738/ Published by Elsevier Ireland Ltd.

comparable anatomical areas are represented and legible in the known exemplars") [12].[2]

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

A substantial part of the decision process is the extraction of information from the fingerprints. The decision whether to exclude relies on a series of assessments and subsidiary decisions made by the examiner during analysis and comparison: assessing whether there are areas in the latent and exemplar that can be used to effect a meaningful comparison; assessing the presence and absence of features; assessing whether similarities should be considered correspondences; assessing whether dissimilarities should be considered discrepancies. Each of these assessments must account for uncertainty: the examiner must consider the level of confidence in each assessment. 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 exclude 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; ridge edges and pores (level 3 details) cannot be the sole factor in exclusion determinations [12]. 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., Ref. 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]

In making an exclusion decision, the examiner considers his/her assessment of similarities and dissimilarities, along with his/her level of uncertainty in this assessment, and then determines if the information is sufficient to render an exclusion. The sufficiency threshold is based on an implicit utility function [19,20], in which the examiner considers the relative benefits of making a correct exclusion versus the costs of making a mistake. 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 light of the high erroneous exclusion rate reported on Black Box and other studies [17,18], and the recent interest in exclusions [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 factors associated with erroneous exclusions. To the extent that these associations are causal, they may help to shed light on how decisions are made; however, non-causal associations may also be informative toward understanding the circumstances

[2] 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.

under which errors are more likely to occur and when determinations are less likely to be reproduced by other examiners. The objectives of this research are to explore empirically which factors most influence examiners' exclusion decisions; which are most strongly associated with reproducibility of determinations; how examiners' subjective assessments of similarities and differences vary; and the extent to which we can ascertain this information from examiners' documentation of their conclusions. 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 exclusions.

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 Refs. [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}).

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 Section 4. 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.

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.

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

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 Fig. 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 ***exclusion***, and therefore, voted results provide the best information we have available as to the sufficiency for exclusion.

Fig. 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 Section 4. 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

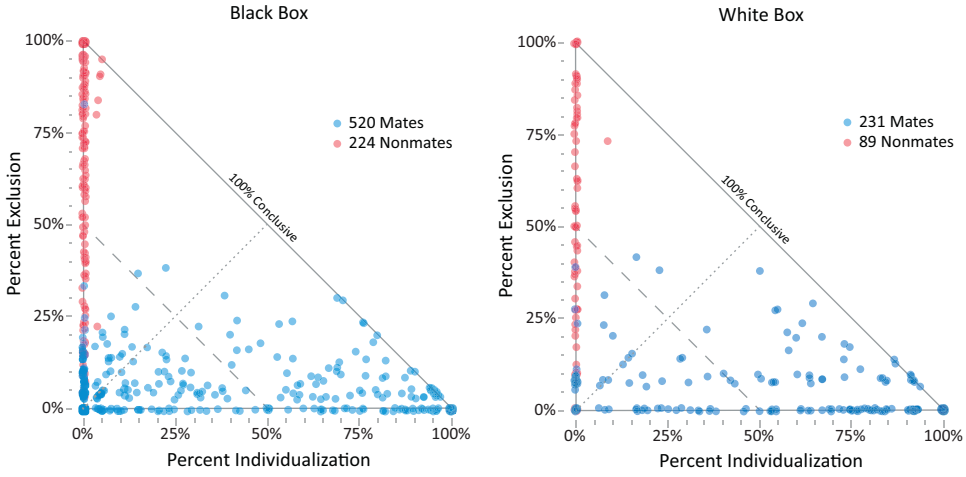

Fig. 1. Determination rates on each image pair in BB (mean of 23 examiners per image pair) and WB (mean 12 examiners per image pair). Points at the origin represent image pairs that examiners agreed unanimously could neither be excluded nor individualized; points at the bottom right were unanimous individualizations; 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 Ref. [2].

