IMPROVING THE RIGOR OF THE LATENT PRINT EXAMINATION PROCESS

Hicklin Austin R.

Université de Lausanne

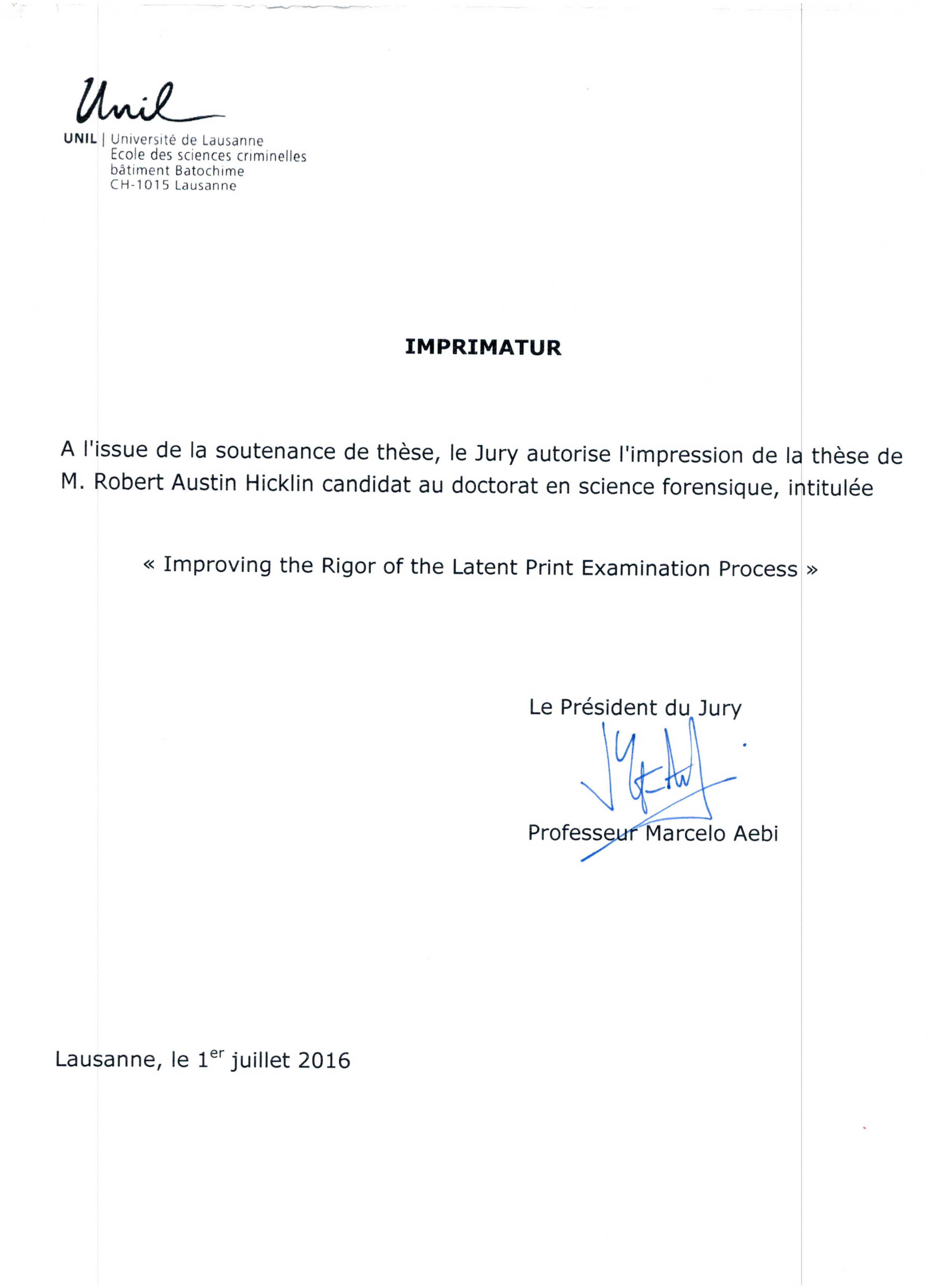
Faculté de droit, des sciences criminelles et d'administration publique
Ecole des Sciences criminelles

Improving the Rigor of the Latent Print Examination Process

Thèse de Doctorat
en Science Forensique

R. Austin Hicklin

5 February 2017



UNIL | Université de Lausanne
Ecole des sciences criminelles
bâtiment Batochime
CH-1015 Lausanne

IMPRIMATUR

A l'issue de la soutenance de thèse, le Jury autorise l'impression de la thèse de M. Robert Austin Hicklin candidat au doctorat en science forensique, intitulée

« Improving the Rigor of the Latent Print Examination Process »

Le Président du Jury

Professeur Marcelo Aebi

Lausanne, le 1er juillet 2016

Abstract

This PhD thesis is a synthesis of a portfolio of interrelated previously published work that was conducted to improve the rigor, standardization, transparency, and quantifiability of the latent print examination process. The core of the work relates to the development, adoption, and implications of the *Extended Feature Set* (EFS). EFS is a formal international standard (incorporated in ANSI/NIST-ITL) that defines a method of characterizing the information content of friction ridge impressions — allowing latent print examiners to unambiguously document the bases of their determinations during examination. EFS is the enabling technology that has made all of the other elements of this portfolio of work possible: evaluations of the accuracy and reliability of latent print examiners' determinations, evaluations of the reliability of examiners' feature markup, evaluations of examiners' assessments of sufficiency, evaluations of latent print quality, development of quality and distortion metrics, evaluations of AFIS accuracy, and the development of training materials to assist in improving the uniformity of examiners' annotations of the features and attributes of friction ridge impressions. The thesis summarizes these previous publications, as well as discussing their implications and possible future research and tools that could leverage this body of work.

Résumé

Cette recherche doctorale présente la synthèse d'un portfolio de travaux et de publications ayant pour objectif d'améliorer la rigueur, la standardisation, la transparence et la quantification dans le cadre du processus d'identification des traces papillaires. L'élément fondateur de cette recherche est le développement, l'adoption et les implications du *Extended Feature Set* (EFS). EFS est un standard formel international (incorporé dans ANSI/NIST-ITL) qui définit la méthode de description des caractéristiques présentes sur les impressions papillaires. Il permet aux experts en lophoscopie de documenter de manière non-ambiguë les observations qui sont à la base des conclusions formulées à la suite des examens. EFS a été le facilitateur qui a rendu possible tous les autres éléments de ce portfolio de recherches, à savoir : l'évaluation de la fiabilité et l'exactitude des conclusions des experts en matière de traces papillaires, l'évaluation de la fidélité des annotations des experts, le développement de mesures de qualité et de la distorsion des traces, l'évaluation de l'exactitude des systèmes AFIS et finalement le développement d'une formation visant à améliorer la reproductibilité, entre experts, des annotations des caractéristiques papillaires et de leurs attributs. Cette recherche doctorale présente une synthèse de l'ensemble de ces travaux publiés et discute des implications de ceux-ci, des voies de recherche future ainsi que les outils qui pourraient y être associés.

Table of Contents

Acknowledgments

I would like to thank my committee — Cedric Neumann, Pierre Margot, and especially Christophe Champod — for their insights and assistance in making this possible.

In the body of work discussed here I was privileged to work with an incredible group of colleagues. All of the work discussed here was performed as part of different teams, with colleagues from Noblis, the FBI, NIST, and other organizations. In each of these tasks I was largely responsible for initiating and defining the work, as well as performing analyses, writing for publication, and being responsible for the end results — but I do not claim or wish to imply that I am solely responsible for this work. I would like to thank my colleagues, coauthors, and fellow researchers, without whom this work would have been impossible: Brad Ulery, JoAnn Buscaglia, Toni Roberts, Steve Meagher, Mike Indovina, Nate Kalka, Ted Unnikumaran, Mike Beachler, George Kiebuzinski, Brad Wing, Scott Swann, Melissa Taylor, Bill Fellner, Matt Schwarz, Peter Komarinski, Erik Stanford, Rachel Wallner, Calvin Yeung, Kristen Allen, Celeste Descoteaux, Mark Burge, Vlad Dvornychenko, Will Chapman, Ron Smith, Larry Pantzer, Kerry-Ann White, Corky Korves, Matt Monaco, John Mayer-Splain, and David Vera. All co-authors were given the formal opportunity to review these acknowledgments, and all who responded concurred; none objected.

I also thank the Committee to Define an Extended Fingerprint Feature Set (CDEFFS) for their work in defining the Extended Feature Set: Behnam Bavarian, Vincent Bouatou, John Burt, Christophe Champod, Yi Chen, Vladimir Dvornychenko, Jeri Eaton, Brian Finegold, Jean-Christophe Fondeur, Mike Garris, Ed German, Mike Gilchrist, Paul Griffin, Masanori Hara, Peter Higgins, Tom Hopper, Anil Jain, Creed Jones, Artour Karaguiozian, Peter Komarinski, Debbie Leben, Bill Long, Davide Maltoni, Dana Marohn, Brian Martin, John Mayer-Splain, Mike McCabe, Glen McNeil, Steve Meagher, Dmitry Mikhailov, Elaine Newton, Afzel Noore, Geppy Parziale, Wade Petroka, Ann Punter, Richa Singh, Ron Smith, Greg Soltis, Matt Schwarz, Scott Swann, Elham Tabassi, Cedric Thuillier, Anne Wang, Phillip Wasserman, Kasey Wertheim, Brian Wong, and Stephen Wood.

While the preparation of this thesis was conducted on my own time, the work discussed here was conducted as part of work as a Fellow at Noblis (formerly named Mitretek Systems), a non-profit research, science, technology, and strategy organization. I would like to thank Amr El-Sawy, Roger Mason, and Larry Pantzer for allowing me to take time away to complete this work.

The work was funded under contract awards to Noblis from the FBI CJIS Division, the FBI Biometric Center of Excellence, the FBI Laboratory Division, and the NIST Law Enforcement Standards Office. I am very grateful to the sponsors of this work, especially the individuals who made it possible: JoAnn Buscaglia, Toni Roberts, Scott Swann, Melissa Taylor, Nick Megna, Justin Smith, Charlie Wilson, Steve Meagher, and Nathan Haught.

This thesis describes and builds upon a series of studies and tasks. Here is a summary of my role in these tasks:

- The *Extended Feature Set (EFS)* was developed by the CDEFFS committee; I served as chair, wrote the majority of the draft specification documents, led workshops, coordinated comments, and presented our findings at conferences. My participation was funded by NIST ITL and FBI CJIS. The bulk of the work of the CDEFFS committee itself was in 2005-2009, but I had an extensive role in incorporating EFS into ANSI/NIST-ITL through 2011.
- The *Latent Quality Study* was conducted by a Noblis/FBI Lab team, under funding from the FBI Laboratory and the FBI Biometric Center of Excellence. I had a significant role in leading the work, developing the initial concepts, designing the study, developing software, performing analysis, and writing the publications. The bulk of the work was done in 2007-2008.
- The *Black Box*, *Black Box Repeatability*, *Sufficiency for Value*, *Sufficiency for Individualization, Analysis to Comparison,* and *Interexaminer Variation in Minutia Markup* studies were conducted

from 2008 through the present by a four-person Noblis/FBI Lab team: Brad Ulery, JoAnn Buscaglia, Toni Roberts and myself. The work was truly a team effort, in which I had a significant role in designing, guiding, conducting, and writing the studies and resulting publications. Noblis participation was funded by the FBI Biometric Center of Excellence and the FBI Laboratory.

- *EFS Markup Instructions*, *EFS Profiles*, and *LITS* were developed by a Noblis team under funding from the NIST Law Enforcement Standards Office. The bulk of the work was done in 2011-2012. I came up with the initial concepts and the first drafts of each document, but development of each was a team effort, led by George Kiebuzinski.
- *ELFT-EFS* was performed by a NIST-ITL/Noblis team: Mike Indovina and I did most of the experimental design, analysis, and writing the report; Mike Indovina ran all of the data. My participation was funded by FBI CJIS. The work was done in 2008-2012.
- The *Distortion Study* was performed by Nate Kalka and myself, under funding from the FBI Biometric Center of Excellence. I developed the initial concept and provided guidance, but Nate Kalka did the bulk of the analyses and writing of the report. The bulk of the work was done in 2012-2013.

I have been involved in other, related tasks that I mention in passing, but which are not central to this thesis:

- The *Universal Latent Workstation (ULW)* has been designed, developed, and maintained by Noblis under contract to FBI CJIS from 1998 through the present. I was initial designer and developer, and continue to provide guidance to the ULW team, which is currently led by Ted Unnikumaran.
- The *EFS Training Tool* was developed by a Noblis team under funding from the NIST Law Enforcement Standards Office. The bulk of the work was done in 2012-2013. I had a significant role in developing the initial concepts and providing guidance to the team, but development was a team effort.
- The *Latent Quality Metric* project was conducted by a Noblis team (primarily Nate Kalka, Mike Beachler, and myself) under contract to FBI CJIS, from 2012 through 2014. I was responsible for developing concepts and providing guidance to the team, which was led by Nate Kalka.
- The *NGI Latent Best Practices* project was conducted by a Noblis team (primarily Nate Kalka, Mike Beachler, and myself) under contract to FBI CJIS, from 2012 through 2014. I was responsible for developing concepts and providing guidance to the team, which was led by Nate Kalka.
- *ACEware* is being developed by a Noblis team under a grant from the National Institute of Justice (NIJ). I am responsible for developing concepts and providing guidance to the team.

