Preprint of article subsequently published:
Ulery BT, Hicklin RA, Roberts MA, Buscaglia J (2014) Changes in latent fingerprint examiners' markup between analysis and comparison. Forensic Science International 247:54-61. (http://dx.doi.org/10.1016/j.forsciint.2014.11.021)

Changes in latent fingerprint examiners' markup between Analysis and Comparison

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

After the initial Analysis of a latent print, an examiner will sometimes revise the assessment during Comparison with an exemplar. Changes between Analysis and Comparison may indicate that the initial Analysis of the latent was inadequate, or that confirmation bias may have affected the Comparison. 170 volunteer latent print examiners, each randomly assigned 22 pairs of prints from a pool of 320 total pairs, provided detailed markup documenting their interpretations of the prints and the bases for their Comparison conclusions. We describe changes in value assessments and markup of features and clarity. When examiners individualized, they almost always added or deleted minutiae (90.3% of individualizations); every examiner revised at least some markups. For inconclusive and exclusion determinations, changes were less common, and features were added more frequently when the image pair was mated (same source). Even when individualizations were based on eight or fewer corresponding minutiae, in most cases some of those minutiae had been added during Comparison. One erroneous individualization was observed: the markup changes were notably extreme, and almost all of the corresponding minutiae had been added during Comparison. Latents assessed to be of value for exclusion only (VEO) during Analysis were often individualized when compared to a mated exemplar (26%); in our previous work, where examiners were not required to provide markup of features, VEO individualizations were much less common (1.8%).

Keywords: Forensic science, Biometrics, Decision, Identification, Fingerprints, Quality assurance

Introduction

As one part of casework, a latent[1] print examiner compares a latent (friction ridge impression from an unknown subject) to an exemplar (print deliberately collected from a known source) in a process known as

[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*

ACE: Analysis, Comparison, and Evaluation [1]. In Analysis,[2] the examiner assesses the quantity, quality, and distinctiveness of the latent's features and determines whether the latent is of sufficient value for comparison with an exemplar. In Comparison, the examiner determines the extent of corresponding or contradictory information; this serves as the basis for the Evaluation determination of identification, exclusion, or inconclusive. Here, we report the results of a large-scale study describing how examiners' markup of features, clarity, and value made during Analysis of a latent were changed during Comparison with an exemplar.

Some agencies and researchers recommend a "linear ACE" procedure [2, 3, 4], in which "examiners must complete and document analysis of the latent fingerprint before looking at any known fingerprint" and "must separately document any data relied upon during comparison or evaluation that differs from the information relied upon during analysis" [2]. Others argue that a recurring, reversible and blending ACE model is preferable [5]. The rationale for linear ACE is based on concerns regarding circular reasoning [6,7,8,9]. When comparing prints, after establishing anchor points of potential similarity, an examiner looks back and forth between the latent and exemplar to assess the similarity of each potentially corresponding region. This process has been described as involving "forward" comparison (from the latent to the exemplar) and "reverse" comparison (from the exemplar to the latent) [8,10]. Reverse comparison in search of potential discrepancies is necessary, but the examiner needs to be alert to the risk of confirmation bias or circular reasoning. Once an examiner sees a possible alignment between the latent and exemplar, the process of following individual ridges and marking correspondences may result in changes to the latent markup simply because more time and effort is expended, and because additional features may be suggested based on the similarity; when the examiner is unable to find any potentially corresponding areas, there is less basis for such revisions. A notable example of the problem of bias from the exemplar resulting in circular reasoning occurred in the Madrid misidentification, in which the initial examiner reinterpreted five of the original seven Analysis points to be more consistent with the (incorrect) exemplar: "Having found as many as 10 points of unusual similarity, the FBI examiners began to 'find' additional features in LFP 17 [the latent print] that were not really there, but rather suggested to the examiners by features in the Mayfield prints"[8].

There is a general lack of formal guidelines for ACE procedures and documentation. The Scientific Working Group on Friction Ridge Analysis, Study and Technology's (SWGFAST's) *Standard for the Documentation of ACE-V* directs examiners to document both the Analysis of a latent and any "re-analysis" of the latent that occurs during the Comparison phase "such that another qualified examiner can determine what was done and interpret the data" [1]. That said, the details of how to document Analysis and Comparison are mostly unspecified, and SWGFAST's standards are unenforced, leaving the details to be sorted out by agency standard operating procedures or by the examiners' judgments. In the past, detailed documentation of Analysis was often limited to that required for searches of Automated Fingerprint Identification Systems (AFIS), with instructions on which features to mark varying substantially by vendor. Agencies vary on whether they require detailed markup to document the features of a latent in Analysis, and whether they require markup of features in Comparison. Those agencies that do require markup vary substantially on how that markup is effected, including pinpricks in physical photographs, color-coding approaches (e.g., GYRO [11]), and a variety of software-based solutions (e.g., the FBI's Universal Latent Workstation (ULW) [12], Mideo Latentworks® [13], PiAnOS [14], and Adobe® Photoshop®). Our research uses the Extended Feature Set (EFS) format as defined in the *ANSI/NIST-ITL* standard [15] and supporting guidelines for examiners [16], which attempt to standardize the syntax and semantics of markup. However, although EFS is broadly used as a non-proprietary format for searches of an AFIS, it is not yet frequently used for markup of non-AFIS casework. Because documentation is not standardized in practice, the extent to which examiners revise their markup during Comparison is difficult to ascertain, either in casework or in research.

There is not extensive previous work regarding how examiners revise their Analysis assessments of a latent during Comparison. Dror et al [17] discussed how the presence of a exemplar affected the features observed in the latent: the number of minutiae marked in the presence of a mated exemplar was generally greater than

standard terminology to maintain consistency with our previous and future papers in this series [19,20,21,22,23,24]. See Glossary, Appendix A.1.

[2] In this report, we capitalize Analysis and Comparison to indicate that we are referring to the ACE phases. For brevity, we use Comparison to refer to the Comparison/Evaluation phase.

the number marked when the exemplar was not present; much of this effect was limited to a subset of the latents. Evett and Williams [18] describe how UK examiners working under a 16 point standard used the exemplar to "tease the points out" of the latent after reaching an "inner conviction." Neumann, et al. [19], in a discussion of interexaminer reproducibility, provide examples showing changes to minutiae markup made during Comparison.

This study was conducted to better understand the latent print examination process as currently practiced, in order to identify potential areas for improvement. This work builds on our previous studies of the latent print examination process, which have included analyses of fingerprint quality [20,21], analyses of how examiners make value determinations [22], analyses of the accuracy and reliability of examiners' analysis and comparison determinations [23, 24], and analyses of the sufficiency of information for individualizations [25]. In this study we assess how the examiner's assessment of a latent print changes when the examiner compares the latent with a possible mate. We describe changes in feature markup, clarity markup and value assessments between the Analysis and Comparison phases of ACE:

- How pervasive were changes in latent print markup and value assessments?
- How were changes in latent markup associated with the comparison conclusion reached by the examiner, the examiners' ratings of comparison difficulty, and the examiner's clarity markup?
- Were changes in latent markup affected by whether the comparison was (unbeknownst to the examiner) to a mated or nonmated exemplar? How were changes in latent markup associated with low-minutia-count individualizations?
- How were changes in latent value assessments associated with changes in markup?

Materials and Methods

This paper presents analyses of portions of the data collected in the "White Box" study, in which practicing latent print examiners annotated features, clarity, and correspondences in latent and exemplar fingerprints to document what they saw when performing examinations. The White Box study and its results with respect to sufficiency for individualization are described in [24]; the test procedure, participants, and fingerprint data are summarized in Appendix A.

The test procedure was designed to correspond to that part of casework in which an examiner compares a single latent to a single exemplar print (latent-exemplar image pair). The test software enforced a linear ACE workflow. In the Analysis phase, only the latent was presented, and the examiners provided the following markup: local clarity map (produced by "painting" the image using six colors denoting defined levels of clarity [15,21]); locations of features; types of features (minutiae, cores, deltas, and "other" points (nonminutia features such as incipient ridges, ridge edge features, or pores); and value determination (of value for individualization (VID), of value for exclusion only (VEO)[3], or no value (NV)). If the latent print was determined to be VEO or VID, the exemplar was presented for side-by-side comparison with the latent. During this combined Comparison/Evaluation phase, the examiner annotated the exemplar (clarity and features) and assessed its value (VID, VEO, NV), optionally revised the latent markup and value determination, further annotated the pair of images to indicate corresponding and discrepant features, reported the comparison determination (individualization, exclusion, or inconclusive), and assessed comparison difficulty (very easy, easy, moderate, difficult, very difficult). Any modifications of the markup and value determination for the latent after the exemplar was presented were recorded, thus enabling this study.

The fingerprint markup and value determinations complied with EFS, which is an international latent fingerprint data exchange standard [15]; the test instructions were derived from [16], which proposes standard instructions for the markup of latent prints. Any such study can only partly correspond to actual casework across multiple agencies, since operational procedures vary among agencies on documentation of latent print examination and on how latent value is assessed (Appendix A.2). The software application used for our experiment is a variant of the FBI's ULW Comparison Tool [12], which is widely used for operational

[3] *In the software, this option was abbreviated as "Limited value," which in the text and video instructions was clearly defined as being of value for exclusion only.*

casework by local, state, and federal agencies in the United States, as well as by some international agencies. Participants were instructed in the test objectives, procedures, and software usage through a short video, a detailed instruction document, and practice exercises.

During the Comparison phase, examiners moved or deleted some of the features marked during Analysis, and marked additional features. For each pair of latent markups (Analysis and Comparison phases), we classify features as ***retained***, ***moved***, ***deleted***, or ***added***. A retained feature is one that is present at exactly the same pixel location in both markups; a moved feature refers to one that was deleted during Comparison and replaced by another within 0.5 mm (20 pixels at 1000 ppi, approximately one ridge width); a deleted feature is one that was present in the Analysis markup only (no Comparison feature within 0.5 mm); an added feature is one that was present in the Comparison markup only (no Analysis feature within 0.5mm). We discuss the effectiveness of this classification approach in Appendix A.5.

We generally report clarity results by aggregating the six levels specified by the examiners (described in Appendix A.4) into two levels: Clear and Unclear. Clear areas (painted by the examiners as green, blue, or aqua) are those where the examiner can follow individual friction ridges and is certain of the location, presence and absence of all minutiae. Unclear areas (painted as yellow, red, or black) include background as well as areas where the examiner was confident in the continuity of ridge flow, but any minutiae were at best debatable.

Participation was open to practicing latent print examiners and included a broad cross-section of the fingerprint community. A total of 170 latent print examiners participated: 33% were Certified Latent Print Examiners (an International Association for Identification certification); an additional 56% had other certifications or qualifications as latent print examiners, generally by their employers or non-US national accreditations; 82% were from the United States. Participant survey responses are summarized in [24, Appendix S10].

The fingerprints for the study included prints collected under controlled conditions, and prints from operational casework (described in Appendix A.3). The fingerprint pairs were selected to vary broadly over a four-dimensional design space: number of corresponding minutiae, image clarity, presence or absence of corresponding cores and deltas, and complexity (based on distortion, background, or processing). The test dataset included 320 image pairs (231 mated and 89 nonmated), constructed from 301 latents and 319 exemplars.

Each examiner was assigned 17 mated image pairs and 5 nonmated image pairs; these proportions were not revealed to participants. We received a total of 3,740 responses [24]. Our analyses of changes in value determinations are limited to a subset of 3,709 responses, which omits 10 responses with invalid images and 21 responses with incomplete data due to software problems. Our analyses of markup changes are limited to 2,957 comparisons of 313 image pairs, which also omits 703 NV responses that did not proceed to Comparison, 43 latents changed to NV during Comparison, and 6 exemplar NV determinations made during Comparison (Table 1).

Latent Value		**Inconclusive**		**Exclusion**		**Individualization**		**No conclusion**		**Total**		
Analysis	*Comparison*	*Mate*	*Nonmate*	*Mate*	*Nonmate*	*Mate*	*Nonmate*	*Mate*	*Nonmate*	*Mate*	*Nonmate*	*Overall*
NV								457	246	457	246	703
VEO	NV							15	8	15	8	23
VID	NV							14	6	14	6	20
VEO	VEO	251	78	25	97			3		279	175	454
VEO	VID	5		2	4	103				110	4	114
VID	VEO	22	7	3	4					25	11	36
VID	VID	276	65	100	322	1,592	1	3		1,971	388	2,359
Subtotal (conclusions)		554	150	130	427	1,695	1			2,379	578	2,957
Total		554	150	130	427	1,695	1	492	260	2,871	838	3,709

Table 1. Summary of responses, showing associations between changes in latent value determinations and Comparison conclusions. Changed value determinations are highlighted. "No conclusion" indicates that either the exemplar or latent was NV.

Results

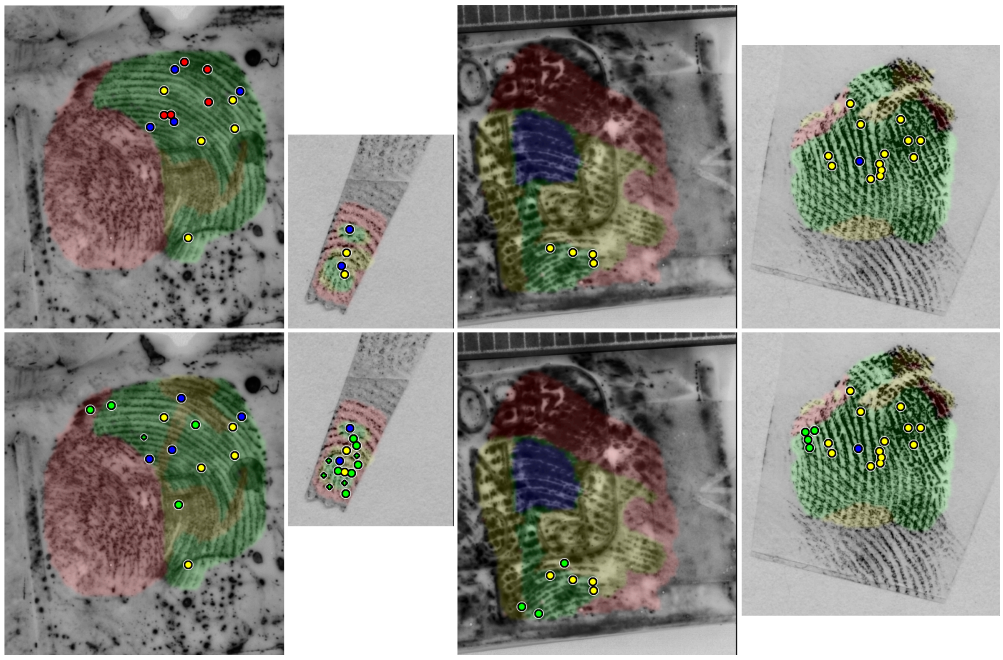


Figure 1: Example markups of four latents from Analysis (top) and from Comparison (bottom). Retained features are in yellow, moved in blue, deleted in red, and added in green. Other (non-minutiae) features are shown as crosses. These latents are shown without markup with their mated exemplars in Appendix A.3.

Figure 1 shows examples of changes between Analysis and Comparison. Table 2 shows an overview of the changes in markup by feature type (details in Appendix B.1). The rates of change were similar for minutiae, cores, and deltas, but notably higher for other features. A high rate of added "other" features was expected because the marking of such features was optional during Analysis [24, Appendix S22], but necessary for features that they used as the basis for Comparison determinations. Most of the features marked were minutiae; this study focuses primarily on changes in minutia markup.