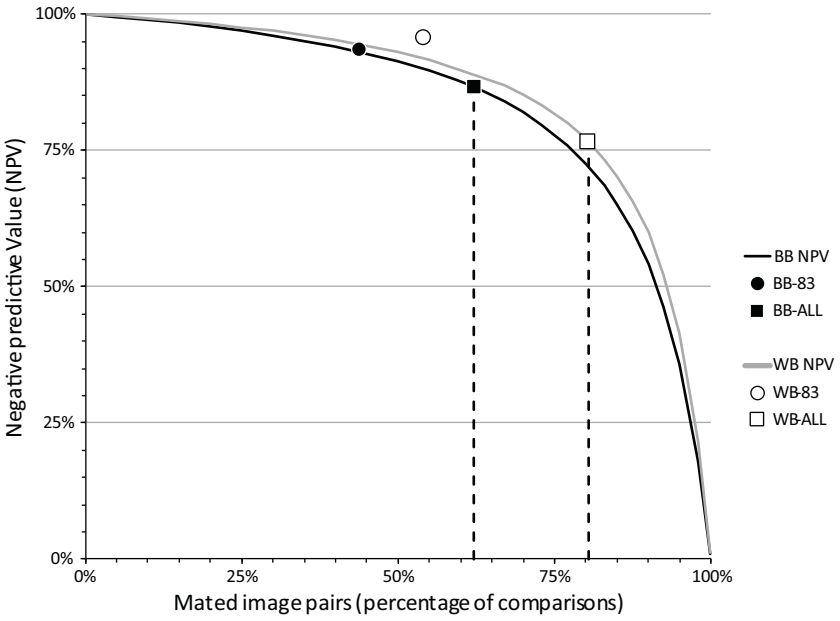

Fig. 2. NPV as measured at actual test proportions of mate and nonmate comparisons (markers), and as extrapolated as a function of the mating proportion (curves). A subset of 83 image pairs included in both tests is also indicated (which allows comparing the tests while controlling for differences in 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, $NPV_{80} = 76.6\%$.

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 Fig. 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).

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

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.

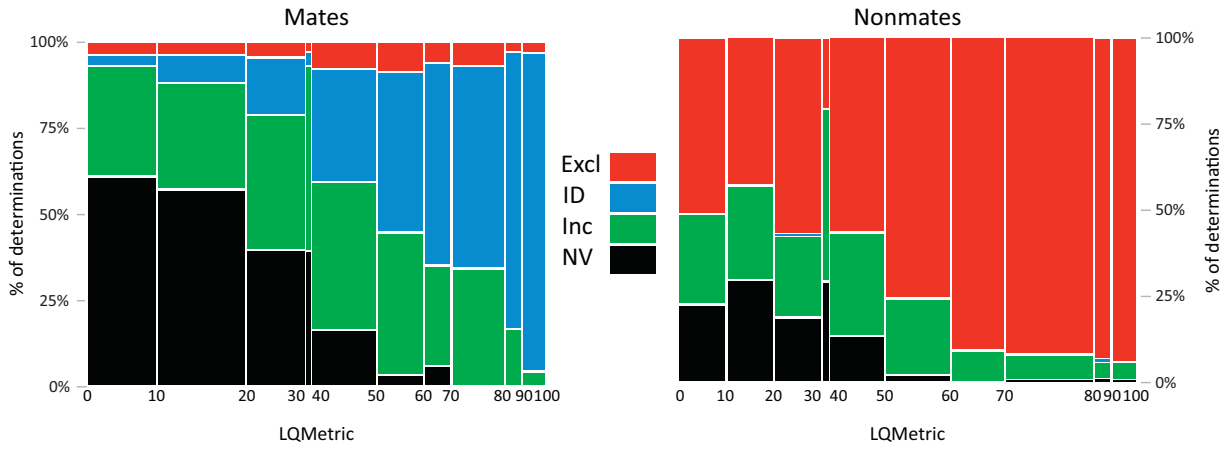

Fig. 3. Mosaic plots of the distribution of determinations by LQMetric, for mates and nonmates (BB, n = 11,578 mated and 5543 unmated image pair presentations).

The FBI's Latent Quality Metric (LQMetric)[3] 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.

Fig. 3 shows the relations between LQMetric and examinerdeterminations (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 exclude. 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}).

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 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

Fig. 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).