I was not involved in the writing or reviewing of the President's Council of Advisors on Science and Technology (PCAST) report (discussed in Section 2.3). I participated in a small workshop in October 2015 held on behalf of PCAST, chaired by Eric Lander (PCAST Co-Chair) and Jennifer Mnookin (Dean of UCLA Law School), and a followup teleconference in May 2016.

As this thesis is a synthesis of previous work, I do include some text in this thesis verbatim from my previous publications without specific attribution.

The views expressed are my own and do not necessarily reflect the official policy or position of any agency of the U.S. Government, or of the Organization of Scientific Area Committees (OSAC). This document was approved for publication by the FBI.

Chapter 1 Introduction

Forensic latent[1] fingerprint and palmprint examination is at a critical juncture. In the past few years fingerprint identification systems have increased in size and accuracy, and forensic use of fingerprints has expanded its usage beyond criminal justice into military, counterterrorism, and intelligence uses. At the same time, there have been hundreds of legal challenges to the admissibility of fingerprints as evidence in court,[2] high-profile errors,[3] and scathing criticisms of forensic science (including fingerprints).[4] Investigations are impeded because automated fingerprint identification systems (AFIS) at state, local, and federal agencies across the US and around the world have little or no interoperability,[5] and state and local agencies take relatively little advantage of the nationwide latent print system in the US.[6] Latent print evidence is often unused in "minor" crimes, even though studies indicate that processing property crimes is an effective means of identifying individuals potentially associated with major crimes.[7,8] Forensic laboratories must simultaneously adapt to new technology and new uses of fingerprints while addressing legal challenges and attempting to overcome backlogs. The forensic latent print discipline requires tools, processes, and procedures to address legal issues, provide standardization and transparency to the latent print examination process, and make effective and efficient use of new technology — in short, to improve the rigor of current latent print examination processes.

Since 2005 I have been involved in developing a portfolio of work that includes a variety of previously published journal articles and government reports,[9] addressing development of standards, scientific research and analysis, and technical evaluations of commercial systems. These are not separate projects: from the outset I proposed and developed these as interrelated facets of a concerted whole, focused on the problem of increasing the rigor of latent print examination. This thesis is a synthesis of this portfolio of previously published work, providing a summary as well as discussing implications and possible future research and tools that could leverage this body of work.

In all of the work here I had key roles in conceptualizing, designing, conducting, and being responsible for the end results — but all of these projects were team efforts, and I do not claim or wish to imply that I am solely responsible for this work. This thesis goes beyond the conclusions of these projects and includes my perspective on purpose, context, and implications; please note that this thesis represents my point of view, and does not necessarily represent those of my coauthors or my sponsors.

[1] *For discussion of terminology, see Section 1.3.*

[2] *[German05]*

[3] *e.g. [Mayfield06, Cole05, McKie11]*

[4] *e.g. [NRC09], [Saks05]*

[5] *[NSTC15]*

[6] *Although it varies from agency to agency, state and local agencies only search a few percent of their casework against the FBI's NGI system.*

[7] *Personal communication, Ken Moses (San Francisco Police Department, retired), describing an unpublished 1990s study in which it was shown that increased processing of latent prints at breaking and entering crime scenes resulted in reduction of the serious crime rate.*

[8] *[Roman08, Peterson10]*

[9] *The publications and reports used as the basis for this thesis are listed and summarized in Section 1.2.*

The work I am discussing here is not being done in isolation. Over the past ten to fifteen years — in response to Daubert challenges, the Mayfield misidentification, and the National Research Council's scathing report on forensic science — the discipline has been undergoing a sea change, reevaluating its basic tenets with the goal of improving the scientific basis for latent print examination. I have been proud to be part of that process.

1.1 Rigor in the latent print examination process

Latent print examiners are careful and conscientious in their work. After years of working in the field, it is clear to me that latent print examination as currently practiced is reasonably effective, accurate, and reliable. However, for latent print examination to be truly rigorous, it must demonstrate that it is both accurate and reliable, and have procedures in place to ensure that. I suggest that current practice is often not sufficiently rigorous given its critical importance: I believe that it requires greater ***transparency***, ***standardization***, and ***quantifiability***.

Transparency

A conclusion made by an examiner is *ipse dixit* ("he himself said it" — an unproved assertion) unless the examiner can also delineate the basis for that decision. But conclusions often have limited or no documentation, and what documentation is conducted is often not precise: descriptions of the bases for decisions are limited because there have not been standard, detailed methods for defining and documenting latent print examination. The result is a lack of transparency, as the actions and inner workings of the decision process are neither visible nor accessible. According to Champod et al, "The process by which the expert arrives at an opinion is ultimately obscure. The process relies undoubtedly on extensive and reliable experience, but it is not fully articulated."[10] The examination process involves many micro-decisions including which features were considered and accepted or rejected, the extent of similarities and differences, how features (and configurations of features) were weighted in making conclusions, the difficulty and level of confidence in the conclusion — none of which is conveyed by a categorical conclusion. Greater transparency would provide a provable basis for trust in the discipline, and a response to requests to "show me the evidence."

Standardization

Transparency only addresses part of the problem. The latent print examination process is not standardized across the community — nor are vocabulary and semantics. It is difficult to claim that the examination process is precise given that agencies vary regarding the terminology and implications of conclusions, and given that documentation, when practiced at all, is often cursory, and never in a common format.

In the United States, the organization with the role of defining standards for latent print examination has been the Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST),[11] which is currently being replaced by the Organization of Scientific Area Committees (OSAC) Friction Ridge Subcommittee. SWGFAST has been very successful in developing a variety of documents governing the examination process. However, I believe that most of the SWGFAST standards are more appropriately considered guidelines or best practices. A formal standard defines requirements and specifications

[10] *[Champod04], p 32.*

[11] *Disclosure: I have been a member of SWGFAST since 2009, and am on the Organization of Scientific Area Committees' Forensic Science Standards Board, as the chair of the Physics and Pattern Evidence Scientific Area Committee.*

completely and precisely enough to ensure that it can be used consistently,[12] and can be evaluated through conformance testing to ensure that the requirements have been met. The SWGFAST documents were generally not defined completely or precisely enough to ensure consistent use or to enable conformance testing. In addition, there is no mechanism to enforce the implementation of SWGFAST standards, leading to scattered adoption among agencies: a meaningful standard requires not just a document, but uniform usage across the community. OSAC hopes to improve this situation.

I believe that the rigor of the latent print examination process would be greatly enhanced by scientifically-based consensus and uniform usage regarding the definitions of determinations, the definitions of features, and a common electronic exchange format enabling a standard for detailed documentation of latent examination. I concur with Champod et al when they say, "Quality will be achieved by the publication and endorsement of transparent and detailed procedures describing the identification process and associated quality assurance measures."[13]

Quantifiability

Rigorous processes require validation and should make quantitative analysis possible. Some aspects of friction ridge content are difficult to represent, and the process is difficult to model, but that does not mean the entire process needs to be treated holistically, as it generally is today. Quantifiable representations and models can be effective tools in describing and allowing review of the process even if they do not attempt to address the entirety of the examination process (as George Box said, "all models are wrong, but some are useful"[14]). It is important to understand and not misrepresent the limitations of any representation or model: for example, the "apparent transparency" of point-based standards for sufficiency "was an illusion", because there was no common standard of what constituted a point[15]; using an oversimplified or inappropriate representation or model and treating the results as fact would not be an improvement on the status quo.

The current latent print examination process (at least in the US) is holistic and qualitative. The process is so unspecified that having an examiner explain to other human beings the thought process by which a decision was reached is a painstaking process; the report on the Mayfield misidentification shows a particularly problematic example.[16] Given that we live in the 21st century, processes that do not lend themselves to machine readable representations and automated analysis seem archaic. I believe that the rigor of the current examination process needs to be enhanced in order to be more effective and efficient, to address legal issues, and to interact effectively with technology. As long as examiners make determinations, there needs to be a transparent, standard, and quantifiable means of documenting and communicating how they make these decisions.

[12] *ISO definition of standard (http://www.iso.org/iso/home/standards.htm): "A standard is a document that provides requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for their purpose."*

[13] *[Champod04], p27*

[14] *[Box87], p424.*

[15] *[Champod04], p32*

[16] *[Mayfield06]*

1.2 Overview of thesis and summary of previous work

Criticisms and limitations of the current examination process..Chapter 2

This chapter discusses those aspects of the latent print examination process as it is currently practiced that I believe could most benefit from improvements in rigor, transparency, standardization, and quantifiability. I do not attempt to provide an overview of the latent examination process as a whole: for that I refer the reader to the variety of overviews already available, such as [Ashbaugh99] or [FPSourceBook].

PREVIOUS WORK..Chapter 3 - Chapter 7

These chapters provide a summary of each of the elements of this portfolio, explaining how each has made progress toward the goal of improving the rigor of the latent print examination process, and summarizing the purposes, results, and impact of each. The thesis itself does not attempt to fully restate the design or results of the component works, which have all been previously detailed in published journal articles or government reports, and are included here as appendices. Note that this section defines abbreviations for each of these previous publications, which I use throughout this thesis.

The Extended Feature Set (EFS)..Chapter 3

The ***Extended Feature Set*** is the core of this portfolio, and the enabling technology that makes all of the other elements of this portfolio of work possible. EFS provides a standard definition for features and other attributes of latent or exemplar impressions, as defined in the ANSI/NIST-ITL standard:

[ANSI/NIST] *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.*[17]

Standardizing examiners' annotation procedures...Chapter 4

The ***EFS Markup Instructions*** project took materials and lessons learned from the implementation and use of EFS, resulting in a NIST Special Publication designed to assist in the adoption and use of EFS by latent examiners:

[EFSMI] *Chapman W, et al (2013) Markup Instructions for Extended Friction Ridge Features. National Institute of Standards and Technology, Special Publication 1151.*

Two software-development projects built upon these written instructions. The ***EFS Training Tool*** (Section 4.2) implements the EFS markup instructions in free, web-based software. ***ACEware*** (Section 4.3) is a task in progress that will use EFS as the basis for training and evaluating examiners in detailed documentation of ACE, and will facilitate documentation of operational casework.

Defining and measuring quality, clarity, and distortion..Chapter 5

The ***Latent Quality Study*** involved conducting a detailed survey of how quality and clarity are assessed within the latent fingerprint community, developing guidelines and metrics for describing the clarity of friction ridge impressions, and developing software tools to provide objective, reproducible methods for assessment of friction ridge impression clarity. The study resulted in the definition of ridge clarity/confidence used in EFS, and two publications:

[17] Applies to any ANSI/NIST-ITL version 2011 or later: 2013 and forthcoming 2015/16 revisions of the ANSI/NIST standard have not made significant changes relevant to EFS.

[LQSurvey] *Hicklin RA, et al (2011) Latent fingerprint quality: a survey of examiners. J. Forensic Identification, 61(4): 385-418.*

[AssessingLC] *Hicklin RA, Buscaglia J, Roberts MA (2013) Assessing the clarity of friction ridge impressions. Forensic Sci Int 226(1):106-117.*

The ***Latent Quality Metric (LQMetric)*** project (Section 5.7) used the results of the *Latent Quality Study* to develop operational latent quality software, incorporating clarity/quality metrics into the FBI's Universal Latent Workstation (ULW). No published description of LQMetric is yet available.

The ***Distortion Study*** (Section 5.9) developed metrics for quantifying and visualizing linear and nonlinear fingerprint deformations, and software tools to assist examiners in accounting for distortion in fingerprint comparisons.

[Distortion] *Kalka, ND, Hicklin RA (2014) On relative distortion in fingerprint comparison. Forensic Science International 244(2014), 78-84.*

Evaluating latent print examiners*... *Chapter 6

The ***Black Box Study*** (Section 6.1) was a large-scale study of the accuracy and reproducibility of latent print examiners' determinations. The follow-on ***Black Box Repeatability Study*** retested examiners to evaluate the repeatability of their determinations.

[BB] *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.*

[BBRR] *Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2012) Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners. PLoS ONE 7:3.*

The ***Sufficiency for Value Study*** (Section 6.2) evaluated how image clarity and feature content are associated with the assessment of latent value by latent print examiners.

[SuffValue] *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.*

The ***White Box Study*** (Section 6.3) was designed to investigate the relationship between examiners' annotations and their determinations. The overall study was published in three reports: ***Sufficiency for Individualization*** (Section 6.3.2), ***Analysis to Comparison*** (Section 6.3.3), and ***Interexaminer Variation in Minutia Markup*** (Section 6.3.4).

[SuffID] *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)*

[A-C] *Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (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)*

[IEVMM] *Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2016) Interexaminer variation of minutia markup on latent fingerprints. Forensic Science International, 264:89–99. (http://dx.doi.org/10.1016/j.forsciint.2016.03.014)*
Supporting information published separately:
Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2016) Data on interexaminer variation of minutia markup on latent fingerprints. Data in Brief, 8:158-190. (http://dx.doi.org/10.1016/j.dib.2016.04.068)

Interacting with AFIS*.. *Chapter 7

This chapter discusses different aspects of how latent print examiners interact with latent AFISs.