	Number of features		% of Analysis features			
	Analysis	Comparison	Retained	Moved	Deleted	Added
Minutiae	41,774	46,083	87%	6%	7%	17%
Cores	1,079	1,174	88%	4%	7%	16%
Deltas	512	567	86%	5%	8%	19%
Other features	378	595	86%	2%	12%	70%
Changed or unknown type	213	216	46%	54%	0%	1%
Total	43,956	48,635	87%	6%	7%	18%

Table 2: Feature changes by feature type (n=2,957 comparisons). The features marked in Analysis are categorized as Retained, Moved, or Deleted (which collectively add to 100%). Features Added in Comparison are reported as a percentage increase over the number marked during Analysis (e.g., the number of minutiae added during Comparison amounted to a 17% increase from 41,774.)

After the completion of the test, a panel of examiners reviewed and discussed a small sample of the participants' responses (including some that were randomly selected, and some with unusually extensive changes). They interpreted the majority of the modifications as appearing to be reasonable reinterpretations from the perspective of the examiner who made the changes (as opposed to miscommunication related to careless markup, failure to follow instructions, software issues, etc.). Potential explanations for these reinterpretations included 1) marking details in Comparison that were seen during Analysis but deemed not

worth marking (e.g., level-2 features within deltas, level-3 features); and 2) understanding subtleties of features based on how they appear in the exemplar (e.g., moving the location of a minutia, marking points that were seen in Analysis but were too Unclear to mark). Some of the changes were more disconcerting, including 3) substantial changes compensating for inadequate Analysis; and 4) (occasionally) marking features in the latent that could not have been detected without use of the exemplar. Based on the review, we can see that a small proportion of the modifications in this test can be considered as outliers. For example, one examiner deleted latent features whenever the determination was an exclusion; another examiner routinely deleted all Analysis markup and started feature markup anew in Comparison; occasionally examiners deleted features that were in areas that did not overlap with the exemplar (discussed in Appendix A.6). Changes to clarity tended to be minor local adjustments, excepting those of a few examiners who routinely redid their clarity markup during Comparison.

Changes in minutia markup were strongly associated with the examiners' Comparison determinations and whether the image pair was (unbeknownst to the examiner) mated or nonmated (Table 3). When examiners individualized, they almost always added or deleted minutiae (90.3% of individualizations[4]). Individualizations were associated with more moved and deleted minutiae than were other determinations, and with strikingly more added minutiae; the rate of change was notably higher for those individualized latents that were initially assessed as VEO (supporting analyses in Appendix B.10). Mated exemplars influenced markup even when the determination was inconclusive or exclusion: minutiae were added far more frequently when the image pair was mated rather than nonmated.[5] The high rates of change for individualizations and determinations on mated exemplars presumably resulted from using the exemplars to focus attention on features that were not marked during Analysis. See Appendices B.1 and B.2 for further details, including results for nonminutia features.

		Number of comparisons	% of comparisons with any added or deleted minutiae	Number of minutiae		% of Analysis minutiae			
				Analysis	Comparison	Retained	Moved	Deleted	Added
Mates	Indiv (True Positive, TP)	1,695	90.3%	28,224	31,945	85.2%	6.9%	7.9%	21.1%
	Inconclusive	544	55.8%	5,866	6,324	88.1%	5.3%	6.6%	14.4%
	Exclusion (False Negative, FN)	130	40.8%	1,645	1,735	93.4%	3.6%	3.0%	8.5%
Nonmates	Indiv (False Positive, FP)	1	100.0%	17	14	0.0%	5.9%	94.1%	76.5%
	Inconclusive	150	23.3%	1,211	1,190	90.0%	3.1%	6.9%	5.1%
	Exclusion (True Negative, TN)	427	28.1%	4,811	4,875	93.4%	3.3%	3.2%	4.6%
All mates		2,379	79.5%	35,735	40,004	86.0%	6.5%	7.5%	19.4%
All nonmates		578	27.0%	6,039	6,079	92.5%	3.3%	4.2%	4.9%
Total		2,957	69.3%	41,774	46,083	87.0%	6.0%	7.0%	17.3%

Table 3: Minutia changes by Comparison determination. Percentages based on fewer than 10 comparisons are shown in gray.

Within any given type of comparison determination, the proportion of comparisons with changed minutiae increased as difficulty increased. For example, among exclusions, the proportion of comparisons with deleted or added minutiae ranged from 14% (Very Easy) to 54% (Very Difficult); for individualizations the proportions ranged from 85% (Very Easy) to 94% (Very Difficult). For a subset of 83 image pairs used both in this study and our previous "Black Box" study [22], examiners rated exclusions and inconclusives as substantially more difficult when markup was required (in this study) than when no markup was required (additional data in Appendix B.3).

Most minutiae were marked in Clear areas. The rates of changed minutiae were higher in low-clarity areas, especially for added minutiae (Table 4). The rates of deleted and added minutiae in Clear areas were surprisingly high given that Clear areas are supposed to indicate that the examiner was certain of the location, presence, and absence of all minutiae. Examiners changed minutiae in Clear areas on 72% of the comparisons that resulted in individualizations (Appendix B.4). The concentration of changes in Unclear

[4] *93.2% of individualizations had moved, deleted, or added minutiae.*

[5] *Although distinct procedures were used to select latent fingerprints for use in mated vs. nonmated pairs (described in Appendix A.3), the general trends observed here hold true after controlling for these differences by limiting the data to latents used in both mated and nonmated pairs (Appendix B.1, Table S6).*

areas is even more pronounced when analyzed by the ***median*** clarity across multiple examiners (compare Table 4 to Appendix B.5, Table S13): for true positives (individualizations on mated pairs), minutiae in median Unclear areas were deleted at a rate of 18%, and added at a rate of 47%. The median assessment of clarity was a better predictor of changes in minutia markup than the individual examiner's subjective assessment of their own certainty.

		Number of minutiae		% of Analysis minutiae			
	Clarity	Analysis	Comparison	Retained	Moved	Deleted	Added
All comparisons (n=2,957)	Unclear	11,068	13,268	84%	7%	10%	30%
	Clear	30,706	32,815	88%	6%	6%	13%
	Total	41,774	46,083	87%	6%	7%	17%
True positives (n=1,695)	Unclear	6,646	8,436	79%	8%	12%	39%
	Clear	21,578	23,509	87%	6%	7%	16%
	Total	28,224	31,945	85%	7%	8%	21%

Table 4: Minutia changes by clarity.

Examiners modified 13% of the latent clarity maps during Comparison, with a higher rate of change for mated data (Appendix B.6). Examiners showed no general tendency toward increasing or decreasing clarity when modifying their clarity maps during Comparison. Examiners rarely changed the clarity for retained minutiae (0.9% changed between Unclear and Clear), but changes in clarity occurred much more frequently in association with moved minutiae (6.2%).

During Comparison, examiners provided markup indicating which features corresponded between the latent and exemplar; changes between Analysis and Comparison were disproportionately associated with corresponding minutiae. Among minutiae that examiners indicated as corresponding, 20% were added and 7% were moved during Comparison. Examiners individualized 140 times with 8 or fewer corresponding minutiae (i.e., minutiae for which the correspondence between the latent and exemplar was annotated). In most cases (93 of 140), at least one of the corresponding minutiae was added after the Analysis phase. In fact, 85 of these individualizations depended on fewer than 6 corresponding minutiae that had been marked during Analysis (Appendix B.12).

All examiners added or deleted minutiae in the Comparison phase. Indeed, most examiners (86%) added or deleted minutiae in the majority of their comparisons, and 97% added or deleted minutiae the majority of the time when individualizing. The frequency of changes varied substantially by examiner: half of all deletions were made by 32 of the 170 examiners; half of all additions were made by 48 examiners (Appendix B.9).

Comparisons tended to result in a net increase in total minutiae. We see a strong subjective component to these changes (Appendix B.9). Based on a model of the change in minutia count as a response to the image pair and examiner (Appendix B.10), examiners account for much more of the variance in net increase in total minutiae than do the images, especially for nonmates (42.8% of variance can be attributed to the examiner, 0.5% image pair, 56.8% residual) as opposed to mates (21.7% examiner, 7.9% image pair, 70.4% residual). The effects of image pairs are greatest among true positives (24.1% examiner, 11.6% image pair, 64.3% residual).

Interexaminer consistency in markup tended to increase as a result of changes made during the Comparison phase (Appendix B.13). For example, among individualizations, the proportion of minutiae in Clear areas (using the median clarity across multiple examiners) that 100% of examiners marked increased from 17% (Analysis phase) to 23% (Comparison phase).

The highest rates of changed minutiae occurred when examiners individualized; the rates of added minutiae were particularly high among those minutiae that the examiners indicated as corresponding between the latent and exemplar (Appendix B.8). Among exclusions and inconclusives, mated pairs had higher rates of change than nonmated pairs, suggesting that high rates of change during individualizations may be due in part to comparisons with mated exemplars drawing further attention to the latents' features, and not simply to examiners feeling particularly motivated to document individualization decisions. Clarity and difficulty were strong factors further explaining change rates: changed minutiae, particularly additions, occurred at much higher rates in Unclear areas regardless of the determination; within any given category of determination, change rates increased substantially with the examiner's assessment of difficulty. As a rule, these factors (determination, mating, correspondence, clarity, difficulty) were complementary, with deletion

rates below 1% in Clear areas on very easy to easy non-individualizations, rising above 10% in Unclear areas on moderate to very difficult individualizations; addition rates ranged from below 5% in Clear areas on non-individualizations for non-corresponding minutiae to nearly 70% in Unclear areas on individualizations for corresponding minutiae.

Changes on the erroneous individualization

The sole erroneous individualization was an extreme example of deleted and added minutiae. The examiner based the individualization conclusion almost entirely on minutiae that had not been detected during Analysis (Figure 2). During Comparison, the clarity markup was completely revised, minutiae in green areas were deleted, minutiae were added in newly green areas, and Clear features in overlapping areas were not marked. Such behavior was highly unusual across examiners, and this instance was the most extreme example of changed minutiae markup between Analysis and Comparison (Appendix B.11). Ten other examiners were assigned this nonmated image pair: eight excluded, two were inconclusive.

The error appears to be a consequence of incautious work by this examiner: in 16 of 22 comparisons, this examiner retained none of the minutiae from Analysis. This examiner also had the highest deletion rate among all participants (7.5 minutiae per comparison, compared with a median of 0.7), and a relatively high addition rate. Other examples of associations between erroneous individualizations and extensive changes between Analysis and Comparison were shown in the Mayfield misidentification [8] and in the Neumann et al. study [18]. However, extensive changes were not uniquely associated with erroneous individualizations: both here and in the Neumann study, examiners sometimes made extensive changes on correct individualizations; in the Neumann study, false positive errors were observed that did not involve extensive changes.

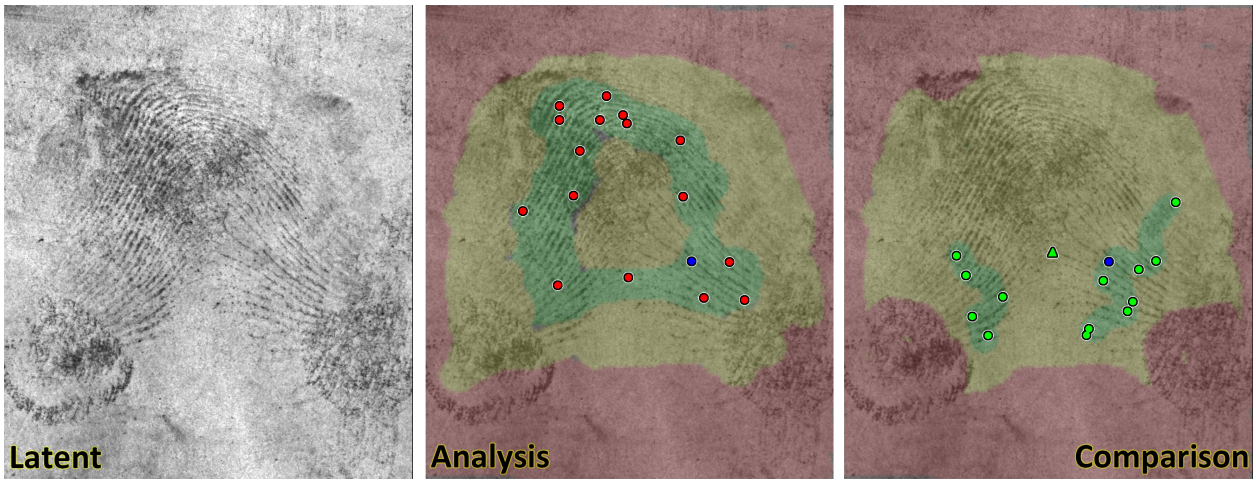


Figure 2: Image pair that resulted in the sole erroneous individualization.[6] Minutiae are shown as circles, deltas as triangles, cores as squares. 16 minutiae were deleted (red), 13 added (green), 1 moved (blue), and 0 retained; 1 delta was added. The examiner rated this comparison Easy.

Changes in latent value determinations

Latents assessed to be VEO during the Analysis phase were often individualized when compared to a mated exemplar: 26% of VEO latents on mated pairs resulted in individualizations (Table 1). The 103 VEO individualizations were not concentrated on a few latents (68 distinct latents), nor limited to a few examiners (69 distinct examiners); most of these latents (43/68) were individualized by the majority of examiners.

[6] *Histogram equalization was used to adjust the grayscale values on the latent image for journal reproduction. The exemplar involved in the erroneous individualization cannot be shown for privacy reasons: fingerprints are protected as Personally Identifiable Information and public release requires permission from the subject, which could not be obtained for this fingerprint.*

On our Black Box test [22], VEO individualizations were much less common: 1.8% of VEO latents on mated pairs were individualized. Because a VEO determination is an assertion that the latent is ***not*** of value for individualization, the contradiction between the initial VEO and the resulting individualization is notable and may indicate inadequate Analysis or inappropriate individualization determinations. We tested whether this difference in VEO individualization rates could be an artifact of data selection: when we control for data selection by limiting to a subset of 83 image pairs used in both tests, we found no substantial difference in the proportion of latents rated VEO in Analysis (22.4% White Box vs. 22.9% Black Box), and the differences in VEO individualizations rates remain (24% White Box vs. 3% Black Box). The most notable difference between the two tests was that examiners were required to provide markup in White Box and not in Black Box. While not conclusive, the results suggest that the striking increase in VEO individualizations may have resulted from requiring examiners to provide markup during Comparison (supporting analyses in Appendix B.14).

As summarized in Table 5 (see also Table 1), White Box examiners increased value determinations from VEO to VID at a much higher rate than they decreased from VID to VEO. They also changed value determinations at a much higher rate when comparing the latent to a mated exemplar than when comparing to a nonmated exemplar. When comparing to a nonmated exemplar, they more often reduced the value determination than increased. We tested whether these patterns could be explained by differences in the selection of latents for mated and nonmated pairs by using a subset of 19 latents, each of which was assigned in both mated and nonmated pairings: the patterns noted in Table 5 continue to hold when tested on that subset.

Latent Value from Analysis	**Latent Value from Comparison**							
	Mates				**Nonmates**			
	NV	*VEO*	*VID*	*Changed*	*NV*	*VEO*	*VID*	*Changed*
VEO	3.7%	69.1%	27.2%	30.9%	4.3%	93.6%	2.1%	6.4%
VID	0.7%	1.2%	98.1%	1.9%	1.5%	2.7%	95.8%	4.2%

Table 5. Latent value in Analysis and Comparison: percentages corresponding to Table 1. Changed value determinations are highlighted. (n=3,006)

In the Black Box study [23], we saw that VEO determinations in particular were not highly repeatable: when retested several months later, examiners changed 45% of their latent VEO determinations (33% when retested within hours or days). Here we see a distinct, but related phenomenon due to the influence of the exemplar where examiners often changed their VEO determinations immediately after the initial Analysis assessment: 30.9% of VEO determinations were changed when the latent was compared to a mate; 6.4% were changed when the latent was compared to a nonmate (Table 5).