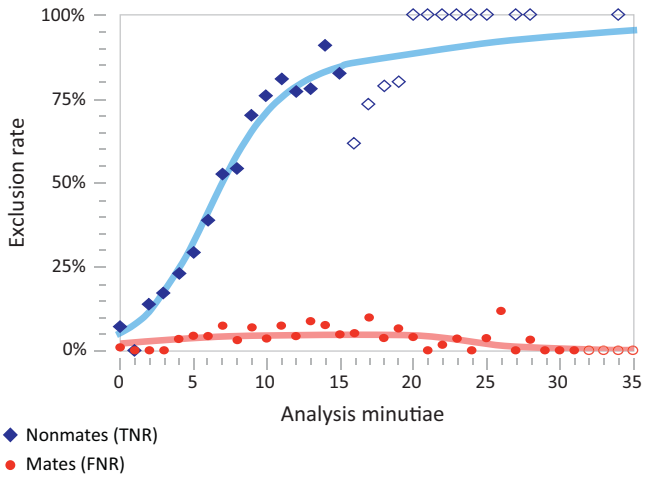

Fig. 4. True and false negative rates by the number of minutiae marked on latents during analysis (WB). Each marker represents an exclusion rate (true or false) calculated over all nonmated or mated presentations for the specified minutia count. Open markers indicate rates measured on fewer than 20 presentations. 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).

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

Table 2
Distribution of exclusion reasons. Categories are defined in Appendix SI-6.[4]

	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	

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. [23] 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.

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 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.

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 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
- 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 Ref. [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%).

[4] One WB exclusion is omitted because no exclusion reason was recorded.

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-1 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

Figs. 5 and 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).

Fig. 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.

As seen in Fig. 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 Fig. 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).

Fig. 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

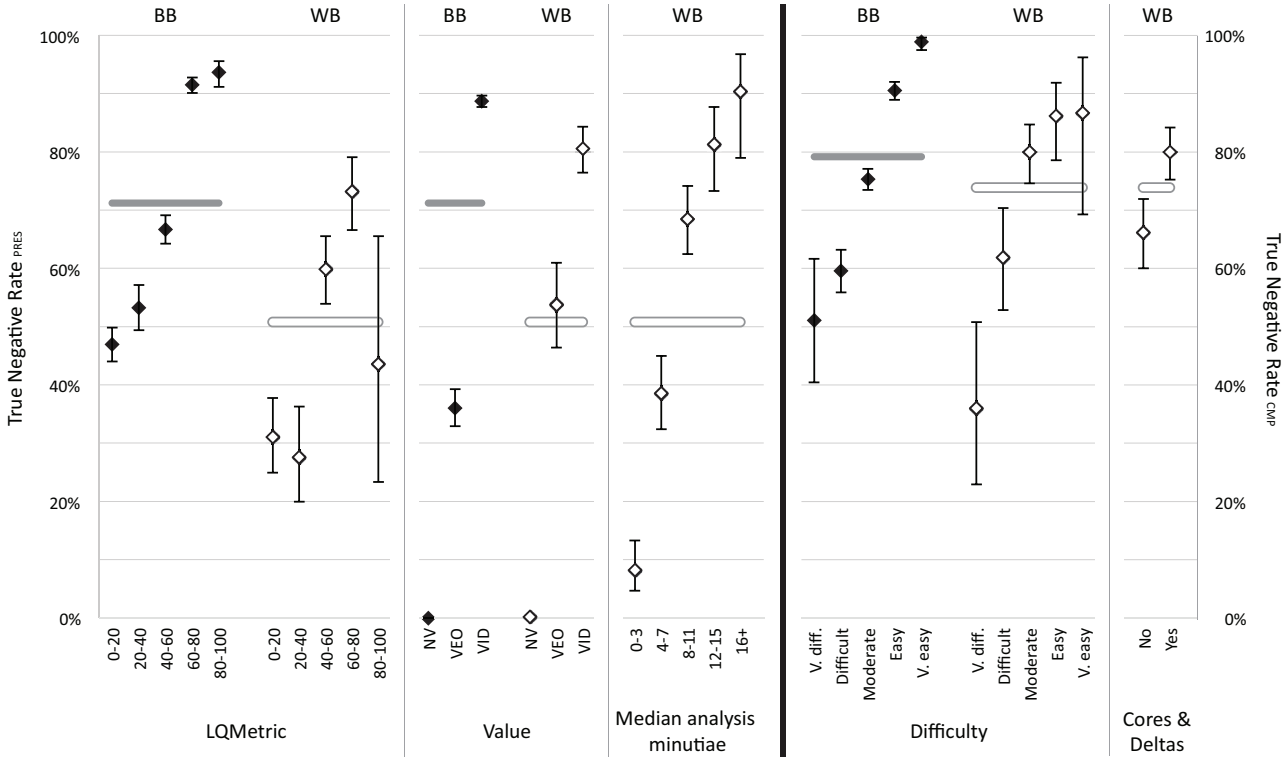

Fig. 5. Associations between TNR and factors. Vertical bars indicate 95% binomial confidence intervals. Factors measured on latents (LQMetric, value, median analysis minutiae) are based on all presentations (TNR_{PRES}); difficulty and cores & deltas are based on comparisons (TNR_{CMP}). Horizontal lines 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).