Under ***Latent AFIS interoperability*** (Section 7.1), I discuss the variety of efforts that are building on EFS to enable vendor-neutral interchange of data among proprietary latent AFISs. This includes two specifications that build on EFS to enable interoperability:

[EFSProfiles]	*Chapman, et al (2013) Extended Feature Set Profile Specification. NIST Special Publication 1134.* *http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1134.pdf*
[LITS]	*Chapman, et al (2013) Latent Interoperability Transmission Specification. NIST Special Publication 1152.* *http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1152.pdf*

The NIST ***Evaluation of Latent Fingerprint Technologies: Extended Feature Sets (ELFT-EFS)*** studies (Section 7.2) were conducted to evaluate the state of the art in latent AFIS matching, using different sets of EFS features marked by experienced latent print examiners.

[ELFT-EFS1]	*Indovina M, Hicklin RA, Kiebuzinski GI (2011) ELFT-EFS Evaluation of Latent Fingerprint Technologies: Extended Feature Sets. National Institute of Standards and Technology Interagency Report #7775.*
[ELFT-EFS2]	*Indovina M, Dvornychenko, V, Hicklin RA, Kiebuzinski GI (2012) ELFT-EFS Evaluation of Latent Fingerprint Technologies: Extended Feature Sets, Evaluation 2. National Institute of Standards and Technology Interagency Report #7859.*

Findings and Recommendations, Implications and Future Possibilities*..............................*Chapter 8

This chapter discusses the implications of the completed work, briefly discussing ongoing projects in which I am involved, and my personal recommendations for possible future research, technology, policies, best practices, and operational procedures.

1.3 Terminology

Although general agreement on terminology would be highly desirable, terminology is unfortunately not used consistently throughout the forensic science and biometrics community. Therefore it is necessary to make explicit how I am using terminology in this thesis: the Glossary defines how I am using a variety of terms, but here I wish to bring your attention to a few specific terms:

- ***Latent and print*** — "Latent" or "latent print" is the preferred term in North America for a friction ridge impression from an unknown subject, and "print" is used to refer generically to known or unknown friction ridge impressions from fingers, palms, toes, or soles. Outside of North America, a friction ridge impression from an unknown source is often referred to as a "mark" or "trace," and "print" is often used to refer only to known impressions. Here I am using the North American terminology to maintain consistency with my previous work, with SWGFAST usage, and with usage in the worldwide AFIS community. It should be noted that both *Merriam Webster's* and the *Oxford English Dictionary* equate "latent" (as a noun) with "latent fingerprint" and "latent print."[18,19] The term "exemplar" is used here to refer to a friction ridge impression that was collected under controlled conditions from a known subject.
- ***Individualization and identification*** — The term "individualization" has been controversial.[20] Here I use it solely as a synonym for "identification" in order to correspond with SWGFAST terminology, defined as "the decision by an examiner that there are sufficient discriminating

[18] *Merriam-Webster, Merriam-Webster.com. Accessed 24 Apr. 2014. <http://www.merriam-webster.com/dictionary/latent>.*

[19] *Oxford English Dictionary, Oxford University Press, OED Online version. Accessed June 2012.*

[20] *[Cole14, Stoney91]*

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."[21]

- ***Correct, appropriate, and consensus determinations*** — Some determinations are provably incorrect, such as an individualization of an image pair that are definitively known to be from different sources (false positive error). However, there is no standard way to determine whether a given determination is "correct" with respect to individualization vs. inconclusive, exclusion vs. inconclusive, or value vs. no value; therefore while I will describe errors as "incorrect", I do not ever refer to determinations as "correct." Given the lack of validated quantitative models, at the moment the best measure of determining whether a given determination is "appropriate" or "inappropriate" is consensus: if a supermajority or unanimous consensus of examiners concurs on a determination, it is reasonable to consider it "appropriate" — the appropriateness of a determination without a supermajority of examiners is at best debatable. Unfortunately, consensus requires multiple independent examinations, which are often only available in research studies.
- ***Rigor*** — I use rigor as a general term to refer to the demonstrable accuracy and reliability of a discipline, with procedures in place to ensure that accuracy and reliability, In particular I emphasize that rigor requires transparency, standardization, and quantifiability.[22]
- ***Accuracy, reliability, repeatability, and reproducibility*** — The terms accuracy and reliability are often used inconsistently in legal contexts[23] and even in some technical dictionaries.[24] I use what I consider to be the predominant definitions: "accuracy" is used here to refer to the correctness of a measured result (variation of a measurement from the true value);[25] "reliability" refers to the precision or consistency among different measurements (without regard to the measurements' relation to true values). I use "reliability" as a general term to refer jointly to its components: "repeatability," which refers to the extent of intra-examiner agreement (the same examiner at different times), and "reproducibility," which refers to the extent of inter-examiner agreement (different examiners).

[21] *[SWGFAST-ID12]*

[22] *Merriam-Webster defines rigorous as "scrupulously accurate: precise" and rigor as "strict precision: exactness" ("rigor" [def. 4] and "rigorous" [def. 3] Merriam Webster Online Dictionary. Merriam-Webster.com, accessed 18 Nov. 2015). Google defines rigorous as "extremely thorough, exhaustive, or accurate" and rigor as "the quality of being extremely thorough, exhaustive, or accurate" ("rigor" [def. 2a] and "rigorous" [def. 1] Google.com, accessed 18 Nov. 2015). American Heritage defines rigorous as "characterized by or adhering to strict standards or methods; exacting and thorough" and rigor as "strictness in adhering to standards or a method; exactitude." ("rigor" [def. 2a] and "rigorous" [def. 1] American Heritage Dictionary of the English Language, Fifth Edition. AHDictionary.com, accessed 18 Nov. 2015).*

[23] *In legal contexts, "reliable" is often used to mean "accurate." [Cole06]*

[24] *For example, the NIST Engineering Statistics Handbook defines accuracy as "In metrology, the total measurement variation, including not only precision (reproducibility), but also the systematic offset between the average of measured values and the true value." (NIST/SEMATECH e-Handbook of Statistical Methods, www.itl.nist.gov/div898/handbook, accessed 14 Jan 2017) Although ISO defines "accuracy as "closeness of agreement between a test result or measurement and the true value" it then adds a note "Accuracy refers to a combination of trueness and precision." ("Accuracy (trueness and precision) of measurement methods and results", ISO TC 69/SC 6 N)*

[25] *Accuracy or measurement accuracy: "closeness of agreement between a measured quantity value and a true quantity value of a measurand" (International vocabulary of metrology — Basic and general concepts and associated terms, JCGM 200:2008)*

- ***Clarity and quality*** — I differentiate between clarity and quality: here, "clarity" refers to fidelity, the extent to which the physical features are faithfully represented in an image being used; "quality" is a more general concept that includes not just clarity, but also the utility of the inherent physical features, and the utility for a specified purpose. See detailed discussion in Section 5.1.
- ***Ground truth*** — It is difficult or impossible to have "ground-truth" (definitive) certainty in many aspects of latent print examination, such as definitive knowledge of mating (c.f.) or friction ridge features.
- ***Image, impression, print*** — All discussion of latents and exemplars in this thesis applies to digital images of friction ridge impressions, and therefore I do not differentiate among the terms "image," "impression," and "print."
- ***Mate, nonmate, and mating*** — "Mating" refers to definitive knowledge of whether two prints share the same source. A "mate" is an exemplar that is known to be from the same source as the latent (i.e. from the same area of skin from a single subject). A "nonmate" is an exemplar that is definitively known to be from a different source as the latent. In general, mating implies ground-truth (c.f.) knowledge of the source of each print (especially when latent print examiners are being evaluated), but in some uses (such as AFIS evaluations) consensus among examiners can serve as a surrogate.
- ***Determination, decision, and conclusion*** — I use "determination" to refer to the Analysis phase value determinations ("Value for individualization" (VID), "Value for exclusion only" (VEO), or "No Value" (NV)), and the Comparison/Evaluation phase determinations (Individualization, Exclusion, or Inconclusive). I use "conclusion" to refer solely to Individualization or Exclusion (not Inconclusive). In the original Black Box study we used the term "decision" but subsequent papers used "determination." I use "determination" rather than "decision" in part because the latter term brings with it implications[26] that would be tangential to my focus here.

[26] *see e.g. [Beidermann08]*

Chapter 2 Criticisms and limitations of the current examination process

Forensic science in general — and especially latent print examination — has been undergoing a period of internal and external scrutiny since the late 1990s, particularly in response to admissibility challenges that began in 1999, the 2004 Mayfield misidentification, and the 2009 National Research Council's report on forensic science (hereafter "NRC Report").

Assertions that were common in the field twenty years ago are not acceptable today, because subsequent information (from research or operational errors) have shown that they could not be supported. At that time, many examiners claimed individualization to the exclusion of all others,[27] or a zero error rate.[28] Some examiners claimed that two examiners, trained to competency, would always reach the same conclusion.[29] Standardized or detailed documentation of conclusions was not seen as important, based on the assumption that as long as examiners reached the same conclusion, it didn't matter how they got there. Claims about the discipline often were presented by practitioners without any justification, and what was made available did not necessarily stand up to scrutiny. For example, the unpublished "50k Study" was presented as a scientific study on fingerprint individuality, in response to the initial 1999 Daubert challenge in the Mitchell case;[30] I concur with a variety of critics[31] that the experimental design did not support its purpose and conclusions, and its statistical analyses were particularly inadequate — instead of supporting the discipline, the 50k study cast into question the case for the discipline. I contend that such problematic statements and research are what made the criticisms and the NRC Report necessary, but that these challenges have made the discipline better: the criticisms made it possible to review and change long-standing assumptions and practices, resulting in improved practices and a climate open to change.

As an illustration of how latent print examination has had to change in response to technology in the past, a colleague shared an anecdote of an examiner's first encounter with AFIS: one examiner made individualizations against both the first and second candidates in a list, even though the candidates were from different subjects; the examiner had never had to make comparisons against unusually similar subjects before. That examiner soon retired.[32]

2.1 Effects of admissibility challenges and the Mayfield error

" Daubert and Kumho Tire invite fresh and critical looks at old habits and beliefs."[33]

The admissibility of fingerprint evidence in the US was not challenged for decades, until after Daubert and Kumho Tire replaced the older Frye standard. Starting with US v. Mitchell[34] in 1999,

[27] *[Cole14]*

[28] *"And you can always bring it home to the fact that a competent examiner correctly following the ACE-V methodology won't make errors. [...] the error rate of the methodology is zero" [Scarborough]*

[29] *[Cole02]*

[30] *[Mitchell99]*

[31] *[Wayman00, Stoney01, Champod01, Pankanti02, Kaye03, Zabell05]*

[32] *Personal communication with Ed German*

[33] *[Havvard01]*

[34] *[Mitchell99]*

there have been dozens of challenges to the admissibility of fingerprint evidence. Of all these challenges, only one (Maryland v. Bryan Rose)[35] denied the admissibility of the testimony of fingerprint examiners (subsequently admitted in federal court), and another (Llera Plaza)[36] resulted in an initial denial that was partially reversed (by the same judge) six weeks later. Even though all of these challenges have not had a direct impact in the cases at hand, they laid the groundwork for a critical review of the latent print examination process that did not come to a head until the Mayfield error. Starting around the same time as (sometimes in association with) the Daubert challenges, a variety of critical reviews of latent print examination were published, such as by Simon Cole, Itiel Dror, Ralph and Lyn Haber, and Jennifer Mnookin.[37] However, prior to the Mayfield error, the criticisms had little effect on actual practice.

In March 2004, the FBI Laboratory misidentified a latent fingerprint from the Madrid train bombings to an Oregon lawyer named Brandon Mayfield instead of the person was subsequently determined to be the true source of the latent, an Algerian named Ouhane Daoud. The primary causes of the error were reported as the unusual similarity of the prints, bias from the known prints of Mayfield, faulty reliance on extremely tiny (Level 3) details, inadequate explanations for difference in appearance, failure to assess the poor quality of similarities, and failure to reexamine the latent following the Spanish National Police's negative conclusion.[38] The US Department of Justice (DOJ) Office of Inspector General (OIG) released a detailed 2006 report[39] [hereafter "Mayfield OIG Report"] with a detailed analysis of how the examination was conducted, the causes of the misidentification, and recommended remediation. A subsequent 2011 report[40] reviewed progress in the intervening five years.

The Mayfield misidentification had a massive effect on the forensic science community in general, and especially the area of latent print examination. The impact was so great in part because it was an error in a high-profile case by esteemed experts at a major laboratory, in which no one could blame the error on inadequacies of the examiners in question — the error had to be attributed to the process. Some of the impact can be attributed to the transparency of the resulting Mayfield OIG report, which provided extensive detail in describing what happened, and allowed a new level of scrutiny into the latent print examination process.

Much of the work discussed here is a direct result of the Mayfield misidentification. In response to the Mayfield case, senior management of the FBI Laboratory tasked a review committee to evaluate and recommend research to test the scientific basis of friction ridge examination; that committee's findings and recommendations were published in 2006.[41] That committee recommended a program of research needed to assess the scientific basis of identification using latent print evidence, including four high-priority projects: quality, quantity, performance, and exclusion. Much of the portfolio of work incorporated in this thesis were part of that research program, including the Latent Quality study (quality), the Black Box study (performance), and the White Box study (quantity and exclusion). Note that the 2006 review committee provided only a general overview of the need for each research project: conceptualizing and designing the studies was the responsibility of our team.