Discussion

We observed frequent changes in markups of latents between the Analysis and Comparison phases. All examiners revised at least some markups during the Comparison phase, and almost all examiners changed their markup of minutiae in the majority of comparisons when they individualized. However, the mere occurrence of deleted or added minutiae during Comparison is not an indication of error: most changes were not associated with erroneous conclusions; the error rates on this test were similar to those we reported previously [22].

Extensive or fundamental changes between Analysis and Comparison ("reanalysis") may indicate that the initial Analysis was inadequate, or that the reanalysis was biased by the exemplar, raising a concern about the reliability of the comparison conclusion. The sole erroneous individualization observed in this study was an extreme example in which the examiner's conclusion was based almost entirely on minutiae that had not been marked during Analysis, and clear features marked in Analysis were deleted in Comparison. Unfortunately, such changes do not appear to be sensitive or specific predictors of erroneous individualizations; examiners sometimes made extensive changes on correct individualizations. Current SWGFAST guidance [26] specifies that any changes made to the latent markup during Comparison be documented, but offers no rules as to what should constitute an acceptable markup change; how much change is acceptable is left as a policy issue. The fact that the one erroneous individualization in our study was made by the examiner who had the highest minutia deletion rate among all participants and also a high addition rate suggests that the problem could be addressed proactively, by implementing processes to detect

when examiners exhibit anomalously high rates of changed features. A linear ACE process that requires detailed markup of changes between Analysis and Comparison could be a useful component of training.

Rates of minutia changes were much higher when examiners individualized than when they were inconclusive, and lower still when they excluded. The tendency to make more changes when individualizing may reflect a strong motivation to thoroughly and carefully support and document these conclusions, a practice of revising the latent markup to more accurately show the final interpretation of feature correspondences, and also a lack of clear standards as to how to properly document inconclusive and exclusion determinations. Among inconclusive and exclusion determinations, examiners added minutiae more frequently when the image pair was mated than nonmated, presumably because comparison with a mated exemplar draws attention to additional corresponding features in the latent; deletion rates were not sensitive to mating. Examiners frequently deleted and added minutiae in Clear areas, even when they rated the comparison as "easy." Given that this clarity designation was supposed to indicate certainty in the location or presence of minutiae, one might have expected the associated change rates to be negligible, which was not the case. The median of multiple examiners' clarity markups was a slightly better predictor of feature changes than the examiners' individual clarity markups.

Some changes are understandable and presumably benign. For example, during Analysis, an examiner may be certain of the presence of a feature, but uncertain of its location (as when three ridges become two in a low-clarity area), then revise its location during Comparison without any implication that the examiner necessarily assigned undue weight to that feature; such adjustments may explain most of the features that we classified as "moved," and some of the deleted and added features.

When examiners were required to provide detailed markup and to compare prints that they assessed to be of value for exclusion only (not of value for individualization), they often changed their value determinations and individualized. Such individualizations were much less common in our previous study where examiners did not provide markup (1.8% vs. 26%). This suggests that the detailed process of marking correspondences may have affected the examiners' assessments of the potential for individualization. These changed value determinations were not limited to a small subset of the latents or examiners, and nearly all of these individualizations were reproduced by at least one other examiner. This finding suggests that comparing marginal value latents and providing detailed markup may result in additional individualizations. Whether this should be encouraged may depend on other factors, such as the other prints available in a case, added labor costs, and the potential risk of a higher rate of erroneous or non-consensus conclusions. As we found in [21], "the value of latent prints is a continuum that is not well described by a binary determinations." The community would benefit from improved, standardized procedures to handle these borderline value determinations.

It is important to consider that the high change rates may be due in part to participants' unfamiliarity with test tools and instructions, and to their bringing casework habits to the test. For example, although we instructed participants to mark all minutiae, cores and deltas during Analysis, many examiners have been conditioned by AFIS vendor training that discourages marking certain areas or types of features, such as minutiae near cores or deltas. Many agencies do not annotate at all, or only for AFIS searches. For those that do annotate, the practice in some agencies is for examiners to mark just enough features in Analysis to document the value determination and move on to Comparison. Additionally, we observed some anomalous changes that were unrelated to Analysis, as when an examiner simply deleted features on the latent in areas that did not overlap with the exemplar. Studies such as this as well as reviews of casework are impeded by the lack of standardization in current practice, which makes the interpretation of markup difficult.

For quality assurance and documentation of casework, we believe that there is a need for examiners to have some means of unambiguously documenting what they see during Analysis and Comparison. This need for standardization of ACE-V documentation does not necessarily imply that such documentation should be mandated across all casework, which is a policy decision that entails cost-benefit tradeoffs. Such detailed documentation could enable a variety of enhancements to training and operational casework such as improved resolution of disagreements between examiners and verifiers (conflict resolution), standardized documentation for testimony, and more detailed information available for technical review and (non-blind) verification of casework. Detailed documentation in standard machine-readable formats would enable increased automation of quality assurance procedures, such as automated flagging of examinations with decisions based on marginal or apparently insufficient information, or with extensive changes between

Analysis and Comparison. Flagged examinations could then undergo additional verification, or be reviewed for potentially inappropriate conclusions; examiners whose work is routinely flagged may benefit from additional training. We concur with others [8, 9, 18, 27, 28] who have stated that rigorously defined and consistently applied methods of performing and documenting ACE-V would improve the transparency of the latent print examination process.

Acknowledgments

We thank the latent print examiners who participated in this study, as well as George Kiebuzinski, William Fellner, Erik Stanford, Michael Beachler, and Kristen Allen. This is publication number 14-16 of the FBI Laboratory Division. Names of commercial manufacturers are provided for identification purposes only and inclusion does not imply endorsement of the manufacturer or its products or services by the FBI. This work was funded 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.

References

1 Standard for the Documentation of Analysis, Comparison, Evaluation and Verification (ACE-V) (Latent), Version 2.0 (2012) (http://www.swgfast.org/documents/documentation/121124_Standard-Documentation-ACE-V_2.0.pdf)

2 Office of the Inspector General (2011) A Review of the FBI's Progress in Responding to the Recommendations in the Office of the Inspector General Report on the Fingerprint Misidentification in the Brandon Mayfield Case. US Department of Justice, Washington, DC (http://www.justice.gov/oig/special/s1105.pdf)

3 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 Interagency/Internal Report (NISTIR) 7842 (http://www.nist.gov/customcf/get_pdf.cfm?pub_id=910745)

4 Kassin SM, Dror IE, Kukucka J (2013) The forensic confirmation bias: Problems, perspectives, and proposed solutions. J. Applied Res. in Memory and Cognition 2(1):42-52 (http://dx.doi.org/10.1016/j.jarmac.2013.01.001)

5 Vanderkolk J (2004) "ACE+V: A Model". Journal of Forensic Identification, 54(1):45-52

6 Interpol European expert group on Fingerprint Identification II (2004) Method for Fingerprint Identification, Part 2: II: Detailing the method using common terminology and through the definition and application of shared principles

7 Ashbaugh D (1999) Quantitative-Qualitative Friction Ridge Analysis: an Introduction to Basic and Advanced Ridgeology. CRC Press, Boca Raton, FL.

8 Office of the Inspector General (2006) A review of the FBI's Handling of the Brandon Mayfield case. US Department of Justice, Washington, DC.

9 Langenburg G (2012) A critical analysis and study of the ACE-V process. Doctoral Dissertation, University of Lausanne, Switzerland

10 Busey T, Yu C, Wyatte D, Vanderkolk J (2013) Temporal Sequences Quantify the Contributions of Individual Fixations in Complex Perceptual Matching Tasks. Cognitive Science 37(4):731-756 DOI: 10.1111/cogs.12029

11 Langenburg G, Champod C (2011) The GYRO System - A Recommended Approach to More Transparent Documentation. Journal of Forensic Identification 61(4):373-384

12 Federal Bureau of Investigation; Universal Latent Workstation (ULW) Software. (https://www.fbibiospecs.org/Latent/LatentPrintServices.aspx)

13 Mideo Latentworks Software. (http://www.mideosystems.com/latentworks.htm)

14 PiAnoS Picture Annotation System. (https://ips-labs.unil.ch/pianos)

15 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)

16 Chapman W, et al (2013) Markup Instructions for Extended Friction Ridge Features. NIST Special Publication 1151. US Department of Commerce, National Institute of Standards and Technology, Washington, DC. (http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1151.pdf)

17 Dror I, et al (2011) Cognitive issues in fingerprint analysis: Inter- and intra-expert consistency and the effect of a 'target' comparison. Forensic Sci Int 208(1):10-17.

18 Evett IW, Williams RL (1996) A Review of the Sixteen Point Fingerprint Standard in England and Wales, Journal of Forensic Identification, 46(1) [Also published in Fingerprint Whorld, 21 (82), October, 1995]

19 Neumann C, et al. (2013) Improving the Understanding and the Reliability of the Concept of "Sufficiency" in Friction Ridge Examination. National Institute of Justice NCJ 244231 (https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=266312)

20 Hicklin RA, et al., (2011) Latent fingerprint quality: a survey of examiners, J. Forensic Ident. 61(4):385–418

21 Hicklin RA, Buscaglia J, Roberts MA (2013) Assessing the clarity of friction ridge impressions. Forensic Sci Int 226(1):106-117. (http://www.noblis.org/media/4b209d60-a147-414e-8bba-90c3a1e22c18/docs/article_assessing_clarity_latent_friction_ridge_pdf)

22 Ulery B, Hicklin R, Kiebuzinski G, Roberts M, Buscaglia J (2013) Understanding the sufficiency of information for latent fingerprint value determinations. Forensic Sci Int 230(1):99-106. (http://www.noblis.org/media/3c760709-5971-4efe-8edf-f00435fcdf1b/docs/article_understanding_sufficiency_information_latent_pdf)

23 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)

24 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)

25 Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2014) Measuring what latent fingerprint examiners consider sufficient information for individualization determinations. PLoS ONE 9(11): e110179. doi:10.1371/journal.pone.0110179 (http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0110179)

26 Scientific Working Group on Friction Ridge Analysis, Study and Technology (2013) Standards for Examining Friction Ridge Impressions and Resulting Conclusions (Latent/Tenprint), Version 2.0 (http://www.swgfast.org/documents/examinations-conclusions/130427_Examinations-Conclusions_2.0.pdf)

27 Swofford H, et al. (2013) Inter- and intra-examiner variation in the detection of friction ridge skin minutiae. Journal of Forensic Identification 63(5):553-569.

28 Haber RN, L Haber (2009) Challenges to fingerprints. Lawyers & Judges Publishing Company, Tucson AZ.

Preprint of article subsequently published:

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

Changes in latent fingerprint examiners' markup between Analysis and Comparison Supporting Information

Appendix A Materials and methods

A.1 Glossary

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

ACE	The phases of ACE-V prior to verification: Analysis, Comparison, Evaluation.
ACE-V	The prevailing method for latent print examination: Analysis, Comparison, Evaluation, Verification.
Added feature	A feature that was added during the Comparison phase: present in the Comparison markup only, with no Analysis feature within 0.5mm.
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.
ANSI/NIST-ITL	An electronic file and interchange format that is the basis for biometric and forensic standards used around the world, including the FBI's EBTS and Interpol's INT-I, among others. As of 2011, this incorporates the Extended Feature Set (EFS) definition of friction ridge features used in this study. [1]
Clarity	The clarity of a friction ridge impression refers to the fidelity with which anatomical details are represented in a 2D impression, and directly corresponds to an examiner's confidence that the presence, absence, and details of the anatomical friction ridge features in that area can be correctly discerned in that impression. (Note: The term "clarity" is used here instead of "quality" to avoid ambiguity, since the latter term as used in biometrics and forensic science is often used to include not only clarity but also the quantity or distinctiveness of features.)
Comparison/Evaluation phase	The second and third phases of the ACE-V method. In this test, there was no procedural demarcation between the Comparison and Evaluation phases of the ACE-V method; hence, this refers to the single combined phase during which both images were presented side-by-side.
Comparison determination	The determination of individualization, exclusion, or inconclusive reached in the Comparison/Evaluation phase of the test. SWGFAST [2] refers to this determination as the Evaluation Conclusion.
Corresponding features	A 1:1 relationship between a feature in a latent and a feature in the exemplar in which the feature is present in both images.
Deleted feature	A feature that was marked during the Analysis phase but deleted in the Comparison phase: present in the Analysis markup only, with no Comparison feature within 0.5mm.
Determination	An examiner's decision: the Analysis phase results in a Value determination, and the Comparison/Evaluation phase results in a Comparison determination.
EFS	The Extended Feature Set — fingerprint and palmprint features as defined in ANSI/NIST-ITL.
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.
False negative	An erroneous exclusion of a mated image pair by an examiner.
False positive	An erroneous individualization of a nonmated image pair by an examiner.
Feature	Minutia, core, delta, or "other" point marked by examiners. In this study, a feature has a location (x,y coordinate) but no direction.
IAFIS	The FBI's Integrated Automated Fingerprint Identification System (as of 2013, IAFIS latent print services have been replaced by the FBI's Next Generation Identification (NGI) system).
IAI	International Association for Identification
Image	A fingerprint as presented on the computer screen to test participants. The test software permitted rotating, panning, zooming, tonal inversion, and grayscale adjustment of the image.
Incipient ridge	A friction ridge not fully formed that may appear shorter and thinner in appearance than fully developed friction ridges.
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. Individualization is synonymous with identification for latent print determinations in the U.S. 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."[2,3]
Insufficient	When referring to examiner determinations (response data), "Insufficient" responses include both latent NV determinations (Analysis phase) and inconclusive determinations (Comparison/Evaluation phase).
Latent (or latent print)	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).
Level-3 detail	Friction ridge dimensional attributes such as width, edge shapes, and pores.

Local clarity map	A color-coded annotation of a friction ridge image indicating the clarity for every location in the print, as described in [5] and defined in the ANSI/NIST-ITL standard [1].
Mated	A pair of images (latent and exemplar) known *a priori* to derive from impressions of the same source (finger). Compare with "individualization," which is an examiner's ***determination*** that the prints are from the same source.
Median clarity map	A local clarity map combining the annotations from multiple examiners, based on the median clarity at each location across the clarity maps from all examiners who annotated the clarity of an image.
Minutiae	Events along the path of a single path, including bifurcations and ending ridges. In this study, examiners did not differentiate between bifurcations and ending ridges. Dots are considered minutiae in some uses, but not for AFIS usage; in this study, examiners were instructed to mark dots as "other" features.
Moved feature	A feature that was marked during the Analysis phase but moved in the Comparison phase: the Analysis and Comparison features are not in the same exact location, but are within 0.5mm.
Noncorresponding feature	A discrepancy – a feature that exists in one print and is definitely not present in the other print. Participants were instructed to indicate points in one print that definitely do not exist in the other print as needed to support an exclusion determination.
Nonmated	A pair of images (latent and exemplar) known *a priori* to derive from impressions of different sources (different fingers and/or different subjects).
NV (No value)	The impression is not of value for individualization and contains no usable friction ridge information. See also VEO and VID.
Other point	In this study, features such as scars, dots, incipient ridges, creases and linear discontinuities, ridge edge features, or pores (i.e., features other than minutiae, cores, and deltas).
Repeatability	Intraexaminer agreement: when one examiner provides the same response (annotation or determination) to a stimulus (image or image pair) on multiple occasions.
Reproducibility	Interexaminer agreement: when multiple examiners provide the same response (annotation or determination) to a stimulus (image or image pair).
Retained feature	A feature that was marked during the Analysis phase and was not modified in the Comparison phase: the Analysis and Comparison features are in the same exact location.
Source	An area of friction ridge skin from which an impression is left. Two impressions are said to be from the "same source" when they have in common a region of overlapping friction ridge skin.
Sufficient	An examiner's assessment that the quality and quantity of information in a print (or image pair) justifies a specific determination (especially used with respect to individualization).
True Negative	The exclusion of a nonmated image pair by an examiner
True Positive	The individualization of a mated image pair by an examiner
ULW	The FBI's Universal Latent Workstation software. [6]
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). A latent value determination is made during the Analysis phase. Agency policy often reduces the three value categories into two, either by combining VID and VEO into a value for comparison (VCMP) category or by combining VEO with NV into a "not of value for individualization" (Not VID) category [survey in 7].
VCMP	Value determination based on the analysis of a latent that the impression is of value for comparison (either VEO or VID).
VEO	Value determination based on the analysis of a latent that the impression is of value for exclusion only and contains some friction ridge information that may be appropriate for exclusion if an appropriate exemplar is available. See also NV and VID.
VID	Determination based on the analysis of a latent that the impression is of value and is appropriate for potential individualization if an appropriate exemplar is available. See also VEO and NV.