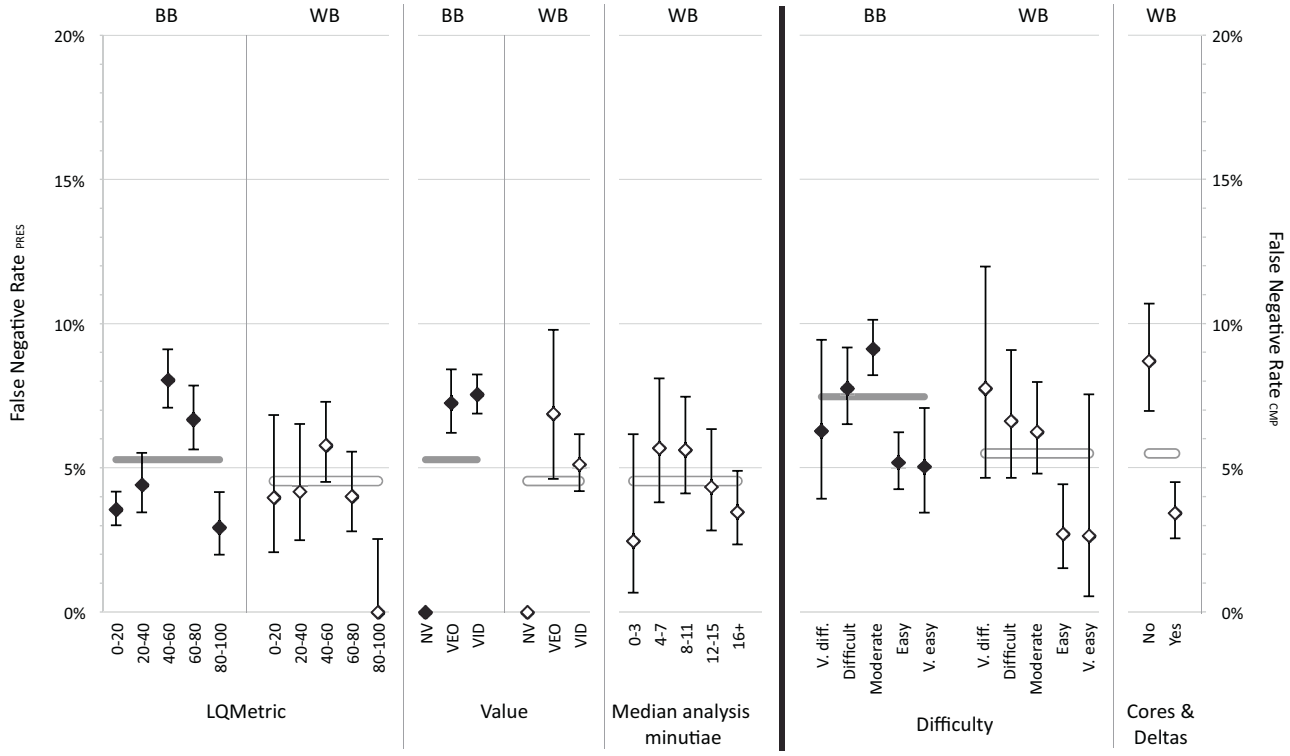

Fig. 6. Associations between FNR and factors. Vertical bars indicate 95% binomial confidence intervals. Factors measured on latents (LQMetric, value, median analysis minutiae) are based on all presentations (FNR_{PRES}); difficulty and cores & deltas are based on comparisons (FNR_{CMP}). Horizontal lines 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).

difficulty level; nevertheless, we found that NPV decreased substantially with difficulty. Additional NPV data is presented in Appendix SI-13 and Appendix SI-15.