[35] *[Rose07]*

[36] *[LleraPlaza02]*

[37] *e.g. [Cole05,Cole06,Dror06a,Dror06b,Haber08,Haber09,Mnookin08b]*

[38] *[Mayfield06]*

[39] *[Mayfield06]*

[40] *[Mayfield11]*

[41] *[Budowle06]*

2.2 The National Research Council's Report on Forensic Science

In 2009, the National Research Council of the National Academies came out with a scathing criticism of the forensic sciences. The NRC report made a series of recommendations, a number of which are directly tied to improving the rigor of latent print examination. The recommendations relevant to latent print examination are summarized here:

- (NRC #1a) Establishing and enforcing best practices for forensic science professionals and laboratories;
- (NRC #1b) Establishing standards for the mandatory accreditation of forensic science laboratories and the mandatory certification of forensic scientists
- (NRC #2) Establish standard terminology to be used in reporting and testifying about the results of forensic science investigations.
- (NRC #3) Research is needed to address issues of accuracy, reliability, and validity in the forensic science disciplines.
  - (NRC #3a) Studies establishing the scientific bases demonstrating the validity of forensic methods.
  - (NRC #3b) Development and establishment of quantifiable measures of the reliability and accuracy of forensic analyses. [...] The research by which measures of reliability and accuracy are determined should be peer reviewed and published in respected scientific journals.
  - (NRC #3c) Development of quantifiable measures of uncertainty in the conclusions of forensic analyses
  - (NRC #3d) Automated techniques capable of enhancing forensic technologies
- (NRC #5) Encourage research programs on human observer bias and sources of human error in forensic examinations. [...] Research on sources of human error should be closely linked with research conducted to quantify and characterize the amount of error.
- (NRC #6) Develop tools for advancing measurement, validation, reliability, information sharing, and proficiency testing in forensic science and to establish protocols for forensic examinations, methods, and practices. Standards should reflect best practices and serve as accreditation tools for laboratories and as guides for the education, training, and certification of professionals.
- (NRC #8) Forensic laboratories should establish routine quality assurance and quality control procedures to ensure the accuracy of forensic analyses and the work of forensic practitioners. Quality control procedures should be designed to identify mistakes, fraud, and bias; confirm the continued validity and reliability of standard operating procedures and protocols; ensure that best practices are being followed; and correct procedures and protocols that are found to need improvement.
- (NRC #12) Launch a new broad-based effort to achieve nationwide fingerprint data interoperability. Convene a task force comprising relevant experts from the National Institute of Standards and Technology and the major law enforcement agencies (including representatives from the local, state, federal, and, perhaps, international levels) and industry, as appropriate, to develop:
  - (NRC #12a) Standards for representing and communicating image and minutiae data among Automated Fingerprint Identification Systems. Common data standards would facilitate the sharing of fingerprint data among law enforcement agencies at the local, state, federal, and even international levels, which could result in more solved crimes, fewer wrongful identifications, and greater efficiency with respect to fingerprint searches; and
  - (NRC #12b) Baseline standards—to be used with computer algorithms—to map, record, and recognize features in fingerprint images, and a research agenda for the continued

improvement, refinement, and characterization of the accuracy of these algorithms (including quantification of error rates).

Much of my reaction to the NRC report remains unchanged from my immediate reaction, which was that its recommendations were very much in line with our work already in progress. In March 2009, I wrote an internal memorandum with my responses, excerpted here:

> *The communities of interest in forensic science include forensic science practitioners, academia, industry, and the legal community. Much of the forensic science research that has been conducted to date has been too balkanized in that it has been designed and conducted by one or two of these communities without considering whether it actually addresses the issues raised by all of the communities. Many of the fingerprint-related recommendations in the NRC Report have a common basis, including AFIS interoperability (#12), standard terminology and reporting (#2), more sophisticated research (#3), more sophisticated analysis of human error and bias (#5), improved tools, protocols, and standards (#6), metrics, tools, guidelines, and procedures for quality assurance and quality control (#8), and the flexibility and interoperability necessary for effective forensic science homeland security work (#13). All of these have been limited in the past by oversimplified or nonexistent means of defining the information content of friction ridge impressions, and by the lack of precise, repeatable, quantifiable, and robust means of defining how fingerprint analysis and comparison are conducted. To address these limitations, we have been specifically targeting enabling technologies for several years, and which are just reaching fruition. A core group of projects that focus on enabling technologies includes the following:*
>
> - *CDEFFS (the ANSI/NIST Committee to Define an Extended Fingerprint Feature Set), an international multi-organization group that has defined an interoperable standard (EFS = Extended friction ridge feature specification) for both AFIS and non-AFIS definition of friction ridge features. The CDEFFS committee was formed almost precisely as defined in NRC #12: created in Dec. 2005 based on recommendations by SWGFAST, sponsored by FBI and NIST, and including many of the world's fingerprint experts, from criminal justice and forensic agencies (federal, state, and local), industry (including all major AFIS vendors), and academia (US and international). The result, however, is not just limited to AFIS interoperability as in NRC #12, but extends into a standard means of defining and exchanging non-AFIS analysis and comparison work. [Status: EFS draft specification 0.3 is in a comment period and will be finalized by 19 March 2009; EFS is the basis for NIST ELFT-EFS, which is starting EFS AFIS interoperability evaluations after a workshop 19-20 March 2009; EFS features are defined with respect to confidence measures developed during the Quality study; EFS is being used in the development of reference data sets, which in turn are being used to define guidelines for the standard markup of friction ridge data content (NRC #2); The Universal Latent Workstation incorporates a reference implementation of EFS (ULW 5.6 to be released in broad-based beta test 19 March 2009); Black box and Quantity studies use EFS as to define the features and characteristics used in analysis.]*
> - *The FBI Laboratory Quality and Quantity Studies (in progress since July 2007) are being conducted to 1) develop policy and procedure guidelines for describing and evaluating the quality and data content of friction ridge impressions and friction ridge comparisons, and 2) to develop rigorous, repeatable and quantifiable software metrics. The Quantity study will further proceed to assess the*

independence and discriminative value of friction ridge features and to devise a data content model for individualization/identification decisions. These guidelines and metrics are critical enabling technologies that will assist in more sophisticated research (#3), more sophisticated analysis of human error and bias (#5), a common basis for improved tools, protocols, and standards (#6), and provides a basis for the metrics, tools, and guidelines necessary to guide procedures for quality assurance and quality control (#8). [Status: Quality guidelines have been incorporated into EFS; Quality software (LQAS=Latent Quality Assessment Software) is being used for research and is being assessed for use in casework; Quantity guidelines and models are in development; Quality and quantity models will be used in the analysis of the Black Box study.]

- *The FBI Laboratory Black Box study (in progress) is an assessment of latent print expert's ability to reach reliable conclusions during analysis and comparison. The guidelines, models, and software from the quality and quantity studies will be used in the data selection and analysis for the Black Box study.*
- *The Universal Latent Workstation (ULW) was developed starting in 1998 specifically as a tool for AFIS interoperability. While most operational use in the thousands of identifications it has made since 1999 have been against IAFIS, it is operationally used as a tool to enable inter-AFIS interoperability in multiple state and local jurisdictions, and additionally serves worldwide as a tool for friction ridge research and exchange. Lessons learned during the development of ULW were used in the development of EFS, and ULW now incorporates a reference implementation of EFS.*
- *The NIST ELFT-EFS (Evaluation of Latent Fingerprint Technologies, Extended feature Sets) is an evaluation of latent AFIS interoperability, based on EFS, and specifically addresses measurement of AFIS interoperability (NRC #12).*

2.3 The PCAST Report on Forensic Science

In 2016 the President's Council of Advisors on Science and Technology (PCAST) issued a report, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*, [PCAST16] with an addendum in 2017 [PCAST17]. The PCAST report addressed how to determine the validity of forensic disciplines the admissibility in US courts of a set of forensic disciplines. The report defined how to establish the validity of forensic disciplines:

> *PCAST noted that the only way to establish the scientific validity and degree of reliability of a subjective forensic feature-comparison method—that is, one involving significant human judgment—is to test it empirically by seeing how often examiners actually get the right answer. Such an empirical test of a subjective forensic feature-comparison method is referred to as a "black-box test." The point reflects a central tenet underlying all science: an empirical claim cannot be considered scientifically valid until it has been empirically tested.*
>
> *If practitioners of a subjective forensic feature-comparison method claim that, through a procedure involving substantial human judgment, they can determine with reasonable accuracy whether a particular type of evidence came from a particular source (e.g., a specific type of pistol or a specific pistol), the claim cannot be considered scientifically valid and reliable until one has tested it by (i) providing an adequate number of examiners with an adequate number of test problems that resemble those found in forensic practice and (ii) determining whether they get the right answer with acceptable frequency for the intended application. While scientists may debate the precise*

> *design of a study, there is no room for debate about the absolute requirement for empirical testing. [PCAST17]*

The PCAST report assessed seven forensic disciplines, and found that its requirements for validity were only met by latent print examination, and DNA analysis of single-source samples and simple mixtures. For the disciplines that PCAST found did not have adequate empirical evidence of validity, they underscored the need to conduct black box studies, and referred extensively to our studies as examples of how such studies should be conducted:

> *PCAST applauds the work of the friction-ridge discipline, which has set an excellent example by undertaking both (i) path-breaking black-box studies to establish the validity and degree of reliability of latent-fingerprint analysis, and (ii) insightful "white-box" studies that shed light on how latent-print analysts carry out their examinations, including forthrightly identifying problems and needs for improvement. PCAST also applauds ongoing efforts to transform latent-print analysis from a subjective method to a fully objective method. In the long run, the development of objective methods is likely to increase the power, efficiency and accuracy of methods—and thus better serve the public. [PCAST17]*

2.4 Limitations, gaps, and needs

Here I summarize my view of the core limitations, gaps, and needs of latent print examination — some of which have been addressed in part in the course of the work described in this thesis. Note that any assessment of the latent print discipline is a snapshot in time, because the discipline has been changing in response to Mayfield and the NRC report. The following sections discuss a need to assess the accuracy and reliability of latent print examiners' determinations (Section 2.4.1), a need to improve the existing ACE-V examination process (Section 2.4.2), and a need for improved procedures (Section 2.4.3).

2.4.1 Assessing the accuracy and reliability of latent print examiners' determinations

As long as latent print examiners make determinations, the minimum threshold for assessing whether examination is rigorous would be assessing how accurate and reliable those determinations are. I see both a need for an overall assessment of the accuracy and reliability of latent print examiners' determinations over a broad range of comparisons (Section 2.4.1a); and a need for assessing the accuracy and reliability of a specific latent print examiner's determination on a specific comparison, based on proficiency testing and difficulty metrics (Section 2.4.1b).

2.4.1a Overall accuracy and reliability

In considering the range of forensic disciplines, I find it surprising how many do not have studies from which an outsider can get a general understanding of the accuracy and reliability of examiners' determinations. If an examiner from a hypothetical forensic discipline wants to make a conclusion, the minimum threshold for validation would require such an assessment. The consumers of an examiner's decisions (laboratory managers, police, prosecution, defense, judge, and jury) need to have at least a basic understanding of the accuracy, (intra-examiner) repeatability, and (inter-examiner) reproducibility of those decisions. In Section 6.1, I discuss the Black Box studies, which were conducted to provide such rates.

2.4.1b Accuracy and reliability for a specific examiner and comparison

Overall rates provide a general understanding of examiners' abilities. However, given the differing skills of examiners and the range of difficulty of latent print comparisons, it is more desirable to understand a specific examiner's abilities to render a decision for a specific comparison. One approach would be to accompany an examiner's decision with corroborating data showing the accuracy, reproducibility and repeatability of decisions for a given examiner proficiency level and comparison difficulty.

There are currently no standardized proficiency tests of examiners. In Section 6.1, I discuss how the Black Box results showed that the skill of latent print examiners is multidimensional, suggesting approaches that could be used in constructing proficiency tests.

The difficulty of a given comparison can be assessed in various ways: as subjective examiner assessments (as used in the Black Box study, Section 6.1), using quality metrics (Chapter 5), or based on the examiners' own markup (as used in the White Box study, Section 6.3).

2.4.2 Improving the existing ACE-V examination process

The core of the fingerprint examination process is called ACE-V[42]: Analysis of the latent print (interpretation based on how it was deposited, developed, etc.), side-by-side Comparison of the two prints (observation of (dis)similarities), Evaluation (determining whether the (dis)similarities are sufficient to support a conclusion), and Verification (examination by a different examiner). My view of ACE-V is that although SWGFAST has significant strides in developing guidelines for ACE-V, it remains underspecified and not standardized in practice. I concur with Mnookin when she says "At root, ACE-V in its current incarnation amounts to no more and no less than a set of procedures to describe the careful comparison of a latent print with a potential source print by an initial examiner and a subsequent verifier."[43]

Here I would like to focus on several key limitations of the current ACE-V process:

- Lack of standard terms and meanings for determinations
- Lack of standardized feature-level documentation
- Limitations of the current holistic determinations
- Implementation in training, operating procedures, accreditation, and certification

2.4.2a Lack of standard terms and meanings for determinations

Latent print examiners make a variety of minor decisions in the ACE process, but two result in key determinations: the analysis determination during the analysis phase that a latent is of value and therefore suitable for comparison, and the comparison/evaluation determination regarding whether two impressions are from the same or different sources.