A.2 Test procedure

Fig. S1 summarizes the test workflow, which conforms broadly to the prevailing ACE methodology.[##][8] This study did not address the Verification phase. Examiners could review and revise their work prior to submitting their results. Examiners were free to modify the markup and value determination for the latent after the exemplar was presented, but any such changes were recorded and could be compared with their Analysis responses. For a more complete description of the test procedure, including the complete test instructions and introductory video, see our previous report [9].

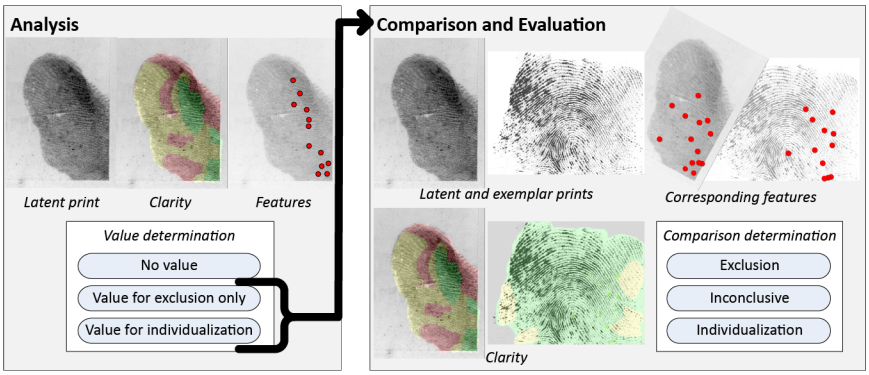


Fig. S1: Test workflow. Each examiner was assigned a distinct, randomized sequence of latent-exemplar image pairs. For each pair, the latent was presented first for a value determination. If the latent was determined to be no value, the test proceeded directly to the latent from the next image pair; otherwise, an exemplar was presented for comparison and evaluation. (From [9])

Usage of latent value determination categories varies by agency, and therefore differed from standard operating procedures for some participants. Table S1 summarizes three general approaches to classifying latent value (suitability) and the prevalence of each among study participants [See 7, Appendix 1.4; 9, Appendix B].

Levels	Definition	Usage
VID vs. not VID (2 levels)	NV and VEO are combined into "not VID." (SWGFAST "Approach 1").	Used by 48% of White Box participants (55% of Black Box participants).
VCMP vs. NV (2 levels)	VEO and VID are combined into value for comparison (VCMP). (SWGFAST "Approach 2").	Used by 8% of White Box participants (14% of Black Box participants).
VID vs. VEO vs. NV (3 levels)	Three distinct categories are used.	Used by 44% of White Box participants (30% of Black Box participants). Some agencies use the VEO category only upon request.

Table S1. Three general approaches to classifying latent value (suitability). Usage statistics were collected via surveys of the Black Box and White Box study participants, who were asked about "the standard operating procedures that you/your agency currently use." The SWGFAST approaches are described in [2].

A.3 Fingerprint data

The fingerprints for the study were collected at the FBI Laboratory and at Noblis under controlled conditions, and from operational casework datasets collected by the FBI. We provide a detailed description of the fingerprint data selection process in [9, Appendix S.5]. All prints were impressions of distal segments of fingers, including some sides and tips. See Fig. S2 for examples.

The latents were processed using a variety of development techniques. The processed latents were captured electronically at 8-bit grayscale, uncompressed, at a resolution of 1000 pixels per inch.

The test workflow differs from the detailed flow diagram published by the Expert Working Group on Human Factors in Latent Print Analysis [8, Figure 1.1], which specifies comparing the latent and known during Analysis to assess whether there is sufficient quality and quantity in common between the two prints to be suitable for comparison.

The exemplars included both rolled and plain impressions captured as inked prints on paper cards or using FBI-certified livescan devices; they were captured at 8-bit grayscale, 1000 or 500 pixels per inch and either uncompressed or compressed using Wavelet Scalar Quantization [10].

We selected nonmated pairs to result in challenging comparisons either by down-selecting among exemplar prints returned by searches of the FBI's Integrated AFIS (IAFIS) or from among neighboring fingers from the same subject.

The assignments of fingerprint images to examiners were randomized based on an Incomplete Block Design (with examiners as blocks, image pairs as factor levels), balanced to the extent possible using the criterion of D-Optimality.

For each image pair assigned to an examiner, the test process saved two data files: one saved upon completion of the Analysis phase (before the exemplar print was presented) and a second upon completion of the Comparison/Evaluation phase. The files complied with the ANSI/NIST-ITL [1] standard, using the COMP transaction described in the Latent Interoperability Transmission Specification [11].

Fig. S2: Example latents (top row) and mated exemplars (bottom row). These latents are shown with markup in Figure 1. The rightmost exemplar is of a fingertip.

A.4 Local ridge clarity

The annotations of local ridge clarity complied with the Extended Feature Set (EFS), which is part of the ANSI/NIST-ITL standard [1]. Fig. S3 summarizes the color-coding method for describing clarity [5]. For minutiae, the primary distinction with regard to clarity is that for green or better areas, the examiner is "certain of the location, presence, and absence of all minutiae" (White Box Instructions [9, Appendix S22]). Yellow areas indicate the opposite, that location, presence, and/or absence are not certain. Black or red areas should not have any marked minutia. When this occurs it is often due to imprecise painting of the clarity, or to not following instructions. For this analysis, we simplified the classification to Clear (green or better) vs. Unclear (yellow or worse).

Unless otherwise stated, we report the (final) Comparison phase clarity as marked by that examiner, except for deleted minutiae where we use the Analysis phase clarity. In some analyses we use the median clarity across multiple examiners, which combines the clarity maps from the examiners who were assigned that pair to represent a group consensus. This reduces the impact of outlier opinions and imprecision. When

constructing the median clarity maps, we excluded four examiners whose clarity markup did not comply with the test instructions.

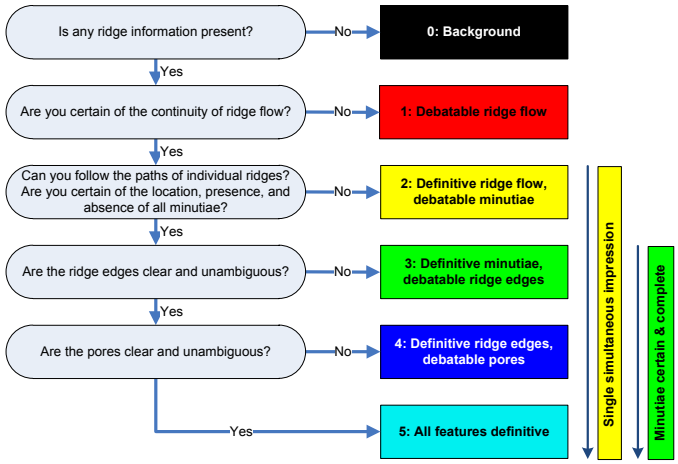


Fig. S3: Definitions of local image clarity. (From [12])

A.5 Determining whether a feature was moved

Because it would have been impractical to manually review the more than 50,000 feature markings in more than 3,000 paired markups, an automated method was necessary. Our technique of classifying features as retained, moved, added or deleted based on a radius of 0.5 mm (20 pixels at 1000 ppi, or approximately the inter-ridge distance) worked satisfactorily for these analyses, but is imperfect. Here we discuss briefly the accuracy of this classification approach.

Fig. S4 shows the distance from where each minutia was marked in the Analysis phase to the nearest minutiae marked in the Comparison phase (for the same latent and the same examiner). Thus, the distributions describe pairs of markings, some of which indicate the same minutia (placement revised) and some of which indicate distinct minutiae. 0.5 mm represents the approximate decision threshold above which pairs of markings are more likely to indicate distinct minutiae, and below which they are more likely to indicate the same minutia. To better show the relevant portion of the histogram, it is truncated at both ends by removing the large spike at zero (minutia marked at exactly the same location in Analysis and Comparison) and minutia pairs whose distance exceeded 2.0 mm. Most of the data in the complete underlying distribution is either in the spike at zero or the long tail to the right.

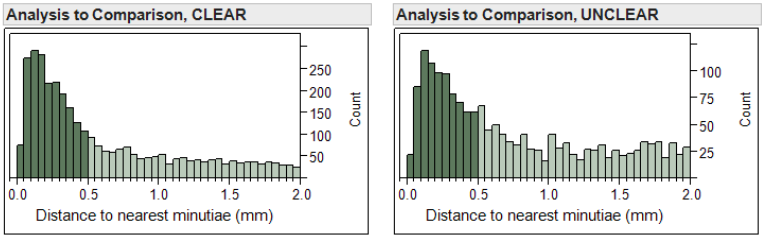


Fig. S4: Distance from Analysis-phase feature markings to the nearest Comparison-phase marking. Feature clarity was obtained from the Comparison-phase median clarity map. 14 examiners who changed their markup often or extensively were omitted.

Fig. S5 shows the distribution of distances between pairs of distinct minutiae as measured on the final (Comparison phase) markups. Through manual review of the markups together with these measurements, we selected 0.5 mm as the threshold for classifying features as moved or distinct. For 98.6% of minutiae, the distance to the nearest neighboring minutia was greater than 0.5 mm.

Although the 0.5 mm rule is generally effective, the resulting classification is not data provided by the examiners: some minutiae that the examiner may have considered as moved we treat as distinct deleted and added minutiae, and vice versa. For example, a very small percentage of features were classified as having their type changed (Table 2), and approximately half of these were also "moved." Our classification criterion may have contributed substantially to those moved features whose types also changed.

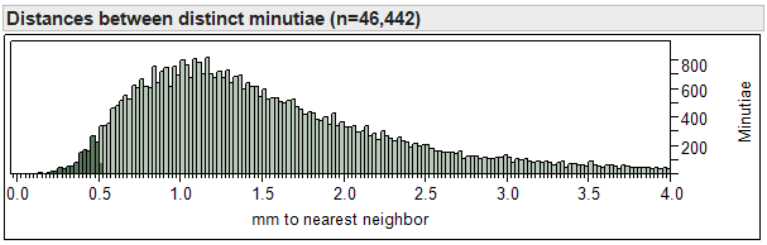


Fig. S5: Distances between ***distinct*** minutiae as marked in Comparison-phase markups: 1.4% of minutiae were within 0.5mm of each other (shown shaded). 465 of 3,008 Comparison phase markups have two minutiae within 20 pixels of each other. Distribution is truncated at 4 mm.

A.6 Anomalous markups

A small fraction of the modifications were anomalous either because the examiner did not follow instructions properly or because the examiner showed unusual behavior. As a rule, we had no objective way to differentiate between extreme (but legitimate) responses and outright errors or a failure to follow instructions. Therefore, unless explicitly noted, we did not exclude anomalous markups from our analyses.

In most cases, the anomalies were specific to one aspect of the markup, with nothing unusual about the remainder of the markup. For example, sometimes examiners applied the clarity colors improperly and marked debatable ridge flow as background.

Examples of unusual behavior include one examiner who deleted latent features whenever the determination was an exclusion. The examiner who made the erroneous individualization routinely deleted all Analysis markup and started feature markup afresh in Comparison. Occasionally examiners deleted features that were not in an area that overlapped with the exemplar. Changes to clarity tended to be minor local adjustments, excepting those of a few examiners who routinely redid their clarity markup during Comparison.

Appendix A References

1 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)

2 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)

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 Hicklin RA, Buscaglia J, Roberts MA (2013) Assessing the clarity of friction ridge impressions. Forensic Sci Int 226(1):106-117. (http://www.noblis.org/media/4b209d60-a147-414e-8bba-90c3a1e22c18/docs/article_assessing_clarity_latent_friction_ridge_pdf)

6 Federal Bureau of Investigation; Universal Latent Workstation (ULW) Software. (https://www.fbibiospecs.org/Latent/LatentPrintServices.aspx)

7 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)

8 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 Interagency/Internal Report (NISTIR) 7842 (http://www.nist.gov/customcf/get_pdf.cfm?pub_id=910745)

9 Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2014) Measuring what latent fingerprint examiners consider sufficient information for individualization determinations. PLoS ONE 9(11): e110179. doi:10.1371/journal.pone.0110179 (http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0110179)

10. Criminal Justice Information Services (1997) Wavelet Scalar Quantization (WSQ) Gray-Scale Fingerprint Image Compression Specification, Version 3.1. (https://www.fbibiospecs.org/docs/WSQ_Gray-scale_Specification_Version_3_1_Final.pdf)

11 Chapman W, et al. (2013) Latent Interoperability Transmission Specification. NIST Special Publication 1152. US Department of Commerce, National Institute of Standards and Technology, Washington, DC. (http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1152.pdf)

12 Chapman W, et al (2013) Markup Instructions for Extended Friction Ridge Features. NIST Special Publication 1151. US Department of Commerce, National Institute of Standards and Technology, Washington, DC. (http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1151.pdf)

Appendix B Results

B.1 Feature modifications by Comparison determination

Table S2 through Table S5 describe how many minutiae, cores, deltas, and other features changed between Analysis and Comparison by mating and Comparison determination on 2,957 paired markups. For the counts, Comparison = Analysis – Deleted + Added. The percentages are relative to the number marked during Analysis. We show percentages based on fewer than 50 features in gray. Features whose type changed from Analysis to Comparison are not reported in these tables.

MINUTIAE	Number of minutiae						% of Analysis minutiae			
	Analysis	Retained	Moved	Deleted	Added	Comparison	Retained	Moved	Deleted	Added
Mate Indiv	28,224	24,038	1,945	2,241	5,962	31,945	85.2%	6.9%	7.9%	21.1%
Mate Inconc	5,866	5,167	310	389	847	6,324	88.1%	5.3%	6.6%	14.4%
Mate Excl	1,645	1,536	59	50	140	1,735	93.4%	3.6%	3.0%	8.5%
Nonmate Excl	4,811	4,494	161	156	220	4,875	93.4%	3.3%	3.2%	4.6%
Nonmate Inconc	1,211	1,090	38	83	62	1,190	90.0%	3.1%	6.9%	5.1%
Nonmate Indiv	17	-	1	16	13	14	*0.0%*	*5.9%*	*94.1%*	*76.5%*
All mates	35,735	30,741	2,314	2,680	6,949	40,004	86.0%	6.5%	7.5%	19.4%
All nonmates	6,039	5,584	200	255	295	6,079	92.5%	3.3%	4.2%	4.9%
Total	41,774	36,325	2,514	2,935	7,244	46,083	87.0%	6.0%	7.0%	17.3%

Table S2: Minutia changes by Comparison determination.