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

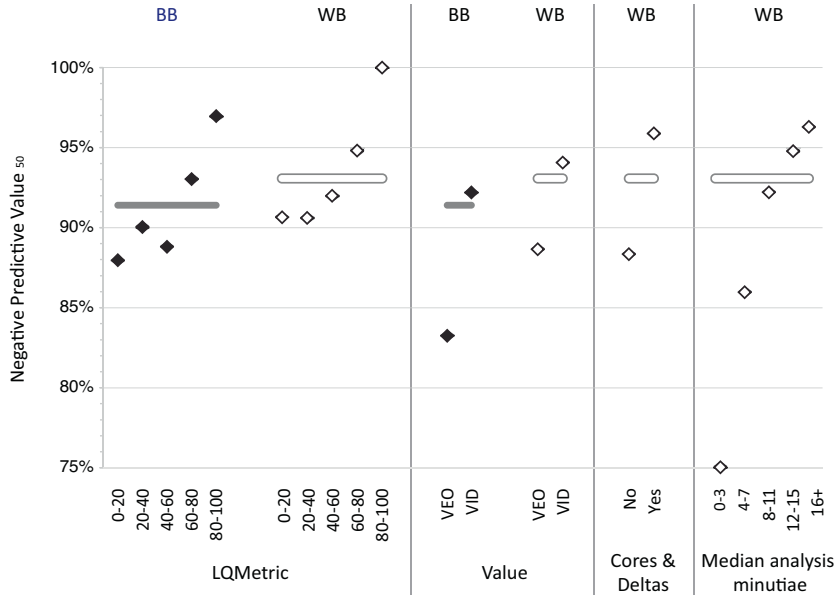

Fig. 7. Associations between NPV_{50} and factors. Horizontal lines indicate overall NPV: Black Box NPV_{50} = 91.4% (n = 4558 exclusions); White Box NPV_{50} = 93.1% (n = 561 exclusions). Confidence intervals were not estimated for lack of a standard method and because of debatable modeling assumptions.

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.,Ref. 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. Discussion

As discussed by Ray and Dechant [16] and Champod et al. [20], relatively little attention has been paid to exclusions as compared to individualizations. The empirical data we have presented describes how exclusions are related to various attributes of latent prints. This information can be used to focus training and proficiency testing, to interpret results that may appear to differ across studies, and to guide the sampling of fingerprints for use in future experimental designs.

Our findings suggest ways to improve training and thus the performance of individual examiners. Although this research focuses on the performance of practicing examiners, with emphasis on their errors and disagreements, it is important to step back and consider the contexts in which those examiners work: many issues related to exclusion arise from a lack of consensus in the community. 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 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 confuse exclusions and non-identifications. Standardization of exclusion terminology, policies, and procedures is needed.

There is no generally-accepted method for documenting the basis for exclusion. The lack of such a standard method contributed to participants not consistently providing the detailed information needed to evaluate the extent of support for exclusions, corroborating the findings of Neumann, et al. [18]. We previously found sufficiency and reproducibility of individualizations to be strongly associated with measures of the number of corresponding minutiae, which can readily be annotated and evaluated (e.g., Ref. [6]); we have nothing analogous to corresponding minutiae to quantify dissimilarities when making exclusions. We assume that limited documentation has an adverse effect on quality assurance, potentially making it difficult to detect questionable exclusions and impeding the verification of difficult decisions. If the reason for exclusion is that the pattern classes differ, detailed markup may not be necessary; otherwise, marking of discrepancies and other reference points is often needed to commmunicate the basis for an exclusion.

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.

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.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.forsciint.2017.02.011.