The Analysis determinations of suitability (value) vary among agencies. In the Black Box study, 55% of participants reported that their standard operating procedures assessed value as a 2-category decision of value for individualization (VID) vs. no value (NV), described by SWGFAST as "Approach 1"; 14% used a 2- category decision of value for comparison (VCMP) vs. NV (SWGFAST "Approach 2"); 30% used a three-category decision of VID, value for exclusion only (VEO), or NV.[44]

Although examiners and agencies appear to be consistent at least in concept with individualization/identification determinations, agencies are notably different regarding exclusion and inconclusive determinations:

- Participants in the Black Box study differed widely in how they use the term "exclusion" as a conclusion in their standard operating procedures: examiners differed on whether exclusion means that the latent did not come from any friction ridge skin for that subject (51%), from any finger from the subject (10%), or from a specific exemplar (e.g., a specific finger) (11%) — 4% said that any comparison that is not an individualization is an exclusion, and 23% said they do

[42] *[Huber59, Ashbaugh99]*

[43] *[Mnookin10]*

[44] *[BB], Appendix 1.4; [SWGFAST-StdExam13]*

not use the term. However, regarding the concept of exclusion, most survey respondents (84%) said that they often conclude that a latent and the exemplars provided definitively did not come from the same source; only 3% never make such a conclusion.

- Black Box participants also differed widely regarding inconclusive determinations: approximately half of the Black Box survey respondents reported that they are either not permitted to make (32%) or discouraged from making (19%) an inconclusive determination if the latent and exemplar are both of value and include a large potentially corresponding area. Discouraging or disallowing inconclusive determinations puts pressure on examiners to make individualization or exclusion determinations that they might not have otherwise made, or results in the examiners retroactively labeling the print(s) as no value.

In short, agencies do not agree on the terms, meanings, or use of the determinations made in the examination process. When combined with the lack of transparency into how these decisions are made, we have a fundamental problem.

2.4.2b Lack of standardized feature-level documentation

Although ACE-V requires detailed evaluation of often complex data, this information is not generally documented in a rigorous, quantifiable, and reproducible manner (at least not in the US). Frequently Analysis and Comparison assessments are not documented in sufficient detail for another qualified latent print examiner to understand what information the examiner used in making determinations.

The Mayfield OIG report and the NRC report were both critical of the lack of detailed documentation of the latent print examination process. The Mayfield OIG report described how the lack of documentation contributes to circular reasoning,[45] and indicated that required documentation may result in more reliable conclusions.[46] The Mayfield OIG report recommends more rigorous documentation (Recommendation 10) and separate documentation of features observed in the latent during the Analysis phase (Recommendation 11). The NRC report also called for required documentation, for transparency and to "provide the courts with additional information on which to assess the reliability of the method for a specific case. "[47]

The absence of detailed documentation indicating which features were used in making a given decision is problematic for several reasons:

- ***Transparency*** — As we have discussed already (Section 1.1), without detailed documentation the actions and inner workings of the decision process are neither visible nor accessible, and

[45] *"This process of circular reasoning infected the process, particularly in the absence of standards or safeguards requiring the examiner to keep distinct which features were seen in the latent fingerprint during the analysis and which were only suggested during the comparison. This error likely would have been avoided had the examiner firmly established and documented which features were clearly discernible in the latent fingerprint in the 'analysis' phase, before conducting a comprehensive side-by-side comparison." [Mayfield06], p191-192*

[46] *"The absence of substantive documentation requirements is a conspicuous shortcoming of the current SOPs. We believe that there is a strong possibility that if the examiner and verifier had been required to document the analysis and comparison phases of their examinations, they might have noticed more dissimilarities and appreciated the cumulative impact of them before reaching their flawed conclusions. They might also have had greater appreciation for the low quality of the admitted similarities between the latent and the Mayfield known prints. We believe that documentation would have facilitated a more objective comparison and evaluation, regardless of the particular standard utilized to declare an identification." [Mayfield06], p202*

[47] *[NRC09] p 5-13*

are not fully articulated for any consumer of an examiner's determinations (e.g. supervisors, other examiners, attorneys, judges).[48]

- ***Potential bias*** — 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. When comparing highly similar prints, the examiner runs the risk of confirmation bias or circular reasoning, in which additional features may be suggested during comparison based on the similarity. In the Mayfield misidentification, the initial examiner reinterpreted five of the original seven Analysis points to be more consistent with the (incorrect) exemplar.[49] Without separately documenting the features observed in the analysis and comparison phases, the examiner may not realize — and the consumers of the decision cannot know — whether such bias affected the decision.
- ***Quality assurance*** — The lack of detailed documentation hampers quality assurance and technical review, which have no method of flagging borderline decisions or potentially biased decisions.

There is a general lack of formal guidelines for documentation, and forensic laboratories vary greatly in how operational latent print examination is documented. 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."[50] 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. Other than for AFIS searches, most agencies do not require detailed markup to document the features of a latent in Analysis, nor corresponding features used in Comparison. Those agencies that do require markup vary substantially on how that markup is effected. Some agencies (even now) document minutiae using pinpricks in physical photographs. Some agencies annotate using general-purpose tools such as Adobe Photoshop: while this is reasonable at a small scale, such a generic format retains none of the semantics associated with the features; while another examiner, agency, or automated analysis tool could open and view the resulting file, the exact intentions of the original examiner would frequently be lost. More rigorous approaches have been implemented in the University of Lausanne's PiAnoS (Picture Annotation System) and Latentworks by Mideo Systems. A limitation of both PiAnoS and Mideo Latentworks is that they are not standards-based systems and therefore detailed annotations in either cannot readily be exchanged with other systems and have no interoperability with AFIS workstations.

There are no generally accepted, rigorous definitions of features or clarity, and therefore no generally accepted systematic approaches to indicate confidence in features, to define ridge detail

[48] *"Conveying levels of certainty, weights of features, and tolerances of the analyst are helpful contributions to the transparency of documenting the ACE-V decision-making process. In cases where variation in analyst opinions can commonly occur, it becomes critical to understand how analysts reach conclusions. GYRO or any similar annotation system (such as PiAnoS) provides a mechanism for enhanced transparency, and additional information is easily conveyed to a reviewing analyst." [Langenburg11], p382*

[49] *"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." [Mayfield06], p7*

[50] *[SWGFAST-Doc12]*

(level-3) features, or even consistently-used definitions of what exactly constitutes a minutia. The lack of such rigorous definitions and systematic approaches contributes to a lack of reproducibility (interexaminer agreement) and repeatability (intraexaminer agreement) of which features are annotated by examiners,[51] complicates attempts to develop quantitative approaches for sufficiency — as well as limiting interchanges among examiners, long-term archiving of casework, validation, evaluation of examiner capabilities, quantitative analysis, and quality assurance.

Much of my work has been a multifaceted approach toward standardization of detailed markup, including precise and detailed data formats (Chapter 3); procedural standards, guidelines, and training materials instructing examiners when and how to use the formats (Chapter 4); software compliant with those formats (Section 4.3 as well as the Universal Latent Workstation); evaluation of the efficacy of those formats (Chapter 7); and enforcement of standards (via my role in the Organization of Scientific Area Committees).

2.4.2c Limitations of the current holistic determinations[52]

Conceptually, determinations could be made using numeric, holistic, or probabilistic approaches.[53] In practice, numeric and holistic approaches are broadly used; as yet, probabilistic approaches are very rarely used operationally.

Some countries use a numeric approach in which a minimum minutia count ("point standard") is used as a criterion for determining that a latent is of value, or for individualization: a 2011 survey of 73 countries by INTERPOL found that 44 countries use a point standard, 24 of which require a minimum of 12 minutiae.[54] Various papers have indicated that a minimum minutia threshold is problematic.[55] The U.K. and most agencies in the U.S. previously used minutia count standards but abandoned them in favor of a nonnumeric, holistic approach.[56] In 1973, the International Association for Identification resolved that there was no basis for requiring a "pre-determined minimum number of friction ridge characteristics" for individualization.[57]

In the holistic approach, an examiner's individualization determination is based on that examiner's assessment of the quantity and clarity of corresponding features, their relationships, and their specificity.[58] ACE relies upon the examiner's skills, training and experience, not upon formal criteria. In the absence of such criteria, and given the lack of standardized markup, the only available method for assessing whether an individualization is more appropriate than an inconclusive determination for a particular comparison is by consensus among examiners.[59]

Ideally, the examiners' decisions would be augmented or replaced with an estimate of the probability that two prints came from the same source. There have been a number of attempts over more than a century to more precisely articulate and standardize the procedures by which examiner reach determinations.[60] Some of this research has been successfully incorporated into the development of Automated Fingerprint Identification Systems (AFISs), which are effective tools in

[51] *[Langenburg09b, Dror11, Langenburg12a, SuffValue]*

[52] *Derived from [SuffID]*

[53] *[ENFSI15]*

[54] *[Farelo12]*

[55] *[IsraelNP95, Polski11, Su10]*

[56] *[Evett96, McKie11]*

[57] *[IAI73]*

[58] *[SWGFAST-Conclusions13, Locard20]*

[59] *[SWGFAST-ErrorRates12]*

[60] *Surveys in [Stoney01, Neumann13a]*

matching finger- and palmprints in very large databases. For latents, AFISs generate candidates for human examiners to compare, and do not make automated decisions[61] — for exemplars, which are generally larger, higher quality, and less distorted than latents, AFISs can make fully automatic determinations without involving human examiners for all but the poorest quality images.[62] A variety of research proposes the use of statistical models[63] to augment or replace the determinations of latent print examiners with probabilistic estimates of the strength of evidence; these models are not yet generally accepted for operational use. As I will discuss in Chapter 8, I see augmenting the examiner with probabilistic approaches as a desirable medium-term solution, and replacing the examiner (at least for easier comparisons) as a desirable long-term solution. For the near term, we need to consider how to improve the existing holistic determinations, beyond standardizing terminology/usage (Section 2.4.2a) and detailed documentation (Section 2.4.2b).

2.4.2d Implementation in training, operating procedures, accreditation, and certification

Note that I refer to lack of "standardization in practice." I would like to make a distinction between a document that calls itself a standard, and actual standardization. Many formal "standards" are rarely used, including many formally developed through Standards Development Organizations (SDOs).

My view is that a true standard must be

- Incorporated into training programs, and reviewed in the accreditation of training programs
- Incorporated into agency operating procedures, and reviewed in the accreditation of agencies
- Evaluated in the certification of individual examiners
- Broadly and consistently implemented in operational use

The primary limitation to the SWGFAST standards has been lack of enforcement. My hope for NIST OSAC will be able to surmount this obstacle to true standardization by enforcing the incorporation of standards to training programs, accreditation of agencies, and certification of individual examiners.

2.4.3 Improved procedures outside of ACE-V

It is important to note that ACE-V does not describe the entire latent print process. Prior to ACE-V latents are detected, processed, and captured, and candidate exemplars are selected (as suspects or via AFIS search). Procedures after ACE-V include quality assurance, technical review, and reporting. These procedures are often agency specific: again, SWGFAST has developed guidelines for some aspects of these procedures, but are underspecified and not standardized in practice. Two of the areas outside of ACE-V warrant especial attention:

- There is a need for improved quality assurance procedures to minimize the risk of error, to ensure more reproducible results, and to minimize the risk of bias. In particular I see a need for incorporating automated tools into quality assurance procedures.
- I see a need to balance the risk of error with the need to effectively conduct casework. Although I generally agree with many of the points in the NRC report, I do believe that one problematic effect has been a singular focus on erroneous identifications. A forensic scientist could avoid erroneous identifications completely by doing no work whatsoever. In engineering or scientific analyses, a report that shows a single error rate generally is based on oversimplified assumptions, when in reality there are almost always tradeoffs between errors. In practice, there are not one but two types of error that need to be considered: the risk of identifying the

[61] *[Komarinski04, ELFT-EFS1, ELFT-EFS2]*

[62] *[Wilson04]*

[63] *e.g. [Abraham13, Neumann13a, Neumann12, Pankanti02, Neumann07, Su09]*

wrong person versus the risk of not identifying the perpetrator. In enhancing the latent print examination process, we need to consider not just methods of limiting erroneous identifications, but also methods of making the process more efficient, and increasing the probability and number of correct identifications.

Chapter 3 The Extended Feature Set (EFS)

EFS is the core of this portfolio, and the enabling technology that makes all of the other elements of this portfolio of work possible. Here I describe how EFS was developed, its purpose, my guiding philosophy for EFS, and summarize the content and use of EFS.

3.1 The ANSI/NIST-ITL standard

The Data Format for the Interchange of Fingerprint, Facial, & Other Biometric Information (ANSI/NIST-ITL) is the most recent revision of a series of standards that began in 1983; first version 1986. These ANSI/NIST standards have been extensively used as the primary method of communicating biometric information for law enforcement and other large-scale identification purposes. The ANSI/NIST ITL standards are the basis for biometric and forensic specifications used around the world, including the FBI EBTS, DOD EBTS, DHS IXM, Interpol's INT-I, and a wide variety of national, state, and local application profiles.[64]

3.2 Committee to Define an Extended Fingerprint Feature Set (CDEFFS)

At the ANSI/NIST-ITL 1-2000 Standard Workshop I[65] in April 2005, various participants noted that the fingerprint feature definitions in the ANSI/NIST-ITL standard are oversimplifications of the more extensive set of features used by human fingerprint experts. Use of these oversimplified feature definitions limits the performance of AFIS, and limits the value of ANSI/NIST-ITL files as a format for communication between human fingerprint examiners. SWGFAST was tasked to identify, define and provide guidance on additional fingerprint features beyond the traditional ending ridges and bifurcations defined in the ANSI/NIST-ITL-2000 standard. SWGFAST drafted a memo[66] to NIST in response, enumerating the features used by expert human latent examiners that were not currently addressed in fingerprint standards. SWGFAST stated its concern: "AFIS technology, since its onset, has utilized a very limited amount of fingerprint detail. Latent print experts must rely on far more information in effecting individualizations/exclusions than just ending ridges and bifurcations, i.e., the Type-9 minutiae record. SWGFAST is attempting to educate and provide to the vendor community the additional features and how they are utilized by these experts."