CORES	Number of cores						% of Analysis features			
	Analysis	Retained	Moved	Deleted	Added	Comparison	Retained	Moved	Deleted	Added
Mate Indiv	680	588	39	53	108	735	86.5%	5.7%	7.8%	15.9%
Mate Inconc	194	175	5	14	32	212	90.2%	2.6%	7.2%	16.5%
Mate Excl	36	32	-	4	9	41	*88.9*	*0.0*	*11.1*	*25.0*
Nonmate Excl	136	127	2	7	25	154	93.4%	1.5%	5.1%	18.4%
Nonmate Inconc	33	31	-	2	1	32	*93.9*	*0.0*	*6.1%*	*3.0%*
Nonmate Indiv	-	-	-	-	-	-	-	-	-	-
All mates	910	795	44	71	149	988	87.4%	4.8%	7.8%	16.4%
All nonmates	169	158	2	9	26	186	93.5%	1.2%	5.3%	15.4%
Total	1,079	953	46	80	175	1,174	88.3%	4.3%	7.4%	16.2%

Table S3: Core changes by Comparison determination.

DELTAS	Number of deltas						% of Analysis features			
	Analysis	Retained	Moved	Deleted	Added	Comparison	Retained	Moved	Deleted	Added
Mate Indiv	337	281	23	33	67	371	83.4%	6.8%	9.8%	19.9%
Mate Inconc	82	76	-	6	14	90	92.7%	0.0%	7.3%	17.1%
Mate Excl	21	20	-	1	4	24	*95.2%*	*0.0%*	*4.8%*	*19.0%*
Nonmate Excl	59	53	3	3	11	67	89.8%	5.1%	5.1%	18.6%
Nonmate Inconc	13	12	1	-	1	14	*92.3%*	*7.7%*	*0.0%*	*7.7%*
Nonmate Indiv	-	-	-	-	1	1	-	-	-	-
All mates	440	377	23	40	85	485	85.7%	5.2%	9.1%	19.3%
All nonmates	72	65	4	3	13	82	90.3%	5.6%	4.2%	18.1%
Total	512	442	27	43	98	567	86.3%	5.3%	8.4%	19.1%

Table S4: Delta changes by Comparison determination.

OTHER FEATURES	Number of other features						% of Analysis features			
	Analysis	Retained	Moved	Deleted	Added	Comparison	Retained	Moved	Deleted	Added
Mate Indiv	209	165	7	37	234	406	78.9%	3.3%	17.7%	112.0%
Mate Inconc	79	73	1	5	24	98	92.4%	1.3%	6.3%	30.4%
Mate Excl	9	9	-	-	-	9	*100.0%*	*0.0%*	*0.0%*	*0.0%*
Nonmate Excl	58	55	-	3	4	59	94.8%	0.0%	5.2%	6.9%
Nonmate Inconc	23	22	-	1	1	23	*95.7%*	*0.0%*	*4.3%*	*4.3%*
Nonmate Indiv	-	-	-	-	-	-	-	-	-	-
All mates	297	247	8	42	258	513	83.2%	2.7%	14.1%	86.9%
All nonmates	81	77	-	4	5	82	95.1%	0.0%	4.9%	6.2%
Total	378	324	8	46	263	595	85.7%	2.1%	12.2%	69.6%

Table S5: Other feature changes by Comparison determination.

Table S6 describes the subset of 19 latents that were presented in both mated and nonmated image pairings. This table shows substantially different change rates for minutiae marked on these 19 latents depending on whether the comparison print was mated or not.

19 latents used in both mated and nonmated image pairs	Number of comparisons	Number of minutiae		% of Analysis minutiae			
		Analysis	Comparison	Retained	Moved	Deleted	Added
Mate Indiv (True Positive)	92	1,135	1,441	76.4%	10.7%	12.9%	39.8%
Mate Inconc	71	664	704	82.5%	6.6%	10.8%	16.9%
Mate Excl (False Negative)	15	204	210	96.6%	2.0%	1.5%	4.4%
Nonmate Excl (True Negative)	103	1,116	1,151	91.7%	4.9%	3.4%	6.5%
Nonmate Inconc	30	236	224	87.7%	0.8%	11.4%	6.4%
Nonmate Indiv (False Positive)	0	-	-	-	-	-	-
All mates	178	2,003	2,355	80.5%	8.5%	11.0%	28.6%
All nonmates	133	1,352	1,375	91.0%	4.2%	4.8%	6.5%
Total	311	3,355	3,730	84.7%	6.8%	8.5%	19.7%

Table S6: Minutia changes by Comparison determination, limited to a set of 19 latents that were presented in both mated and nonmated image pairs; compare with Table 3. Use of this subset of data allows us to investigate the effects of exemplar mating without confounding due to population differences between mated and nonmated latents. This corroborates the general trends shown for the full dataset: higher rates of change for mated data, especially for added features on true positives and mated inconclusives.

In 49 instances, the examiners proceeded to the Comparison phase but did not report Comparison conclusions either because the exemplar was NV or because the examiner changed the latent value determination to NV in Comparison. Table S7 shows the minutia changes for those comparisons. Compare with Table 3.

49 instances without a Comparison conclusion	Number of comparisons	Comparisons with any added or deleted minutiae	Number of minutiae		Analysis minutiae			
			Analysis	Comparison	Retained	Moved	Deleted	Added
Latent changed to NV (mates)	29	19	241	244	178	18	45	28
Latent changed to NV (nonmates)	14	6	109	100	97	0	12	3
Exemplar NV (mates)	6	0	44	44	43	1	0	0

Table S7: Minutia changes for the 49 comparisons that did not result in comparison conclusions because the examiner either changed the latent value determination to NV or determined the exemplar to be NV.

B.2 Proportions of comparisons with modified minutiae

Table S8 shows patterns of change associated with entire markups. For example, examiners usually added one or more minutiae to their latent markups when individualizing (86.0% of true positives), but not when excluding (22.4% of true negatives and 38.8% of false negatives).

		Number of Comparisons						
		No modified minutiae	Any retained minutiae	Any moved minutiae	Any deleted minutiae	Any added minutiae	Any deleted or added minutiae	Total
Exclusions	All	347	533	99	86	145	173	557
	Mated	71	127	28	22	50	53	130
	Nonmated	276	406	71	64	95	120	427
Inconclusives	All	336	658	157	175	308	344	704
	Mated	227	522	142	152	284	309	554
	Nonmated	109	136	15	23	24	35	150
IDs	All	115	1,650	842	869	1,458	1,531	1,696
	Mated	115	1,650	841	868	1,457	1,530	1,695
	Nonmated	-	-	1	1	1	1	1
	Mated	413	2,299	1,011	1,042	1,791	1,892	2,379
	Nonmated	385	542	87	88	120	156	578
	Total	798	2,841	1,098	1,130	1,911	2,048	2,957

		% of Comparisons						
		No modified minutiae	Any retained minutiae	Any moved minutiae	Any deleted minutiae	Any added minutiae	Any deleted or added minutiae	Number of comparisons
Exclusions	All	62%	96%	18%	15%	26%	31%	557
	Mated	55%	98%	22%	17%	38%	41%	130
	Nonmated	65%	95%	17%	15%	22%	28%	427
Inconclusives	All	48%	93%	22%	25%	44%	49%	704
	Mated	41%	94%	26%	27%	51%	56%	554
	Nonmated	73%	91%	10%	15%	16%	23%	150
IDs	All	7%	97%	50%	51%	86%	90%	1,696
	Mated	7%	97%	50%	51%	86%	90%	1,695
	Nonmated	*0%*	*0%*	*100%*	*100%*	*100%*	*100%*	*1*
	Mated	17%	97%	42%	44%	75%	80%	2,379
	Nonmated	67%	94%	15%	15%	21%	27%	578
	Total	27%	96%	37%	38%	65%	69%	2,957

Table S8: Counts and percentages of comparisons having any retained, moved, deleted, or added minutiae, by conclusion. "Modified" refers to markups on which at least one minutiae was moved, deleted or added. Percentages based on fewer than 10 comparisons are in gray.

B.3 Minutia modifications by difficulty, determination, and mating

Table S9 shows the percentage of minutiae retained, moved, deleted, and added by comparison determination, mating, and difficulty.

Table S10 shows the percentage of comparisons having any deleted or added minutiae by comparison determination, mating, and difficulty.

Table S11 compares examiner's ratings of difficulty from the Black Box and White Box studies. For each combination of determination and mating, it shows the distribution of comparisons by difficulty level.

		Difficult	Comparisons	# of Minutiae		% of Analysis minutiae			
				Analysis	Comparison	Retained	Moved	Deleted	Added
Exclusions	Mate	No	78	1,016	1,070	95%	4%	1%	7%
		Yes	52	629	665	91%	3%	6%	11%
	Nonmate	No	333	3,760	3,784	94%	3%	3%	4%
		Yes	94	1,051	1,091	90%	6%	5%	8%
Inconclusives	Mate	No	210	2,190	2,339	93%	3%	4%	11%
		Yes	344	3,676	3,985	85%	7%	8%	17%
	Nonmate	No	71	504	510	95%	1%	3%	4%
		Yes	79	707	680	86%	4%	9%	6%
IDs	Mate	No	1,334	23,404	26,110	87%	6%	7%	19%
		Yes	361	4,820	5,835	77%	10%	13%	34%
	Nonmate	No	1	17	14	*0%*	*6%*	*94%*	*76%*
Subtotal		No	2,027	30,891	33,827	89%	5%	6%	16%
		Yes	930	10,883	12,256	82%	8%	10%	22%
Total			2,957	41,774	46,083	87%	6%	7%	17%

Table S9: Minutia modifications by comparison determination, mating, and difficulty. Here, "Difficult" combines the examiners' assessments of Difficult and Very difficult (yes), as opposed to Moderate, Easy, and Very easy (no). Note that the majority of inconclusives are difficult, which is not true of any other determination. Percentages based on fewer than 50 minutiae are in gray.

		Number of comparisons						% of comparisons with deleted or added minutiae					
		V.Easy	Easy	Mod.	Diff.	V.Diff	Total	V.Easy	Easy	Mod.	Diff.	V.Diff	Total
Exclusions	All	29	115	267	111	35	557	14%	19%	30%	43%	54%	31%
	Mated	3	15	60	34	18	130	*33%*	47%	35%	41%	56%	41%
	Nonmated	26	100	207	77	17	427	12%	15%	29%	44%	53%	28%
Inconclusives	All	7	38	236	267	156	704	*29%*	21%	43%	58%	51%	49%
	Mated	3	23	184	219	125	554	*33%*	30%	49%	64%	57%	56%
	Nonmated	4	15	52	48	31	150	*25%*	7%	21%	29%	26%	23%
IDs	All	107	515	713	272	89	1,696	85%	87%	92%	94%	94%	90%
	Mated	107	514	713	272	89	1,695	85%	87%	92%	94%	94%	90%
	Nonmated	--	1	-	-	-	1	*0%*	*100%*	*0%*	*0%*	*0%*	100%
	Mated	113	552	957	525	232	2,379	82%	83%	80%	78%	71%	80%
	Nonmated	30	116	259	125	48	578	13%	15%	27%	38%	35%	27%
	Total	143	668	1,216	650	280	2,957	68%	71%	69%	70%	65%	69%

Table S10: Comparisons with any deleted or added minutia by comparison determination, mating, and difficulty (n=2,957). Percentages indicate the proportion of comparisons of that difficulty that had any deleted or added minutiae. Percentages based on fewer than 10 comparisons are in gray.

Table S11 compares results from this study (WB) to our previous Black Box (BB) study [1] based on a common subset of data; because the examiners were anonymous, the extent of overlap between the participants in the two tests is unknown. Examiners rated exclusions and inconclusives as substantially more difficult when markup was required (WB) than when no markup was required (BB). The test conditions for BB and WB were similar, but in BB examiners did not provide markup. Note that in the percentages Table S11 total to 100% on each row (percentage of comparisons at each difficulty level), whereas those in Table S10 do not total to 100% (percentage of comparisons with added or deleted minutiae within each difficulty level).

			Number of comparisons						% of comparisons				
			V.Easy	Easy	Mod.	Diff.	V.Diff	Total	V.Easy	Easy	Mod.	Diff.	V.Diff
Exclusions	Mated	BB	3	8	24	7	3	45	7%	18%	53%	16%	7%
		WB	0	0	5	2	3	10	0%	0%	50%	20%	30%
	Nonmated	BB	75	192	283	84	12	646	12%	30%	44%	13%	2%
		WB	16	48	103	49	9	225	7%	21%	46%	22%	4%
Inconclusives	Mated	BB	3	15	100	76	12	206	1%	7%	49%	37%	6%
		WB	0	1	36	33	18	88	0%	1%	41%	38%	20%
	Nonmated	BB	1	63	162	76	11	313	0%	20%	52%	24%	4%
		WB	3	11	26	27	25	92	3%	12%	28%	29%	27%
IDs	Mated	BB	50	142	200	83	21	496	10%	29%	40%	17%	4%
		WB	26	88	108	43	14	279	9%	32%	39%	15%	5%
	Nonmated	BB	-	-	-	-	-	-					
		WB	-	1	-	-	-	1	*0%*	*100%*	*0%*	*0%*	*0%*
		BB	132	420	769	326	59	1,706	8%	25%	45%	19%	3%
		WB	45	149	278	154	69	695	6%	21%	40%	22%	10%

Table S11: Comparison of Black Box and White Box difficulty assessments on a common set of 83 image pairs used in both tests. The assignments of image pairs to examiners differed on the two tests: each image pair was assigned to more examiners on Black Box than White Box, however, mated image pairs were assigned proportionally more often in White Box.

B.4 Markups with minutia modifications by clarity

Examiners usually deleted or added one or more minutiae during Comparison when they individualized (Table S12). The percentage of comparisons with deleted or added minutiae was lower among exclusions and inconclusives, and lower on nonmated pairs than mated pairs. The majority of these changed markups had deleted or added minutiae in Clear areas.

			Number of comparisons		% of comparisons	
		Total Comparisons	with any deleted or added minutiae	with deleted or added Clear minutiae	with any deleted or added minutiae	with deleted or added Clear minutiae
Exclusions	Mated	130	53	38	41%	29%
	Nonmated	427	120	73	28%	17%
	All	557	173	111	31%	20%
Inconclusives	Mated	554	309	184	56%	33%
	Nonmated	150	35	22	23%	15%
	All	704	344	206	49%	29%
IDs	Mated	1695	1530	1217	90%	72%
	Nonmated	1	1	1	*100%*	*100%*
	All	1696	1531	1218	90%	72%
	Mated	2379	1892	1439	80%	60%
	Nonmated	578	156	96	27%	17%
	Total	2957	2048	1535	69%	52%

Table S12: Comparisons with deleted or added minutiae by determination, and mating. Percentages based on fewer than 10 comparisons are in gray.

B.5 Minutia modifications by clarity

Table S13 shows the percentages of minutiae retained, moved, deleted, and added by median clarity; the percentages are calculated relative to the number of minutiae marked during Analysis. Rates of deletions and additions were substantially higher in Unclear areas than in Clear areas.

Table S14 and Table S15 show the percentage of minutiae retained, moved, deleted , and added by clarity level. Minutiae marked in black or red areas are usually the result of improper or imprecise clarity painting. These tables present the same data as in Table 4, except further decomposed by each clarity level.