References

[1] R.A. Hicklin, et al., Latent fingerprint quality: a survey of examiners, J. Forensic Identif. 61 (4) (2011) 385–419.
[2] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Accuracy and reliability of forensic latent fingerprint decisions, Proc. Natl. Acad. Sci. U. S. A. 108 (19) (2011) 7733–7738, doi:http://dx.doi.org/10.1073/pnas.1018707108.
[3] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Repeatability and reproducibility of decisions by latent fingerprint examiners, PLoS One 7 (3) (2012) e32800, doi:http://dx.doi.org/10.1371/journal.pone.0032800.
[4] R.A. Hicklin, J. Buscaglia, M.A. Roberts, Assessing the clarity of friction ridge impressions, Forensic Sci. Int. 226 (1) (2013) 106–117, doi:http://dx.doi.org/10.1016/j.forsciint.2012.12.015.
[5] B.T. Ulery, R.A. Hicklin, G.I. Kiebuzinski, M.A. Roberts, J. Buscaglia, Understanding the sufficiency of information for latent fingerprint value determinations, Forensic Sci. Int. 230 (1) (2013) 99–106, doi:http://dx.doi.org/10.1016/j.forsciint.2013.01.012.
[6] B.T. Ulery, R.A. Hicklin, M.A. Roberts, J. Buscaglia, Measuring what latent fingerprint examiners consider sufficient information for individualization determinations, PLoS One 9 (11) (2014) e110179, doi:http://dx.doi.org/10.1371/journal.pone.0110179.
[7] B.T. Ulery, R.A. Hicklin, M.A. Roberts, J. Buscaglia, Changes in latent fingerprint examiners' markup between analysis and comparison, Forensic Sci. Int. 247 (2014) 54–61, doi:http://dx.doi.org/10.1016/j.forsciint.2014.11.021.
[8] N.D. Kalka, R.A. Hicklin, On relative distortion in fingerprint comparison, Forensic Sci Int. 244 (2014) 78–84, doi:http://dx.doi.org/10.1016/j.forsciint.2014.08.007.
[9] B.T. Ulery, R.A. Hicklin, M.A. Roberts, J. Buscaglia, Interexaminer variation of minutia markup on latent fingerprints, Forensic Sci. Int. 264 (2016) 89–99, doi:http://dx.doi.org/10.1016/j.forsciint.2016.03.014.

[10] National Research Council, Strengthening Forensic Science in the United States: a path forward, National Academis Press, Washington, DC, 2009. https://www.ncjrs.gov/pdffiles1/nij/grants /228091.pdf.
[11] SWGFAST, Standard Terminology of Friction Ridge Examination (Latent/Tenprint Document #19) Ver. 4.1, (2013) . http://swgfast.org/documents/terminology/121124_Standard-Terminology_4.0.pdf.
[12] SWGFAST, Standards for Examining Friction Ridge Impressions and Resulting Conclusions (Latent/Tenprint Document #10), (2013) . http://www. swgfast.org/documents/examinations-conclusions/130427_Examinations-Conclusions_2.0.pdf.
[13] M. Triplett, The Need to Validate Principles and the Value of Reproducible Results, IDentification News, 2012, pp. 42–43 August.
[14] J.I. Thornton, One-dissimilarity doctrine in fingerprint identification, Int. Crim. Police Rev. 306 (1977) 89–95.
[15] Expert Working Group on Human Factors in Latent Print Analysis, Latent Print Examination and Human Factors: Improving the Practice Through a Systems Approach, U.S. Department of Commerce, National Institute of Standards and Technology, 2012. http://www.nist.gov/oles/upload/latent.pdf.
[16] E. Ray, P.J. Dechant, Sufficiency and standards for exclusion decisions, J. Forensic Identif. 63 (6) (2013) 675–697.
[17] G. Langenburg, 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, J. Forensic Identif. 59 (2) (2009) 219–257.
[18] C. Neumann, C. Champod, M. Yoo, T. Genessay, G. Langenburg, Improving the Understanding and the Reliability of the Concept of "Sufficiency" in Friction Ridge Examination, National Institute of Justice, Washington DC, 2013. https://www.ncjrs.gov/pdffiles1/nij/grants/244231.pd.
[19] A. Biedermann, Decision theoretic properties of forensic identification: underlying logic and argumentative implications, Forensic Sci. Int. 177 (2008) 120–132.
[20] C. Champod, C.J. Lennard, P. Margot, M. Stoilovic, Fingerprints and Other Ridge Skin Impressions, 2nd edition, CRC Press, 2016.
[21] R.A. Huber, Expert witness, Crim. Law Q. 2 (1959) 276–296.
[22] D. Ashbaugh, Quantitative-qualitative Friction Ridge Analysis: An Introduction to Basic and Advanced Ridgeology, CRC Press, New York, 1999.
[23] I. Pacheco, B. Cerchiai, S. Stoiloff, Miami-Dade Research Study for the Reliability of the ACE-V Process: Accuracy & Precision in Latent Fingerprint Examinations, (2014) . https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=270637.
[24] G. Langenburg, A Critical Analysis and Study of the ACE-V Process (Unpublished Doctoral Dissertation), Université de Lausanne, Lausanne, 2012. http://www.unil.ch/files/live//sites/esc/files/shared/Langenburg_Thesis_Critical_Analysis_of_ACE-V_2012.pdf.
[25] C.J. Stone, C.Y. Koo, Additive splines in statistics, Proc. Stat. Comp. Sect. Am. Statist. Assoc. 27 (1985) 45–48.