In response to the SWGFAST memo, at the ANSI/NIST-ITL 1-2000 Standard Workshop II[67] in December 2005, Steve Meagher (FBI) and I gave a presentation entitled "Extended Fingerprint Feature Set",[68] and proposed that a committee be convened to define an Extended Feature Set (EFS) as an addendum to the next revision of the ANSI/NIST-ITL standard. The Committee to Define an Extended Fingerprint Feature Set (CDEFFS) was chartered for that purpose, and I was voted the chair; I served in that role until the formal adoption of EFS in 2011. The committee included 47 representatives from various US Federal Agencies, SWGFAST, the US and international latent fingerprint community, and engineers from a variety of AFIS vendors. The final version of EFS was the result of agreements reached among the members of CDEFFS during multiple workshops held in 2006, extensive electronic interactions and document reviews from 2005 through 2009, and a

[64] *[NIST-AppProfiles]*

[65] *[NIST-Workshops]*

[66] *[SWGFAST-Memo05]*

[67] *[NIST-Workshops]*

[68] *[Hicklin05]*

smaller number of electronic interactions from 2009 through 2011. CDEFFS was dissolved when EFS was formally incorporated into ANSI/NIST-ITL in November 2011.

3.3 EFS purpose and uses

The purpose of EFS is to define a quantifiable, standard method of characterizing the content of a fingerprint or other friction ridge image. While the initial impetus for EFS was to have an interchange format by which the features discerned by examiners could be communicated to AFIS, it was immediately apparent that defining the content of an impression in a standard, machine-readable format would serve a variety of uses.

EFS provides a standard definition for features and other attributes of latent or exemplar impressions, correspondences and discrepancies among two or more impressions, and determinations. These definitions can be used to document content either as discerned by an examiner, or as extracted by an AFIS or other automatic process. EFS was designed from the start to include all friction ridge data, including palms, lower finger joints, and plantars (feet and toes).

Uses of EFS include:

- AFIS searches and responses (feedback to examiners from automated processing)
- Documentation of ACE-V casework, for archiving, interchanges among examiners, and courtroom presentations
- Review of examiners' work, for quality assurance, validation and technical review, and conflict resolution
- Automated quality assurance
- Data interchange between automated systems
- Quantitative analysis for research

3.4 EFS philosophy

In leading CDEFFS and guiding the broader EFS process, my purpose was to find a consensus on how to represent the data content in a way interpretable by examiners and by software, and to avoid being a proponent for any specific approach. The process was iterative, with five draft specifications from 2007 through 2010 incorporating substantial revisions based on feedback.

As we considered which features and other attributes were appropriate, decisions for inclusion were based on providing a means to capture all content in an impression that an examiner would consider substantive during examination. Automated representations of features that did not correspond to attributes interpretable by examiners were omitted (e.g. Fourier representations). Much discussion often preceded group consensus on whether to include features. For example, the consensus of the group was that the position of pores should be included but not the size and shape of pores; the Chatterjee[69] categorization of ridge edge shapes was considered but rejected as excessively complex and difficult to reproduce; 3-dimensional ridge features were not included because they could not be interpreted using the 2-dimensional latent and exemplar images typically in use.

It is important to note that EFS itself is intended to be a spanning set of all of the features that an examiner can discern in an image. This does not mean that all of these features are appropriate for every use. When used for AFIS searching, there are implicit tradeoffs between examiner time, accuracy, and the number of successful searches. A business process analysis would optimize procedures to maximize the number of successful searches and/or overall accuracy, while minimizing examiner markup time. For AFIS searching, we developed the concept of EFS profiles: different EFS profiles define the extent of examiner markup used for a search, to provide different

[69] *[Chatterjee62]*

levels of incremental tradeoffs between examiner markup time and matcher accuracy.[70] When used for casework documentation, again not all EFS features are needed for every use: they are there so that if an examiner uses a feature as a basis for a decision, there is a standard representation; it is only necessary to mark pores if the examiner uses pores to justify a decision.

In my view it has not been my role solely to write a standards document, but to take whatever steps possible to ensure its adoption and effective use. Across a variety of disciplines, there are many purported "standards" that may have the imprimatur of a formal standards organization, but are unused or rarely adopted, and I feel it would a major missed opportunity if this were to happen with EFS. My plan for EFS was that for it to be effective it should have not just a technical specification, but also instructions for examiners, reference data and software, evaluations of its efficacy, and processes to aid in its adoption. The portfolio of work described here implements much of that plan.

1. ***Technical specification*** — A precise definition of a machine-readable data interchange format in a technical specification designed for engineers, for use in developing software tools. The specification should not just define the syntax (how each feature or other attribute is saved in a computer file), but also the semantics (when and how features are marked, defining not just what the presence of a feature means but also its absence). After multiple drafts starting in 2007, EFS was formally adopted by ANSI in 2011.[71]
2. ***Instructions for examiners*** — A technical specification is not sufficient to ensure appropriate use: instructions should be provided for latent print examiners on how to understand and use features. This would complement (and not contradict) the technical specification, but would provide more explanation and examples. These started as instructions for the examiners providing markup for the reference data in 2009, and after many revisions were published as a NIST Special Publication in 2013,[72] discussed in Section 4.1. The instructions have been implemented in software for training in the EFS Training Tool (Section 4.2), and are being included in the forthcoming ACEware, discussed in Section 4.3.
3. ***Reference data*** — Reference datasets are invaluable to implementers of standards, as they provide concrete examples and limit the likelihood of misunderstandings. For nascent standards, the development of reference data is important as a test of the completeness of the specification, and for EFS the process of building the reference datasets identified ambiguities in the draft EFS standards that were then clarified. The reference dataset was used for the NIST ELFT-EFS tests (Section 7.2).
4. ***Reference software implementation*** — To develop reference data, a software implementation is needed for both examiners and engineers to use. As with the reference data, the process of implementing drafts of the standard in software served to test the completeness of the specification and flagged areas for improvement. The FBI's Universal Latent Workstation (ULW) serves as the EFS reference implementation.[73]
5. ***Evaluations of AFIS use of EFS*** — The effectiveness of EFS should not merely be assumed, but evaluated, with the evaluations serving to provide feedback to improve future revisions of the standard. AFIS use of EFS can be evaluated in terms of both accuracy and interoperability. The NIST ELFT-EFS tests (Section 7.2) served as early evaluations of EFS, but future evaluations should be conducted in order to assess ongoing improvements to algorithms.

[70] *See Section 7.1.1 for further discussion of EFS profiles.*

[71] *[ANSI/NIST]*

[72] *[EFSMI]*

[73] *Disclosure: I worked with Tom Hopper in the initial design and development of ULW starting in 1998, and I still provide technical oversight for the Noblis team that enhances and maintains ULW.*

6. ***Evaluations of human use of EFS*** — As with evaluations of AFIS, evaluating human use of EFS must be ongoing, because the effectiveness of such a new approach will presumably change as it becomes familiar and incorporated into standard training and operating procedures. While none of these studies was intended specifically to evaluate EFS, each of these served as an evaluation of some aspects of EFS. Evaluations of EFS human markup included the Sufficiency for Value study (Section 6.2), the Sufficiency for Individualization study (Section 6.3.1), the Analysis to Comparison study (Section 6.3.3), and the Inter-Examiner Variation of Minutia Markup study (Section 6.3.4). EFS determinations were evaluated to some extent in each of the studies evaluating markup, and also in the Black Box study (Section 6.1.1). If EFS is incorporated into training and operational procedures, future evaluations should be conducted to determine their efficacy, and to suggest improvements to the standard and instructions.
7. ***Operational use*** — The true value of any standard is, of course, being used operationally. To make operational implementation possible, EFS was incorporated into the FBI's Electronic Biometric Transmission Specification (EBTS), and the EFS Profile Specification and Latent Interoperability Transmission were developed (Section 7.1). EFS is used by the FBI's Next Generation Identification (NGI) system and various other systems (e.g. the Western Identification Network (WIN, which serves 9 US states) and Orange County, California).

3.5 Standardization of features and attributes [74]

EFS provides a broader, more complete, and more detailed set of friction skin features than any other fingerprint features standard. It defines how and when to annotate a variety of level-1 (pattern class and ridge flow), level-2 (ridge path), and level-3 (ridge edge and pore) features, as summarized in the following tables. Figure 1 shows examples of EFS markup.

[74] *Derived from [ANSI/NIST]*

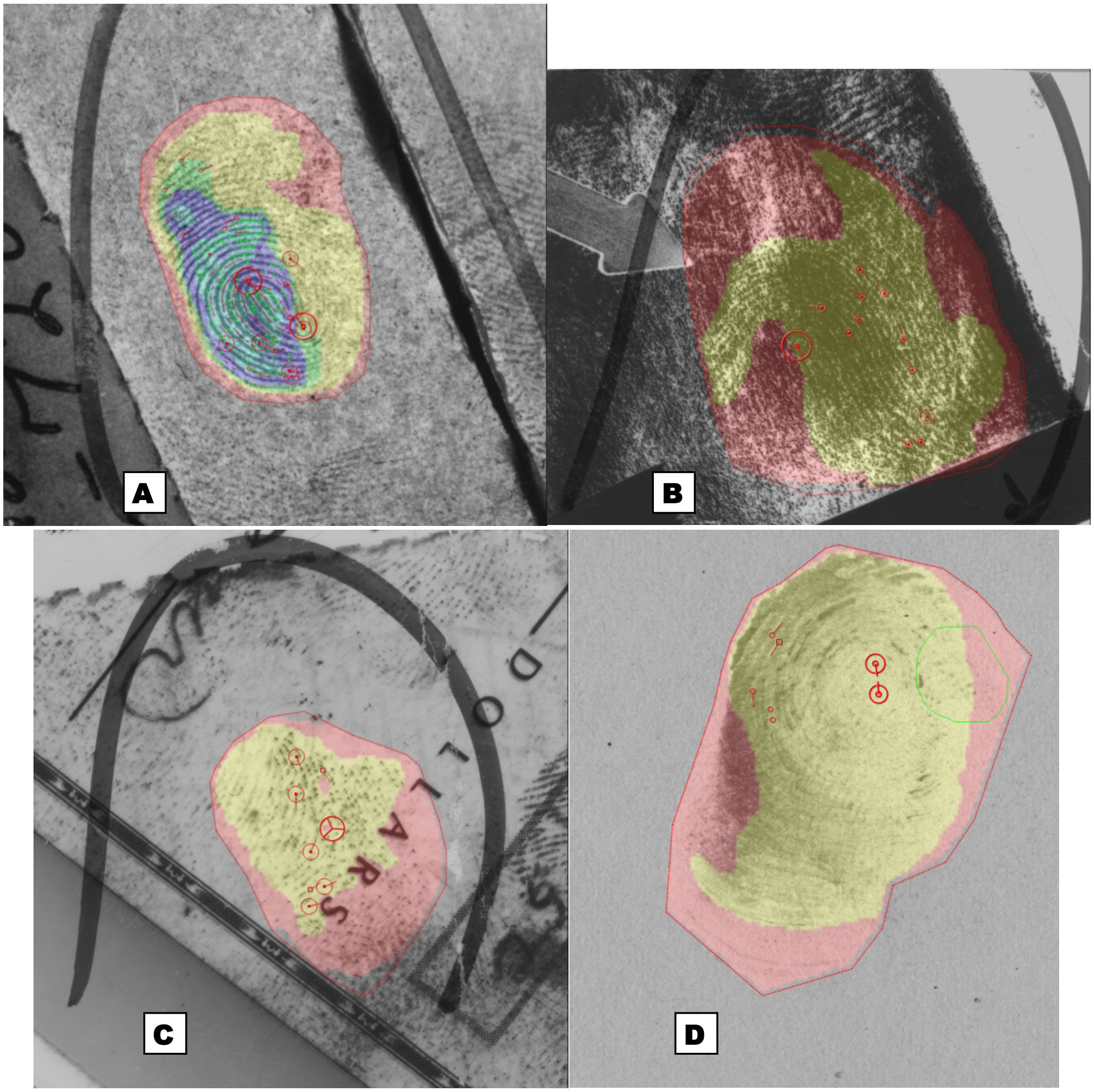


Figure 1: Examples of latents with EFS markup. In addition to minutiae, cores, and deltas, image A includes markup for dots (red circles), pores (green points), and incipient ridges (red lines); image B and C include minutiae with radii of uncertainty; image D includes dots and an area of overlapped impressions (green local quality issue polygon). For these four latent prints, the specific clarity and feature markups shown here were performed by examiners who assessed each latent as VID. Among all examiners, value determinations varied: A (100% VID; 5 examiners); B (57% VID, 14% VEO, 29% NV; 7 examiners); C (14% VID, 14% VEO, 71% NV; 7 examiners); D (2% VID, 47% VEO, 51% NV; 59 examiners).[75]

[75] *Images and value assessments from [SuffValue].*

3.5.1 Location and orientation fields

Region of interest (ROI)	Rectangle or polygon bounding region of interest within the image, separating a single contiguous impression from the background or other impressions. If a single image contains multiple impressions (overlapping or not), multiple EFS feature sets can be defined for an image, each with its own ROI.
Orientation	Indicates unknown orientation, or deviation from upright with orientation uncertainty. Indicating direction is wiser than rotating the image to upright, because rotating images (other than in multiple of 90°) necessarily degrades the image quality; note that software can temporarily rotate to upright for display but retain the unrotated image in the file without degradation.
Finger, palm, plantar position	The known or possible area of friction ridge skin from the fingers, palm, or foot. EFS was designed from its onset to accommodate not just fingerprints, but palm and plantar prints.