MEDIAN CLARITY	Clarity	Number of minutiae		% of Analysis minutiae			
		Analysis	Comparison	Retained	Moved	Deleted	Added
All comparisons (n=2,957)	Unclear	9,752	11,578	80%	6%	14%	32%
	Clear	32,022	34,505	89%	6%	5%	13%
	Total	41,774	46,083	87%	6%	7%	17%
True positives (n=1,695)	Unclear	5,197	6,673	73%	8%	18%	47%
	Clear	23,027	25,272	88%	7%	6%	15%
	Total	28,224	31,945	85%	7%	8%	21%

Table S13: Minutia changes by ***median*** clarity across multiple examiners. Compare with Table 4, which shows minutia changes by the clarity of individual examiners.

Clarity	Analysis	Comparison	Retained	Moved	Deleted	Added
Unclear	11,068	13,268	83.6%	6.5%	9.9%	29.7%
Clear	30,706	32,815	88.2%	5.8%	6.0%	12.9%
Total	41,774	46,083	87.0%	6.0%	7.0%	17.3%
Black	759	829	75.1%	11.6%	13.3%	22.5%
Red	910	1,261	82.9%	4.9%	12.2%	50.8%
Yellow	9,399	11,178	84.4%	6.2%	9.4%	28.3%
Green	24,397	26,114	87.1%	6.3%	6.6%	13.7%
Blue	5,160	5,501	92.1%	4.3%	3.7%	10.3%
Aqua	1,149	1,200	92.8%	3.8%	3.4%	7.8%

Table S14: Minutia modifications by clarity (limited to 2,957 markups where examiners performed comparisons). Percentages calculated relative to number of minutiae marked in Analysis phase. Clarity is from the Comparison phase, except for Deleted features, where it is from the Analysis phase.

Clarity	Analysis	Comparison	Retained	Moved	Deleted	Added
Unclear	6,646	8,436	79.3%	8.2%	12.4%	39.4%
Clear	21,578	23,509	87.0%	6.5%	6.6%	15.5%
Total	28,224	31,945	85.2%	6.9%	7.9%	21.1%
Black	509	578	68.8%	15.9%	15.3%	28.9%
Red	518	780	77.6%	5.4%	17.0%	67.6%
Yellow	5,619	7,078	80.4%	7.8%	11.8%	37.7%
Green	16,748	18,325	85.8%	7.0%	7.2%	16.6%
Blue	3,950	4,262	91.0%	4.8%	4.2%	12.1%
Aqua	880	922	91.7%	4.1%	4.2%	9.0%

Table S15: Minutia modifications by clarity, limited to 1,695 markups of true positives.

B.6 Changes in clarity markup

The data in this section excludes from the 2,957 comparisons 41 where clarity was not painted or left substantially incomplete during the Analysis phase, or was "erased" during Comparison; possibly these examiners were unaware that the software allowed for clarity layer of the markup to be toggled on and off. Most of these exclusions were associated with two examiners.

Examiners modified 13% of the clarity maps (16% associated with individualizations) (Table S16). These changes affected the clarity of 1.3% of retained minutiae and 9.2% of moved minutiae (Table S17). There was no strong general tendency toward either increasing or decreasing clarity assessments.

		Number of comparisons				Percentage of comparisons		
		Total	Modified clarity map	Increased clear area	Decreased clear area	Modified clarity map	Increased clear area	Decreased clear area
Exclusions	Mates	123	12	9	3	10%	7%	2%
	Nonmates	424	37	15	15	9%	4%	4%
IDs	Mates	1,670	259	97	124	16%	6%	7%
	Nonmates	1	1	-	1	*100%*	*0%*	*100%*
Inconclusives	Mates	549	67	15	36	12%	3%	7%
	Nonmates	149	12	2	7	8%	1%	5%
Subtotal	Mates	2,342	338	121	163	14%	5%	7%
	Nonmates	574	50	17	23	9%	3%	4%
Total		2,916	388	138	186	13%	5%	6%

Table S16: Counts and percentages of comparisons where the examiner modified clarity during Comparison. We measure "modified clarity" as change in total Clear area (green+) or Unclear (yellow) area.

Number of Retained Minutiae								% changed
Analysis clarity	Comparison clarity							
	Black	Red	Yellow	Green	Blue	Aqua	Total	
Black	426	0	3	12	1	1	443	3.8%
Red	0	732	5	21	4	0	762	3.9%
Yellow	1	17	7,768	135	5	0	7,926	2.0%
Green	0	4	148	20,817	82	2	21,053	1.1%
Blue	1	1	2	30	4,657	4	4,695	0.8%
Aqua	0	0	0	0	0	1,048	1,048	0.0%
Total	428	754	7,926	21,015	4,749	1,055	35,927	1.3%

Number of Moved Minutiae								% changed
Analysis clarity	Comparison clarity							
	Black	Red	Yellow	Green	Blue	Aqua	Total	
Black	12	0	2	6	1	2	23	47.8%
Red	0	27	2	3	0	1	33	18.2%
Yellow	4	9	502	43	1	1	560	10.4%
Green	5	9	73	1,405	28	0	1,520	7.6%
Blue	2	0	1	23	190	4	220	13.6%
Aqua	0	0	0	0	1	34	35	2.9%
Total	23	45	580	1,480	221	42	2,391	9.2%

Table S17: Changes of clarity for (A) retained and (B) moved minutiae. (n=2,916 responses.

B.7 Minutia modifications by conclusion, difficulty, and clarity

		Difficult	Clarity	# of Minutiae Analysis	# of Minutiae Comparison	% of Analysis minutiae Retained	Moved	Deleted	Added
Exclusions	Mates	No	Clear	762	793	95%	4%	1%	5%
			Unclear	254	277	94%	3%	3%	12%
		Yes	Clear	388	393	90%	4%	7%	8%
			Unclear	241	272	93%	3%	4%	17%
	Nonmates	No	Clear	2,844	2,837	94%	3%	3%	3%
			Unclear	916	947	96%	1%	3%	7%
		Yes	Clear	653	663	88%	7%	5%	6%
			Unclear	398	428	92%	4%	4%	12%
Inconclusives	Mates	No	Clear	1,648	1,723	95%	2%	3%	7%
			Unclear	542	616	89%	4%	7%	21%
		Yes	Clear	2,082	2,158	84%	7%	9%	12%
			Unclear	1,594	1,827	86%	6%	8%	23%
	Nonmates	No	Clear	340	352	96%	2%	2%	6%
			Unclear	164	158	95%	0%	5%	1%
		Yes	Clear	394	373	85%	6%	9%	4%
			Unclear	313	307	88%	2%	10%	8%
IDs	Mates	No	Clear	18,415	19,940	88%	6%	6%	14%
			Unclear	4,989	6,170	82%	7%	11%	34%
		Yes	Clear	3,163	3,569	80%	10%	10%	23%
			Unclear	1,657	2,266	72%	11%	18%	54%
	Nonmates	No	Clear	17	14	*0%*	*6%*	*94%*	*76%*
Total				41,774	46,083	87%	6%	7%	17%

Table S18: Minutiae changes by conclusion, difficulty, and clarity.

B.8 Factors associated with minutiae deletions and additions

Table S19 and Table S20 highlight several factors strongly associated with deleted and added minutiae.

		Analysis phase			Deleted			% Deleted		
Clarity	**Difficulty**	**Indiv**	**Inconc**	**Excl**	**Indiv**	**Inconc**	**Excl**	**Indiv**	**Inconc**	**Excl**
Clear	Very easy	2,030	66	223	101	0	0	5.0%	0.0%	0.0%
	Easy	8,114	216	1,145	405	1	10	5.0%	0.5%	0.9%
	Moderate	8,254	1,660	2,162	600	53	77	7.3%	3.2%	3.6%
	Difficult	2,546	1,713	859	255	146	44	10.0%	8.5%	5.1%
	Very difficult	628	803	183	69	69	13	11.0%	8.6%	7.1%
Unclear	Very easy	422	17	128	26	3	1	6.2%	*17.6%*	0.8%
	Easy	1,769	74	302	180	2	7	10.2%	2.7%	2.3%
	Moderate	2,832	661	816	328	42	28	11.6%	6.4%	3.4%
	Difficult	1,090	1,110	475	173	96	21	15.9%	8.6%	4.4%
	Very difficult	556	757	163	120	60	5	21.6%	7.9%	3.1%
Total		**28,241**	**7,077**	**6,456**	**2,257**	**472**	**206**	**8.0%**	**6.7%**	**3.2%**

Table S19: Percentage of minutiae deleted by clarity, difficulty and conclusion (n=41,774 Analysis phase minutiae among the 2,957 markups); estimates from small samples (n<50 minutiae during Analysis) shown in gray.

			% Added				
			Mates			Nonmates	
Correspond	**Clarity**	**Difficulty**	**Indiv**	**Inconc**	**Excl**	**Excl**	**Inconc**
Corresp.	Clear	Difficult	29.6%	20.5%	31.4%	15.4%	*(n=27) 11.1%*
	Clear	Not difficult	17.1%	13.1%	10.3%	3.4%	*(n=22) 31.8%*
	Unclear	Difficult	69.6%	38.8%	*(n=32) 40.6%*	*(n=17) 23.5%*	*(n=18) 16.7%*
	Unclear	Not difficult	46.2%	42.6%	*(n=25) 40.0%*	*(n=15) 26.7%*	*(n=7) 0.0%*
Not	Clear	Difficult	9.7%	8.5%	4.5%	5.5%	3.5%
	Clear	Not difficult	5.9%	5.2%	4.5%	2.5%	4.1%
	Unclear	Difficult	29.7%	17.1%	12.4%	11.3%	7.1%
	Unclear	Not difficult	11.5%	14.0%	8.7%	6.2%	1.3%
Total			**21.1%**	**14.4%**	**8.4%**	**4.6%**	**5.1%**

Table S20: Minutiae addition rates by whether corresponded, clarity, and difficulty (n=41,774 Analysis phase minutiae among the 2,957 markups). Rates calculated as (minutiae added)/(minutiae during Analysis). Single false positive omitted; estimates from small samples (n<50 minutiae during Analysis) shown in gray.

B.9 Distributions of deleted and added minutiae by examiner and by latent

Fig. S6 through Fig. S10 show the distributions of deleted and added minutiae across examiners and latents. Each histogram describes the average number of minutiae added or deleted by examiners or by latents. Much of the variability by image pair (Fig. S7) is associated with the varying proportions of determinations on each image pair; the results for true positives (Fig. S9) eliminate this source of variability.

A small proportion of the examiners account for a large proportion of the deletions. Modeling the number of deleted minutiae as a response to the image pair and examiner (modeled as random effects), we find that examiners account for 24% of the variance and image pairs account for 8% (n=3,006). Modeling the number of added minutiae as a response to the image pair and examiner (random effects), we find that together examiners and image pairs account for 40% of the variance and contribute approximately equally.

Most deletions and additions occur on mated image pairs, especially in association with individualizations (true positives). Much of the variability in deletions and additions by image pair (Fig. S7) is associated with mating (responses on nonmates are shown shaded) and examiners' determinations (cf. Fig. S9, which is limited to markups associated with true positives).

The numbers of minutiae deleted and added are correlated by examiner and also by image pair (Fig. S10).

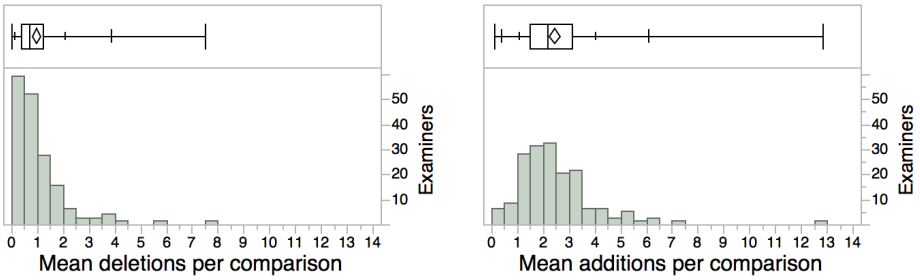


Fig. S6: Distributions of mean number of minutiae deleted and added by each examiner. Data limited to 2,957 comparisons on which 2,935 minutiae were deleted and 7,243 added. Half of all deletions were made by 32 examiners; half of all additions were made by 48 examiners. The median deletion rate across examiners was 0.7 per comparison; the median addition rate was 2.2.

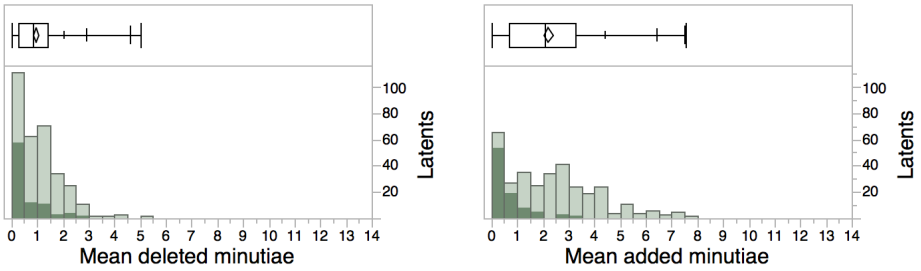


Fig. S7: Distributions of mean number of minutiae deleted and added on each latent. Rates for latents compared to nonmates are shaded; rates for the 19 latents that were paired with both mated and nonmated exemplars are calculated separately for each pairing. Data is limited to 2,957 comparisons (NV determinations excluded) on which 2,935 minutiae were deleted and 7,243 added. Half of all deletions were made on 62 latents; half of all additions were made on 69 latents. Results for nonmated image pairs are shown shaded.

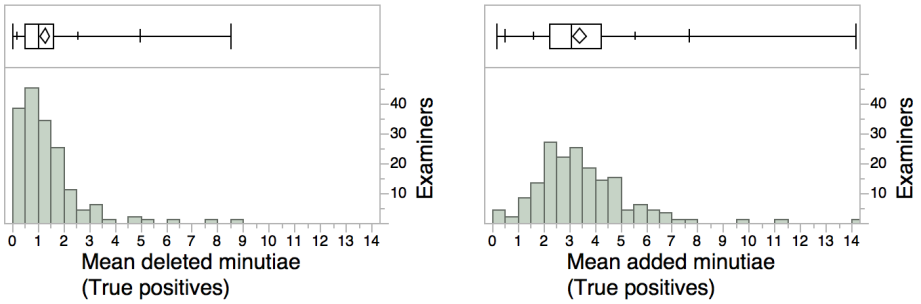


Fig. S8: Distributions of minutia deletions and additions by examiner, limited to 1,695 comparisons with true positive outcomes (individualizations of mated pairs). Based on 2,241 deletions and 5,962 additions. The median deletion rate across examiners was 1.0; the median addition rate was 3.1.

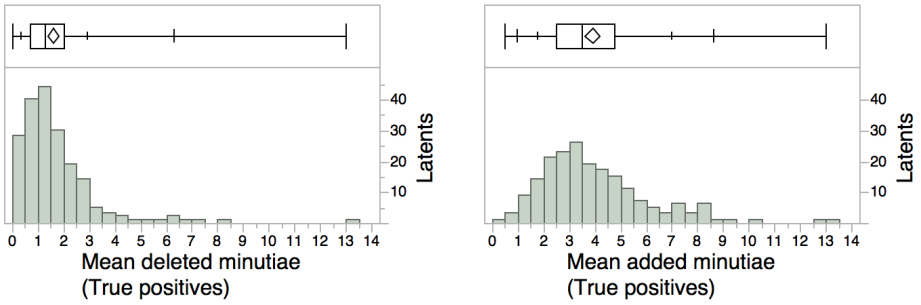
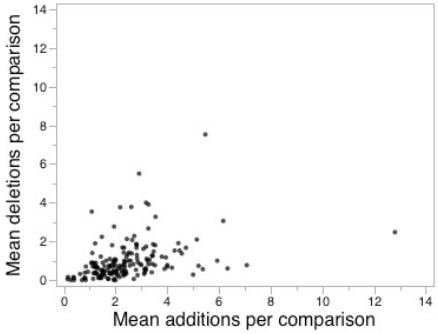


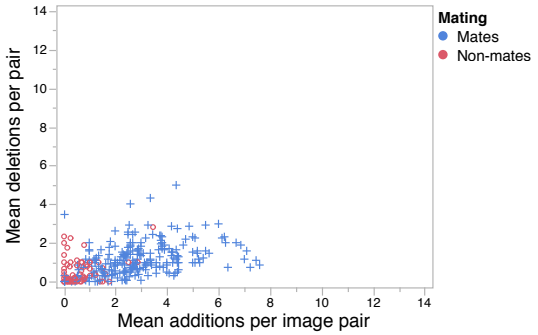
Fig. S9: Distributions of minutia deletions and additions by latent, limited to 1,695 true positive outcomes (individualizations of mated pairs). The median deletion rate across these latents was 1.3; the median addition rate was 3.5. Half of all deletions were made on 49 latents; half of all additions were made on 57 latents.

A) Delete rate by add rate (n=170 examiners)



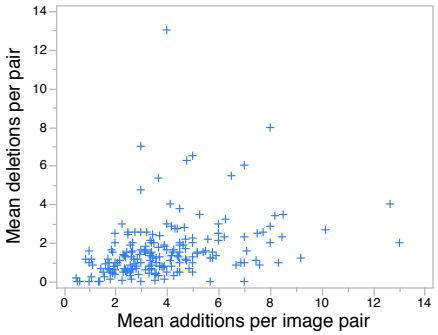
2,935 deleted minutiae; 7,244 added minutiae
2,957 comparisons of 314 image pairs

B) Delete rate by add rate (n=314 image pairs)



2,935 deleted minutiae; 7,244 added minutiae
2,957 comparisons of 314 image pairs

C) Delete rate by add rate on TPs (n=194 mated pairs)

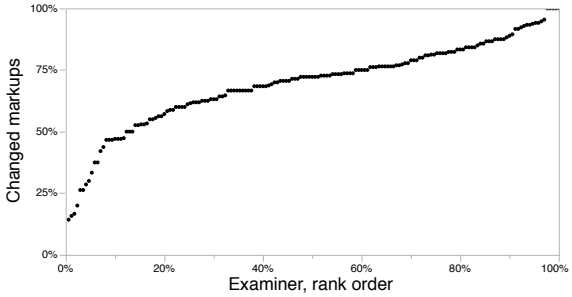


2,241 deleted minutiae; 5,962 added minutiae
True positives only: 1,695 comparisons of 194 mated image pairs

Fig. S10: Mean counts of minutiae deleted and added per comparison (A) by examiner, (B) by image pair, and (C) by image pair limited to true positive outcomes. Mated pairs shown as blue crosses; nonmates as red circles. The examiner who committed the false positive error had the highest deletion rate (7.5 minutiae per comparison). The nonmated image pair on which the false positive error occurred was compared by eleven examiners who in total deleted 31 minutiae (mean 2.8 deletions per examiner) and added 38 (mean 3.5), of which the examiner who committed the false positive error deleted 16 and added 13.

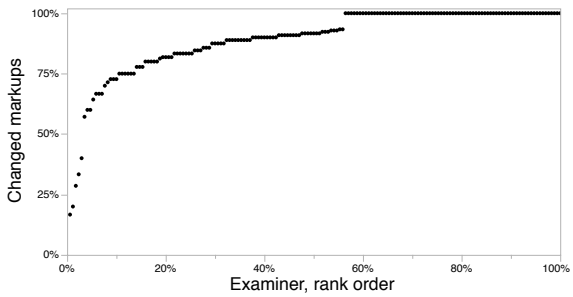
Fig. S11 shows for each examiner the proportion of comparisons in which the examiner added or deleted any minutiae. A small percentage of examiners are outliers in that they have relatively stable latent markup.

A)Markups with added or deleted minutiae, by examiner



n=170 examiners, 2,957 markups, 49,018 total minutiae

B)TP Markups with added or deleted minutiae, by examiner



n=170 examiners, 1,695 true positive markups, 34,186 total minutiae

Fig. S11: Proportion of markups with any additions or deletions, by examiner: A) for all comparisons (median of 17 comparisons per examiner); B) for individualizations of mated pairs (true positives; median of 10 true positives per examiner). All examiners added or deleted minutiae in the Comparison phase. Most examiners (86%) added or deleted minutiae in the majority of their comparisons; 97% added or deleted minutiae the majority of the time when individualizing. 44% of examiners added or deleted minutiae every time they individualized.

B.10 Mean change in minutiae counts

Examiners tended to add more minutiae during Comparison than they deleted. Fig. S12 shows the variability in this net change among individualizations. This section further describes this variability by modeling the net increase in minutiae for each comparison as a response to the image pair, the examiner, and the examiner's determinations.

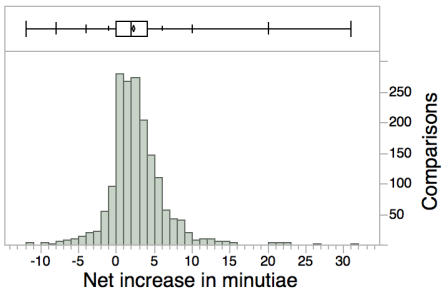


Fig. S12: Distribution of the net increase in minutiae (Added – Deleted) on individualizations (n=1,696).

We use a mixed model of the following form to describe change in minutia count as a response not only to the examiner's determinations (fixed effect), but also to the image pair and examiner (random effects):

$$Minutiae_{Comparison}[i,j,k] - Minutiae_{analysis}[i,j,k] = \beta_0 + \beta_{Determinations}[i] + \beta_{imagePair}[j] + \beta_{Examiner}[k] + \varepsilon[i,j,k],$$

where i indexes the examiner determinations, j indexes image pairs, and k indexes examiners; the betas are unknown parameters for an intercept, the examiner determinations, each image pair, and each examiner.

One purpose of constructing a model, rather than simply reporting mean changes in minutia counts, is to obtain confidence intervals for comparing the means for each of the fixed effects (determination types). Much of the uncertainty with respect to the mean is due to limited numbers of examiners and image pairs. Because of the large numbers of image pair and examiner parameters, they were analyzed as if they were random samples from populations of images pairs and examiners, respectively. This "random effects" model was

analyzed using Restricted Maximum Likelihood Estimation (REML). The implicit "larger populations" should be thought of as a generalization from the volunteer participants and the data selected for the test, not necessarily representative of specific real-world populations: the confidence intervals are provided to help describe the observed associations, not for making inferences beyond the test conditions.

Table S21 shows the percentage of variance in net increase in total minutiae attributable to each of the random effects. Some of the overall variance explained by image pair simply relates to whether or not the pairs are mated: separately, on mates and on nonmates, we see a strong subjective component as examiners account for much more of the variance than do image pairs. Subjectivity accounts for a greater percentage of the variance for nonmates than for mates. Image pairs account for almost none of the variance on nonmates: the effects of image pairs are greatest among true positives, but even here subjectivity accounts for more than twice as much of the variance as do images.

Predictor	All (n=3,006)	Nonmates (n=592)	Mates		
			All Mates (n=2,414)	TP (n=1,695)	Not TP (n=719)
Examiner	17.5%	42.8%	21.7%	24.1%	16.1%
ImagePair	11.9%	0.5%	7.9%	11.6%	3.0%
Residual	70.6%	56.8%	70.4%	64.3%	80.9%

Table S21: Variance component estimates for predicting change in minutia count. TP indicates true positives (individualizations on mated pairs).

Table S22 shows that the tendency toward a net increase in total minutiae is specifically associated with mated pairs and much more pronounced when either the examiner individualizes or the examiner compares a VEO latent. The greatest increase occurs on latents whose value determination is changed from VEO to VID. On nonmate comparisons, no significant increase was observed in association with any of the value determination pairs.

		All (n=3,006) Min_{cmp} - $Min_{Analysis}$			Mates (n=2,414) Min_{cmp} - $Min_{Analysis}$			Nonmates (n=592) Min_{cmp} - $Min_{Analysis}$		
Analysis	**Comparison**	**Mean**	**Lower95**	**Upper95**	**Mean**	**Lower95**	**Upper95**	**Mean**	**Lower95**	**Upper95**
VEO	NV	-1.08	-2.22	0.05	-1.63	-3.13	-0.14	-0.05	-1.01	0.91
	VEO	0.66	0.33	1.00	0.93	0.50	1.37	0.18	-0.08	0.44
	VID (not TP)	2.73	1.12	4.33	2.82	0.70	4.95	0.82	-0.52	2.16
	VID (TP)	3.51	2.93	4.08	3.54	2.93	4.16	--	--	--
VID	NV	-0.12	-1.34	1.10	0.05	-1.49	1.60	-0.63	-1.80	0.53
	VEO	-0.29	-1.22	0.63	0.08	-1.10	1.25	-1.77	-2.63	-0.91
	VID (not TP)	0.42	0.13	1.71	0.88	0.49	1.27	0.11	-0.16	0.27
	VID (TP)	2.01	1.75	2.26	2.03	1.74	2.31	--	--	--
VCMP	**All**	1.30	1.06	1.55	1.70	1.42	1.98	0.05	-0.14	0.25

Table S22: Associations between changed minutia counts and latent value determinations. Least squares means were estimated for each level (paired latent value determinations and whether or not Comparison determination was an individualization) by modeling examiners and latents as random effects. VID results differed depending on whether they were true positives (TP) or not.

B.11 Erroneous individualization feature modifications

The examiner who made the erroneous individualization completely revised that latent markup during Comparison: 16 minutiae were deleted, 13 added, 1 moved, and 0 retained; 1 delta was added. As shown in Fig. S13 (highlighted in red), this examiner's pattern of frequently deleting and adding minutiae was unusual; the modifications associated with the false positive were an extreme example of this unusual behavior.

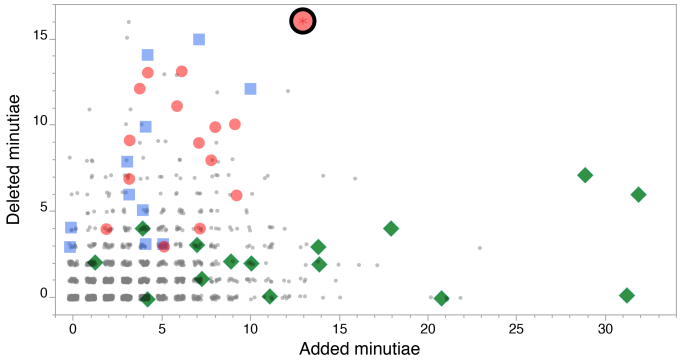


Fig. S13: Distribution of deleted by added minutiae (n=1,696 individualizations). No minutiae were deleted or added on 10% of markups; no more than two minutiae were deleted or added on 34%. Results for three examiners are color-coded (red, blue, green) to show that anomalous responses tend to associate with specific examiners. The change behavior on the false positive markup (outlined red circle) was extreme, and that examiner often deleted and added many minutiae. To maintain anonymity, the examiner IDs shown are not those assigned during the test.

B.12 Low-count individualizations

Table S23 describes the 140 individualizations having eight or fewer corresponding minutiae.[§§] The majority of these (85/140, non-highlighted cells) relied on six or fewer minutiae that had been marked during the Analysis phase. As compared to other individualizations, the percentages of changed minutiae were not notably higher or lower for low-count individualizations.

We provide further description of those individualizations with six or fewer corresponding minutiae in [2, Appendix S17]. The 140 low-count IDs were made by 73 examiners on 82 pairs. Thirty-one examiners made two or more low-count IDs; four examiners made six low-count IDs. As stated in [2, Appendix S17], we believe that the five individualizations with four or fewer corresponding minutiae were due to improper (incomplete) markup; the individualization with five corresponding minutiae also had two corresponding level-3 features. Most of these image pairs (60/82) were individualized by the majority of examiners to whom they were assigned. Typically, other examiners individualized these with a greater number of corresponding minutiae or were inconclusive.

[§§] The set of 140 low-count individualizations was identified by the number of "definitely corresponding" minutiae [9, Appendix S22]. For the purpose of this analysis, and for consistency with our previous paper, we have omitted the responses of five examiners who did not annotate correspondences.

	Added during Comparison phase								
CMin	**0**	**1**	**2**	**3**	**4**	**5**	**7**	**8**	Total IDs
0	2	-	-	-	-	-	-	-	2
2	1	-	-	-	-	-	-	-	1
4	-	2	-	-	-	-	-	-	2
5	1	-	-	-	-	-	-	-	1
6	5	3	1	1	-	-	-	-	10
7	18	17	5	3	-	1	3	-	47
8	20	17	15	12	5	5	1	2	77
Total	47	39	21	16	5	6	4	2	140

Table S23: Low-count individualizations: corresponding minutiae added during Comparison. Cell counts indicate the number of individualizations for a given total corresponding minutiae count (rows) and number of those corresponding minutiae that were added during comparison (columns). For example, 15 IDs were made with eight corresponding minutiae, where two of those minutiae were added during Comparison.

B.13 Effects of changed markup on interexaminer consistency

We measured interexaminer consistency by estimating the proportion of examiners who marked each feature. We evaluated the results by comparing the distributions of these proportions, which range from 0% (only one examiner marked the feature) to 100% (every examiner marked the feature).

To estimate agreement, for each marked feature we first determined the proportion of other examiners who marked within 0.5 mm of that location. To summarize the results as an overall distribution, we applied a weight to each estimated proportion: 1/(number of examiners who marked near that location). The purpose of this weighting is to count each feature equally, rather than counting each feature marking equally; this technique accomplishes that to a good approximation. For example, if an examiner marked a feature and no other examiners marked within 0.5 mm, then the proportion is zero and the weight is 1. On the other hand, if six of ten other examiners marked within 0.5 mm of the first examiner's mark, then the proportion is 6/10 and the weight is 1/7. In the latter example, each of the 7 examiners who marked this feature contributes with a weight of 1/7 with the result that the entire group of 7 examiners contributes the 6/10 proportion once for this feature. This weighting method is inexact because nearness is not a transitive property: examiners who mark a feature precisely may have a higher proportion of agreement than those who mark a feature somewhat away from the center of the cluster of markings.

Fig. S14 describes interexaminer agreement on features marked in Clear areas. The proportion of features marked by only one examiner (far left) drops from 20.2% (Analysis-phase markups) to 16.7% (Comparison-phase markups), while the proportion of features that were not marked by every examiner (far right) decreases from 82.7% to 76.6%. That is, examiners unanimously marked 17.3% of features during Analysis, and 23.4% after Comparison.

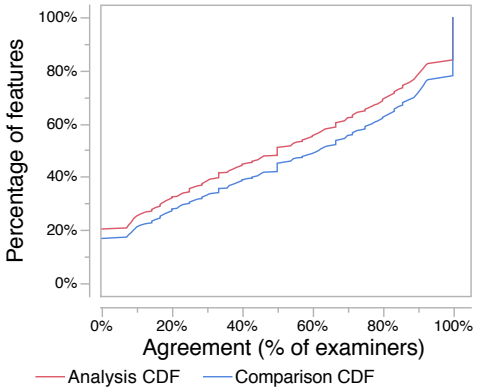


Fig. S14: Cumulative distribution functions describing agreement on the presence of features across examiners, limited to markups of mated image pairs by examiners who individualized, and to features in Clear areas as determined by the Comparison-phase median clarity maps.

B.14 Changes in latent value determinations

The VEO individualization rate was much higher on this study than on our previous Black Box study (26% vs. 1.8%). Because the two studies were not specifically designed to investigate this unexpected result, this presented several analysis challenges that we address here.