3.5.2 Ridge flow (Level-1) characteristics

Pattern classification	Henry/NCIC pattern classification; all possible classes should be indicated. Includes general class (arch, whorl, left/right loop), subclass (e.g. tented arch, central pocket loop), core-delta ridge counts, and whorl delta relationship (i.e. inner, outer or meeting whorl tracing). Only applies to prints from the distal segment of each finger or thumb.
Ridge flow map	Assessment of the direction of ridge flow for each location within the image, as shown in Figure 2. A ridge flow map is a more complete way of indicating level-1 data than pattern classification, as it indicates the shape of the pattern, can be used more effectively on partial fingerprints, and allows defining ridge flow for lower joints, palms, and feet. It can be used by AFIS for pattern-level screening of fingerprints.
Ridge wavelength map	Assessment of the distance between ridges for each location within the image.

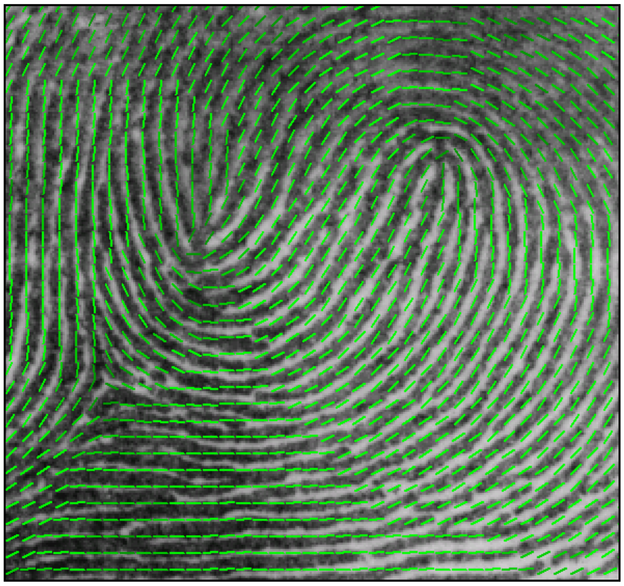


Figure 2: Ridge flow map.[76]

[76] *Figure from [EFSMI]*

3.5.3 Reference points

Cores	Defines the core or cores of a fingerprint, as well as core-like patterns in palm or plantar impressions. EFS cores are marked at the focus of the innermost recurving ridge. The location differs from the Henry/*Science of Fingerprints*[77] location: the EFS definition is more resilient to variations among different latents from the same finger, and is much more amenable to use within AFIS. The core location in *Science of Fingerprints* was highly dependent on minor variations in ridge detail within the core (e.g. the number of inner "rods", or the presence of "appendages" on the innermost recurving ridge)[78]; these details often do not repeat among different latents (or even exemplars). The complexity of the rules meant that automated core detection in AFIS could not accurately and reliably detect the *Science of Fingerprints* cores. Cores are defined by location (with optional radius of uncertainty) and direction (with optional range of uncertainty).
Deltas	Defines deltas for fingerprint, palm, or plantar impressions. Deltas are defined by location (with optional radius of uncertainty), three directions (with optional ranges of uncertainty for each), and type (e.g. left, right, or interdigital position number).
Center points of reference	The center point of reference of a fingerprint is used to define how centered a fingerprint is, as a feature, or for registration or orientation. Although the core may serve some of the same purposes, a center point of reference is defined for arches and provides a single center location for complex whorls, unlike cores. The center point of reference applies to any segment on the fingers, but does not apply to palmprints. The center point of reference is the sole EFS feature that can be located outside of the ***Region of interest***; this allows the estimated center of the finger to be marked even for an extreme side or tip, providing more information than merely indicating e.g. "left side". Center points of reference are defined by location (with optional radius of uncertainty) and type (e.g. lateral center for an arch).
Distinctive features	Distinctive features are unusually discriminating features that are not fully defined using other EFS features, such as scars, dysplasia/dissociated ridges, warts, blisters, or other abnormalities that interfere with normal ridge flow as distinctive features. These features are physical aspects of the friction skin itself, not issues specific to the impression (such as smudging, which is addressed in ***Local quality issues***). This field also provides for indicating unusual features such as oddly shaped cores: this allows an examiner to indicate particularly distinctive features relied on in making a determination. Distinctive features are defined as a polygon, with type and an optional comment.

[77] *[FBI85]*

[78] *[FBI85] p 14*

3.5.4 Minutiae

Minutiae	Minutiae are marked as bifurcations or ridge endings; complex types are marked as combinations of these types. All minutiae in the impression are to be marked, not omitting short ridges or minutiae near cores and deltas. This is a critical requirement, because otherwise if examiners see the same features but mark them differently, there is a false indication of differences. If a receiving AFIS does not use such minutiae, it is incumbent on that system to ignore them. EFS provides guidance on how to differentiate short ridges from dots and incipient ridges, and short bifurcations from spurs or protrusions. The location for bifurcations (at the "Y" of the ridge) and ridge endings (at the "Y" of the valley) was a consensus decision among AFIS vendors for interoperability: previous proprietary formats varied, especially for the placement of ridge endings. If the precise location for a minutia cannot be determined, a radius of uncertainty includes the area of possible locations; this is appropriate in "2 ridges in, 3 ridges out" instances, as well as when the type is unknown (i.e. could be either a bifurcation or ridge ending). This allows for the differentiation of minutiae with precise vs approximate locations: without such, any automated review or analytical tools would have to treat all locations as if they were precise. EFS provides guidance on how to determine minutia direction (theta), and allows for indicating direction uncertainty. Unusual ridge path features (such as significant deviation in ridge path, right angles in a vestige area, or very sharp changes in direction) can be marked as ***Distinctive features***. Confidence in the presence or absence of minutiae is indicated by ***Local clarity/confidence map***: minutiae in green (or better) clarity areas are definite, and the examiner is confident that every minutia in that area is marked; the presence or absence of minutiae in yellow areas is debatable.
Minutiae ridge counts	Indicates the number of ridges crossed between any two minutiae. If the counts cannot be determined precisely, counts can be stated as ranges, minima, or maxima.

3.5.5 Ridge path

Ridge path segments	Defines each segment of a ridge (the portion of a ridge that connects two minutiae) as a feature. Confidence in the ridge path is indicated by ***Local clarity/confidence map***: the ridge path in green (or better) clarity areas is definite, whereas the ridge path in yellow areas is debatable. Using ridge path segments as a feature allows documentation of ridges in sequence, defining the topological relationship between minutiae that is lost in a minutia-only representation. In performing comparisons, examiners have often underscored that they rely on "ridges in sequence," in which they use the ridges for comparison: the minutiae are really attributes of ridge path, not the other way around.[79]
Skeletonized image	Reduces the impression to an image with thinned representations of each ridge; also known as a ridge tracing. The skeletonized image can be converted to or from a set of ***Ridge path segments***.

[79] *[Ashbaugh99, SWGFAST-Memo05]*

3.5.6 Additional features

Dots	A dot is a single or partial ridge unit that is shorter than local ridge width, marked by location and (optionally) length. Confidence in the presence or absence of dots is indicated by ***Local clarity/confidence map***: dots in blue (or better) clarity areas are definite, whereas the presence or absence of dots in green or yellow areas is debatable.
Incipient ridges	A thin ridge, substantially thinner than local ridge width, defined by its end points. Confidence in incipient ridges is handled just as with ***Dots***: blue (or better) clarity areas are definite, green or yellow areas are debatable.
Creases and linear discontinuities	Includes the permanent (named) flexion creases as well as linear discontinuities (minor creases, cracks, cuts, and thin or non-permanent scars). Defined by end points and type.
Ridge edge features	Protrusions, indentations, and discontinuities are marked by location. Confidence in ridge edge features is handled just as with ***Dots***: blue (or better) clarity areas are definite, green or yellow areas are debatable.
Pores	Pores are marked by location. Confidence in the presence or absence of pores is indicated by ***Local clarity/confidence map***: dots in aqua clarity areas are definite, whereas the presence or absence of dots in blue or green areas is debatable.

3.5.7 Image capture attributes

Ridge quality/confidence map	The ridge quality/confidence map is a standard color-coded means of indicating the clarity of the print. This is the means by which the examiner (or process) indicates feature confidence: whether the features marked at a given location are definitive or debatable. (see full description in Section 5.3)
Local quality issues	Indicates area(s) in the image containing quality or transfer issues that indicate that the anatomical friction ridge features may not have been accurately represented in the image. Examples include digital artifacts (e.g. from image compression), tonal inversion of a portion of the image, overlapped impressions, smeared areas, or unusually distorted areas. These are distinct from ***Distinctive areas***, which indicate physical aspects of the friction skin itself.
Latent processing method	Indicates the technique(s) used to process the latent fingerprint. Multiple methods are indicated if they contributed substantively to the visualization of the image. Codes are defined for a wide variety of processing methods.
Latent substrate	Indicates the type of surface on which the latent was deposited. Codes are defined for a variety of common substrates, as well as descriptive text.
Latent matrix	Indicates the substance that forms the impression. Codes are defined for a variety of common matrices, as well as descriptive text.

3.5.8 Special cases

Tonal reversal	Indicates if all or part of the image is reversed black for white.
Possible lateral reversal	Indicates if it cannot be determined whether the image is flipped left for right, such as in some prints on transparent tape.
Possible growth or shrinkage	Used in the unusual circumstance that the impression is believed to have changed size or scale from potential comparisons (e.g. deceased subjects with desiccated skin, swollen skin due to water exposure, or comparing child and adult fingerprints).

3.5.8 Special cases

Evidence of fraud

Indicates that there is basis for determination that the image may be fraudulent. There are four types of fraud:

- ***Evasion*** includes actions that prevent/lessen the likelihood of matching, such as by degrading or obscuring physical characteristics or mutilating fingers. Examples are acid balding of fingers or use of a knife or laser to alter the fingerprints.
- ***Spoofing*** includes purposefully attempting to be identified as a different person in a biometric system by modifying biological characteristics or using fabricated characteristics. Examples are using a rubber finger, gelatin fingerprint attached to a real finger, or image of a fingerprint to fool a biometric reader.
- ***Forged evidence*** is forensic evidence that was fraudulently placed on the surface from which it was collected, using another mechanism or device than the natural contact with friction ridge skin. An example is using a rubber lifter to move a fingerprint from its actual source to another source.
- ***Fabricated evidence*** is forensic evidence that never existed on the surface from which it was supposedly collected. An example is a crime scene examiner deceitfully mislabeling the source of images or lift cards.

3.5.9 Correspondence

Corresponding points or features

Indicates which features or points correspond among two or more images. Used to document the basis for examiners' comparison/evaluation conclusions.

Correspondences can be indicated for predefined features, including point features (such as minutiae, dots, or pores), areas (such as distinctive characteristics), lines (incipient ridges or creases), or paths (ridge path segments). Correspondences can also indicate arbitrary points not associated with specific features. Types of correspondences include

- ***Definite correspondence***, such as used to justify an individualization.
- ***Possible or debatable correspondence***, such as used to indicate arguable correspondences in inconclusive comparisons (e.g. *I see it, but I'm not convinced*), or to indicate reference points used in exclusions (e.g. *These features may look like they correspond, but that logic leads to discrepancies*).
- ***Definite lack of correspondence***, used to indicate discrepancies; indicates a feature that existed in the other image that is definitely not present, optionally indicating the location where the feature would be expected to be if it were present.
- ***Inconclusive correspondence***, used to indicate features that are not visible because they lie outside the corresponding area or in an unclear region.

Corresponding features are used by AFISs to indicate the points the AFIS determined were in correspondence between a latent search image and each of the returned candidate matches.

Corresponding features enable a variety of models and tools related to comparison, including sufficiency analyses, distortion models, comparison user interface tools, and probabilistic/likelihood models.

Area of correspondence

The area of correspondence allows the definition of polygons to indicate the common region of usable ridge detail present in the images being compared, as shown in Figure 3.

Relative rotation of corresponding print

Indicates the relative rotation necessary for two impressions to be compared. This is most relevant in instances in multiple images are compared, such as indicating how a latent should be rotated to correspond to each of a set of candidate exemplars; this allows the user interface to display the images rotated to correspond to each other.