By including only image pairs that were in both studies, Table S24 shows that the difference in individualization rates between the two tests remains after we control for differences in data selection. We also note that the VID individualization rates were not significantly different on the two tests.

Mating	Latent Analysis Value	Black Box Responses	Black Box IDs	White Box Responses	White Box IDs
Mates	NV	36	0	38	0
	VEO	106	3	62	15
	VID	641	493	317	264
Nonmates	NV	293	0	159	0
	VEO	360	0	139	0
	VID	599	0	184	1
Total		**2,035**		**899**	

Table S24: Comparison of Black Box and White Box determinations on a common set of 83 image pairs used in both tests. The assignments of image pairs to examiners differed on the two tests. We assigned each image pair to more examiners on Black Box than White Box. However, we assigned mated image pairs proportionally more often in White Box.

Examiners changed VEO determinations to VID much more often when comparing the latent to a mated exemplar than when comparing to a nonmated exemplar (Table 1). However, when comparing outcomes on mated and nonmated pairs, it is important to recognize that we selected these two sets of latents by different procedures, resulting in characteristic population differences. For example, the mated latents used in this study tended to have higher minutia counts than the nonmated latents (Fig. S15). Value determinations on latents with very high minutia counts (or very low) are unlikely to change from Analysis to Comparison.

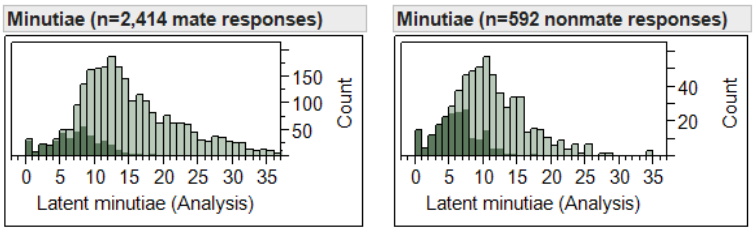


Fig. S15: Analysis-phase distributions of total minutia counts for (A) mated and (B) nonmated latents that were compared (n=3,006). Minutiae counts associated with VEO determinations are shaded. Among compared latents, we generally expect those with very high minutiae counts (e.g., greater than 20) to be associated with more stable value determinations than those with low counts.

Data was available to address the concern over differences in how we selected latents for mated and nonmated pairs. The White Box test included 19 latents each of which was paired with both a mated and a nonmated exemplar for a total of 38 image pairs. These 19 latents yielded 402 Comparison responses. By comparing the results on this subset (Table S25B), we can investigate the effects of exemplar mating on changed value determinations without confounding due to population differences in the selected mated and nonmated latents. As discussed below, we observe the same general patterns of value changes on this subset as on the test overall.

WB: all data

Analysis	Cmp	Mates	Nonmates	Total
NV	NV	457	246	703
VEO	NV	15	8	23
VEO	VEO	279	175	454
VEO	VID	110	4	114
VID	NV	14	6	20
VID	VEO	25	11	36
VID	VID	1,971	38	2,359
Compared		*2,414*	*592*	*3,006*
Total		*2,871*	*488*	*3,709*

WB: subset of 19 latents

Analysis	Cmp	Mates	Nonmates	Total
NV	NV	45	42	87
VEO	NV	1	1	2
VEO	VEO	39	38	77
VEO	VID	11	0	11
VID	NV	0	2	2
VID	VEO	2	2	4
VID	VID	126	93	219
Compared		*179*	*136*	*315*
Total		*224*	*178*	*402*

Table S25: Counts of Analysis- and Comparison-phase latent value determinations: (A) 3,709 paired determinations of 301 latents; (B) subset 402 paired determinations of 19 latents.

While the instructions for the VEO category were the same in both Black Box and White box, the test procedures were different, which may have had an effect on the usage of the VEO category. In Black Box, the examiners were asked first whether the latent was of value for individualization; only if not VID were they asked if it was of value for exclusion only. In White Box, a single menu included the options "Value for ID", "Limited value", and "No value." Since White Box did not explicitly require the examiners to indicate that the latent was not VID in selecting VEO, and since White Box used the abbreviation label "Limited" (as defined in the ANSI/NIST standard [3]), examiners' usage of VEO could have differed between the studies; however, the VEO rates were almost identical between the two studies. When limited to latents that were used in both studies, examiners in White Box were less likely to assess latents as VID, more likely to assess NV, and equally likely to individualize (Fig. S16). These charts show that the greatest shift in the percentage of examiners assessing latents to be VID occurs on those latents where the Black Box assessments had the lowest reproducibility; unanimous Black Box assessments were generally reproduced on White Box. As we reported previously [4], VEO assessments are not highly reproducible.

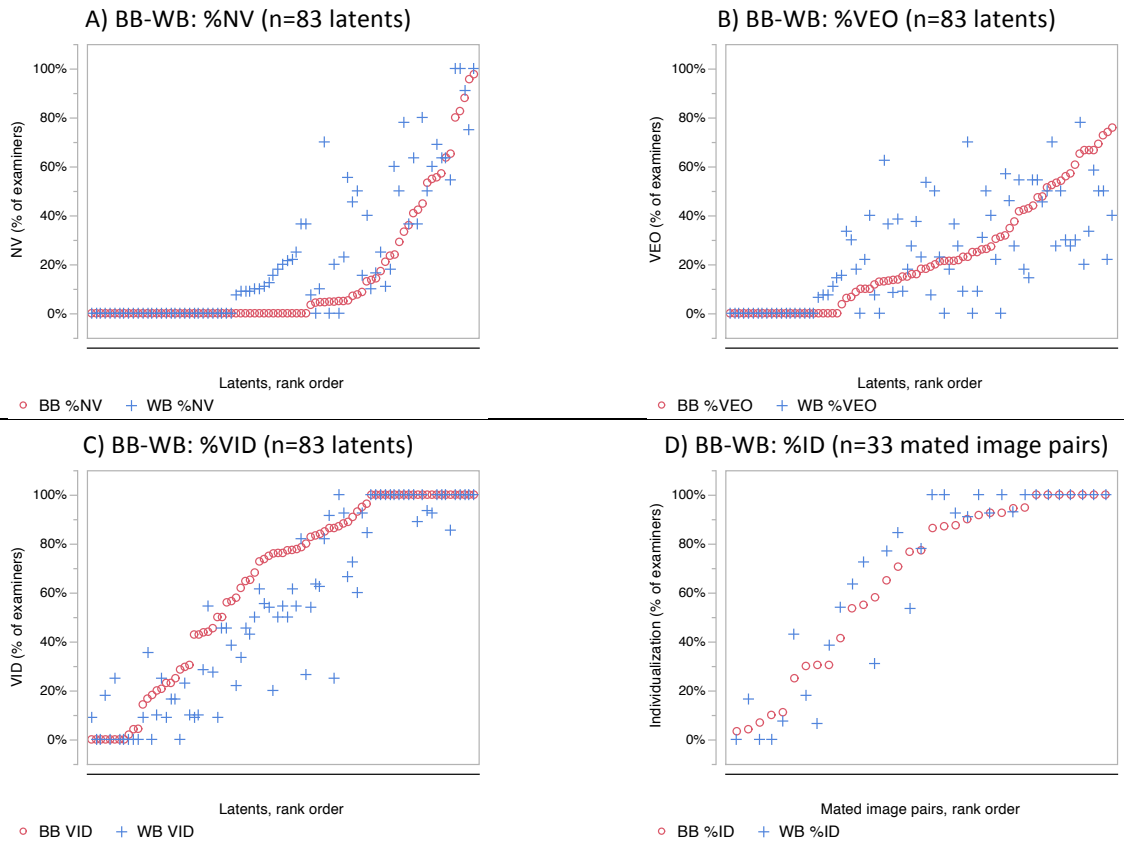


Fig. S16: Examiner determinations on a common set of 83 latents used in both tests, and the 33 mated image pairs used in both tests. Latents (and mated image pairs) are shown sorted in rank order according to the percentages of examiner determinations on the Black Box test. For these latents, mean %NV was 14.6% (Black Box) vs. 23.2% (White Box); mean %VID was 63.8% (Black Box) vs. 53.6% (White Box). These 83 latents were assigned to examiners 2,035 times in Black Box, and 898 times in White Box (mean 25 examiners per latent in Black Box; 11 in White Box).

Unsurprisingly, after further consideration during Comparison examiners sometimes changed their latent value determinations [4]. To better understand this, we formulated several testable hypotheses. Were the changes to value determinations symmetric, i.e., among those determinations that were borderline VEO/VID, did an equal number change in each direction? If the changes were not symmetric, did the value determinations generally tend to increase (or decrease) or was the direction influenced by whether the exemplar was mated or not? If there was bias in the direction of change, can it be attributed to data selection or might it be associated with the values themselves? Was the rate of change influenced by whether the exemplar was mated or not? Lastly, how large are any observed effects?

We use McNemar's test of marginal homogeneity (chi-squared test) to investigate symmetry along the off-diagonal of an Analysis phase by Comparison phase 2x2 matrix of value determination counts. Our null hypothesis is Proportion(1st response is VID) = Proportion(2nd response is VID). We apply this test separately to the mated and nonmated latents. Because examiners were not asked to compare latents that they assessed to be NV, changes from NV to VEO and NV to VID are treated as missing. In order to test for symmetry, we must therefore exclude the corresponding changes from VID to NV and from VEO to NV. McNemar's test on mated latents indicates asymmetry ($p<0.0001$) with 25 changes from VID to VEO and 110 changes from VEO to VID (see Table S25A). The result for nonmated latents is less compelling ($p=0.1185$), but appears to be asymmetric in the opposite direction with 11 changes from VID to VEO and 4 changes from VEO to VID.

Table S26 shows estimates and test statistics used to analyze directional bias in latent value determinations. The notation P(up|mate) denotes the proportion of latent value determinations that increased from VEO to VID from Analysis to Comparison among mated pairs; P(down) refers only to changes from VID to VEO as symmetric data was not available for NV determinations. As previously discussed, we used McNemar's test to evaluate asymmetry; in this case, comparing P(Up) to P(Down) is equivalent to comparing $P(Value_{Analysis} = VID)$ to $P(Value_{Comparison} = VID)$. We use relative risk confidence intervals to compare proportions (by checking whether the interval contains one).

We report exact binomial confidence intervals and the p-values for comparative purposes only. Such statistics presume an inferential framework, but no claims are made as to the representativeness of the sample data.

	All comparisons		Comparisons of 19 latents	
Estimates	**Counts**	**Confidence**	**Counts**	**Confidence**
P(Up|mate)	110/2,414 = 4.6%	95% CI: 3.8% to 5.5%	11/179 = 6.1%	95% CI: 3.1% to 10.7%
P(Up|nonmate)	4/592 = 0.7%	95% CI: 0.2% to 1.7%	0/136 = 0.0%	95% CI: 0.0% to 2.7%
P(Down|mate)	25/2,414 = 1.0%	95% CI: 0.7% to 1.5%	2/136 = 1.5%	95% CI: 1.2% to 8.4%
P(Down|nonmate)	11/592 = 1.9%	95% CI: 0.9% to 3.3%	2/179 = 1.1%	95% CI: 0.3% to 4.8%
Tests				
P(Down|mate) < P(Up|mate)	25 < 110	p < 0.0001 (McNemar)	2 < 11	p = 0.0225 (McNemar)
P(Down|nonmate) > P(Up|nonmate)	11 > 4	p = 0.1185 (McNemar)	2 > 0	p = 0.5000 (McNemar)
P(Up|Nonmate) / P(Up|mate)	0.7% / 4.6% = 0.1	95% RR CI: 0.05 to 0.4	0.0% / 6.1% = 0.0	95% RR CI: 0.0 to 0.6
P(Down|nonmate) / P(Down|mate)	1.9% / 1.0% = 1.8	95% RR CI: 0.8 to 3.8	1.1% / 1.5% = 0.8	95% RR CI: 0.8 to 7.5

Table S26: Test statistics for changed value determinations, by mating: 95% exact binomial confidence intervals for proportions (rows 1-4); McNemar's test of marginal homogeneity (two-sided, exact binomial test; rows 6-7); and exact binomial 95% confidence intervals for relative risk (rows 7-8).

From Table S26 we see that latent value changes differ for mates and nonmates: examiners more often increased their value assessments when comparing mates, and more often decreased their assessments when comparing nonmates. The results of McNemar's tests clearly show this asymmetry on the mates, but were not statistically significant for nonmated comparisons. Since the results for the 19 latents presented in both mated and nonmated pairs agreed in direction with the larger dataset, the differences cannot be explained by differences in how the latent prints were selected for mated and nonmated pairs.

We conclude that the striking differences in the rates of VEO individualizations between the two tests can be attributed primarily to whether or not the examiners provided markup of the prints. The difference in VEO individualization rates remains after controlling for differences in data selection, and other factors such as participants cannot account for the large difference observed. The tendency to increase value assessments is associated with those comparisons where examiners saw corresponding features and the spent additional time studying and documenting the similarities. The asymmetries in changed value determinations (by mating) suggest that examiners make the changes when motivated to support their comparison conclusions, but tend to leave the Analysis determination unchanged when no such motivation is present.

B.15 Comparison with Black Box Repeatability Results

Table S27 shows both Black Box repeatability results [4] and White Box changes to value determinations, on a common set of latents. Black Box within-test repeatability describes two Analysis-phase responses (by the same examiner) to the same latent presented twice within one test (events occurring hours or days apart); across-test repeatability describes two Analysis-phase responses to the same latent presented on two tests (months apart). These comparisons are made on the subset of latents that were common to both tests.

The Black Box repeatability study [4] found that determinations of VEO were much less likely to be repeated than either VID or NV determinations. Here we see that examiners often changed VEO determinations during Comparison (knowing that they were changing their determinations). Whereas the Black Box results indicated that the value determinations for these borderline latents were not of great consequence because those prints were unlikely to be individualized even if compared, these new results indicate that under suitable conditions examiners will individualize many of these borderline latents. Taken together, these

results suggest that borderline latents often are not compared and yet under suitable conditions (or performed by another examiner) such comparisons might result in individualizations.

Black Box within-test repeatability					
	NV	**VEO**	**VID**	**Total**	**Changed**
NV	64	8	2	74	14%
VEO	10	53	16	79	33%
VID	3	19	196	218	10%

Black Box across-test repeatability					
	NV	**VEO**	**VID**	**Total**	**Changed**
NV	44	11	3	58	24%
VEO	12	46	24	82	44%
VID	5	23	210	238	12%

TOTAL White Box Analysis to Comparison					
	NV	**VEO**	**VID**	**Total**	**Changed**
NV	197	N/A	N/A	197	N/A
VEO	7	175	19	201	13%
VID	3	6	492	501	2%

MATES White Box Analysis to Comparison					
	NV	**VEO**	**VID**	**Total**	**Changed**
NV	38	N/A	N/A	38	N/A
VEO	2	44	16	62	29%
VID	0	1	316	317	0%

NONMATES White Box Analysis to Comparison					
	NV	**VEO**	**VID**	**Total**	**Changed**
NV	159	N/A	N/A	159	N/A
VEO	5	131	3	139	6%
VID	3	5	176	184	4%

Table S27: Comparison of Black Box and White Box results on a common set of 68 latents. The Black Box and White Box tests included 83 common image pairs; among these were 68 latents used in the Black Box repeatability tests.

Appendix B References

1 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)

2 Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2014) Measuring what latent fingerprint examiners consider sufficient information for individualization determinations. PLoS ONE 9(11): e110179. doi:10.1371/journal.pone.0110179 (http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0110179)

3 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)

4 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)