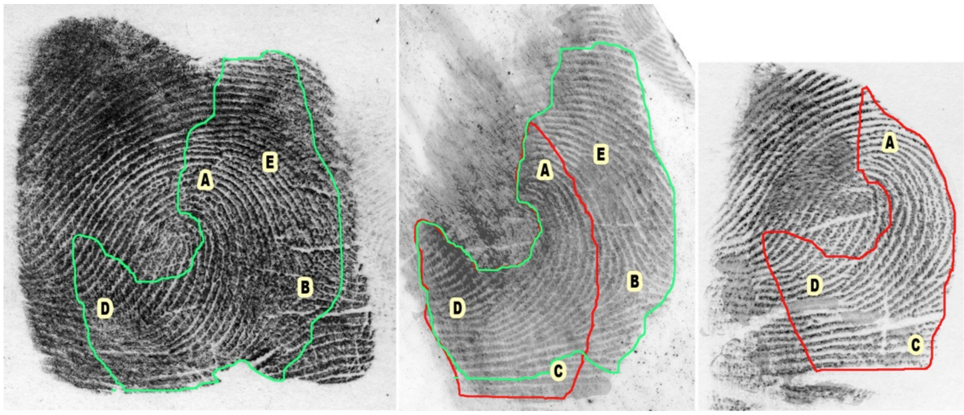


Figure 3: Example of corresponding points and areas for three impressions. The corresponding area between the left two impressions is in green; between the right two impressions is in red. Only a subset of points are marked in this example: minutiae A and D are in all three impressions; minutiae B and E are in the left two impressions; minutia C is in the right two impressions.

3.5.10 Determinations

Examiner analysis assessment

Indicates the examiner's value determination from analysis of the impression:

- ***Value*** — The impression is of value and is appropriate for further analysis and potential comparison; sufficient details exist to render an individualization or exclusion decision. Also known as ***Value for identification (VID)***.
- ***Limited value*** — The impression is of limited, marginal, value; it is not of value for individualization, but may be appropriate for exclusion. Also known as ***Value for exclusion only (VEO)***.
- ***No value*** — The impression is of no value for comparison, is not appropriate for further analysis, and has no use for potential comparison.
- ***Nonprint*** — The image is not a friction ridge impression (this value is necessary for completeness, so that there is an appropriate category for use if examiners have to triage unsorted images prior to examination).

The analysis assessment includes the examiner's name, affiliation, and date.

3.5.10 Determinations

Examiner comparison determination

Indicates the examiner’s determination from comparison/evaluation of two impressions:
Individualization: The two impressions originated from the same source.

Inconclusive includes the rationale for the decision:

- ***Inconclusive, but with corresponding features noted*** — Corresponding features are present, but not to the extent sufficient for individualization. Sometimes described as a qualified conclusion.
- ***Inconclusive, but with dissimilar features noted*** — Non-corresponding features are present, but not to the extent sufficient for exclusion. Sometimes described as a qualified exclusion.
- ***Inconclusive due to no overlapping area*** — There are no potentially corresponding areas, such as when the latent is an extreme tip and the exemplar does not include the tip.
- ***Inconclusive due to insufficient information*** — Used if the specific other types of inconclusive determinations do not apply.

Exclusion includes the extent of the exclusion.

- ***Exclusion of source*** — The two impressions originated from different sources of friction ridge skin, but the subject cannot be excluded (e.g., they could be from different fingers).
- ***Exclusion of subject*** — The two impressions originated from different subjects. This generally means that all potentially corresponding areas of friction ridge skin are available for the subject, or the anatomical location of the latent can be determined.

Comparison determinations can be flagged as “preliminary,” which permits saving of work in progress, or is used to indicate the determination of an examiner prior to verification. Comparison determinations can be flagged as “complex,” based on the quality and quantity of features, low specificity of features, significant distortion, or disagreement among examiners.[80] The comparison determination includes the examiner’s name, affiliation, and date.

[80] *Complex comparisons are defined in SWGFAST Standards for examining friction ridge impressions and resulting conclusions.*

3.5.11 Annotations

Feature presence/absence fields	Allows the examiner to indicate the extent of the examination for a given feature type, defining whether the absence of a particular type of feature means that there are no instances of that type of field present, as opposed to simply not having been marked. For example, this allows to indicate that there are no incipient ridges present in a given impression, as opposed to none having been marked because analysis was not conducted for incipients. There was discussion in CDEFFS of expanding this concept to indicate that additional unmarked features were present, for example to indicate that some minutiae but not all minutiae were marked, such as would be sufficient for a determination of value. This was not included due to two concerns. First, permitting or encouraging examiners to mark only a portion of any one type of feature would prevent the comparison of markup, such as for comparison of Analysis and Comparison minutiae or the features marked by different examiners during conflict resolution. Secondly, the availability of the option might prove to be a temptation for examiners to use excessively: the concern is that enough examiners would use it when they thought they might have missed any features that it would no longer be useful.
Method of feature detection	Indicates the source of feature markup: automated process (including algorithm version) or specific examiner (name and affiliation). For example, can indicate if an automated workstation first performed minutia detection, with subsequent markup by two examiners on different occasions.
Friction ridge quality metric	Stores results of automated quality metrics. This field is not specific to any particular method or algorithm.
Feature color and comment	Enables a latent print examiner to annotate individual features with color for display and/or comment. For example, allows designating colors for minutiae according to the GYRO annotation method,[81] color-coding features based on the extent of examiner agreement during conflict resolution, or assigning an explanation of a specific feature was used in examination.
Comment	Free text comment.
Temporary lines	Used by a latent examiner to annotate a friction ridge image with temporary lines, generally for use as reference points in making a comparison. These lines are solely for the individual examiner's use and reference – there are no implied semantics through the use of this field. This was based on feedback that examiners frequently want to mark e.g. tick marks on intervening ridges, and without a distinct field they were misusing other fields for this purpose.

3.5.12 Feature set profiles

Feature set profile	Indicates an EFS Profile, which defines the sets of EFS features used in latent AFIS searches. Different EFS profiles define the extent of examiner markup used for a search, to provide different levels of incremental tradeoffs between examiner markup time and matcher accuracy. (See Section 7.1 for further discussion of EFS profiles.)

[81] *[Langenburg11]*

3.6 Other changes to the ANSI/NIST standard related to documentation of latents

In addition to EFS I was also instrumental[82] in adding other functionality to ANSI/NIST-ITL to make the standard more appropriate for the interchange of casework documentation, which needs to not only handle the specific images and their annotations, but bundle them with ancillary information. Therefore, along with EFS, the 2011 standard was revised to include source images, context images, and related data. Often, the images used in casework are derived from other images, such as when a latent image is cropped from an original source image that was large, color, and contains multiple impressions; the new representation includes both images and defines the relationship. Context images provide information such as where the latents were found on the evidence. Related data can include information such as case reports. These changes to the standard allow keeping the casework documentation in a single bundle, which can be critically important when interchanging documentation among different organizations. For example, in the review of the Mayfield misidentification, the context of the latent on the evidence would have shown the examiners which finger left the impression;[83] however, since the finger position was unknown, the examiner did not know that the misidentification to Mayfield was using the wrong finger position. Policy and operating procedure need to determine when contextual information is appropriate vs. biasing — but the interchange standard should not make it impossible.

I also chaired the committee to incorporate plantar (foot and toe) impressions into the standard, so that all friction ridge impressions are addressed. While foot impressions are not common, they are key evidence in some cases.

[82] I initiated the concepts, and provided guidance to my colleague John Mayer-Splain, chair of the Source reference representation / associated context data records working group.

[83] Stephen Meagher, personal communication.

Chapter 4 Standardizing examiners' annotation procedures

As discussed in Section 2.2, there are clearly identified needs to improve documentation of casework. There is a need for examiners to be able to unambiguously and consistently document the features they use to justify their decisions in Analysis and Comparison, in a format amenable to automated processing.

Whether such detailed documentation should be mandated for ***all*** casework is a policy issue that should evaluate the tradeoffs between costs (examiner time, possibly fewer examinations being conducted) and benefits (improved transparency and quality assurance procedures). At a minimum, I believe that such documentation should be mandated in specific instances, such as for conflict resolution (when the verifier and original examiner do not concur), for any cases that hinge on a single fingerprint individualization, for examinations that the examiner deems complex or particularly difficult, and for courtroom presentations.

If such detailed documentation were mandated for all casework, that would enable a variety of automated quality assurance tools that would not otherwise be practical. For example, automated flagging of examinations with extensive changes between Analysis and Comparison is only possible if the Analysis and Comparison features are always marked. Quality assurance procedures can take special steps (such as extra verification) for decisions that appear to be high risk, but only if the features used to justify the decision are documented in a machine-readable format. It is not just comparisons that may warrant such documentation: determining that a latent is of no value is a short-circuit decision (since in most agencies no one will ever look at that latent again), but examiners often disagree whether a latent is of value.[84]

While EFS provides a basis for more rigorous documentation, the mere existence of EFS does not magically result in its uniform and widespread usage across the community. Here we discuss tasks conducted to assist in the adoption of EFS by latent examiners:

- *Markup Instructions for Extended Friction Ridge Features*[85] provides instructions to examiners in marking friction ridge impressions to maximize consistency among examiners.
- The *Extended Feature Set Training Tool*[86] implements the EFS Markup Instructions in free, web-based software.
- *ACEware* is a task in progress that will use EFS as the basis for training and evaluating examiners in detailed documentation of ACE, and will facilitate documentation of operational casework.

These tasks collectively will help satisfy some of the requirements of NRC Recommendation #6 by improving and standardizing latent examiner training and practices, providing a basis for improving the proficiency of latent print examiners, providing a means for evaluating the effectiveness of their training, and collecting data that can be used to improve the effectiveness of the latent business process.

Note that operational adoption of EFS is not hypothetical: EFS is in operational use due to its incorporation in several AFISs (most notably the FBI's NGI, starting in 2013).[87] The FBI's Universal

[84] *[SuffValue]; Ron Smith personal communication*

[85] *[EFSMI]*

[86] *[EFSTT]*

[87] *Discussed further in Chapter 7*

Latent Workstation (ULW) has implemented EFS since 2008 (originally based on draft specifications).

4.1 Markup Instructions for Extended Friction Ridge Features [88]

EFS, as incorporated in ANSI/NIST, provides a formal definition of friction ridge features for use by engineers; as a technical specification, it is not (and is not designed to be) very accessible to non-engineers. *EFS Markup Instructions* defines and explains EFS specifically for the use of latent print examiners in order to maximize consistency among examiners, including examples and specific guidance for latent print examiners, and minimizing references to technical details of the file format. No previous document has provided this comprehensive level of detail in defining how to annotate finger- and palmprint features. The *EFS Markup Instructions* also are of value for engineers: they may also be used by feature extraction and matcher algorithm developers as a basis for their expectations as to how examiners mark features, because the goal of markup as conducted for AFIS searching is for the automated algorithms to mirror examiners' markup, and vice versa (discussed further in Chapter 7).

Much of the current document initially started as examples used in CDEFFS workshops (2005-2006), instructions to examiners on marking EFS features for the ELFT-EFS datasets (2007-2009), and guidelines to examiners on marking latent quality developed as part of the Latent Quality Study (2007-2008). The *EFS Markup Instructions* were published as a NIST Special Publication in January 2013. The current version focuses on features that can be used by AFIS and does not address all EFS features (for example, it does not include comparison features, creases, pores, or local quality issues); future revisions should address the rest of EFS.

The ultimate goal of the *EFS Markup Instructions* (and its successor revisions) is to be a standard instruction manual for latent print examiners on how to perform detailed documentation of analysis and comparison, and how to mark features for AFIS searching.

4.2 Extended Feature Set Training Tool

EFS Markup Instructions is merely a document: for it to be effective, the concepts need to be incorporated into training, with extensive examples and exercises. The *EFS Training Tool* does just that: it is an interactive guide to markup using EFS, freely available to anyone over the Internet via a Web browser (at http://www.nist.gov/forensics/EFSTrainingTool/). The *EFS Training Tool* :

- provides an interactive tool for learning the Extended Feature Set
- presents examples and exercises with a range of image clarity and difficulty to provide growth opportunities even for experienced examiners
- trains examiners to use markup consistent with best AFIS accuracy
- develops a framework for future expansion of examiner training
- fosters greater consistency in latent print markups
- furthers the use of a common markup method for the latent print identification community
- provides a training tool that is independent of proprietary AFIS rules

EFS Training Tool is not intended to be a complete training program: it is a self-guided tool to introduce examiners to EFS.

4.3 ACEware (work in progress)

There is a need for greater standardization of ACE-V documentation through more rigorous and consistent training, and through tools for operational casework. ACEware (due to be completed in 2017) seeks to address that problem by providing a platform for standards-based detailed

[88] *Derived from [EFSMI]*

annotation of the latent print examination process. ACEware is an innovative software tool for use in training new latent print examiners in standard, reproducible documentation of examination — as well as for use by experienced case-working latent print examiners in documenting actual casework. ACEware builds upon ULW, which allows users to create, edit, view, and manage latent fingerprint transactions. ACEware extends ULW by providing training functionality, extending its functionality for non-AFIS casework, and providing the capability to create standardized data sets for research and training. ACEware facilitates self-led and instructor-led examiner training and evaluation. Because ACEware documentation is based on EFS, detailed documentation of a complex latent print examination can be exchanged with other examiners, or archived for future review. ACEware is being developed by Noblis under NIJ funding, in collaboration and consultation with several Federal, state, and local law enforcement partners.

The *ACEware* project, when complete, aims to serve two roles in furthering the standardized use of EFS, as software to support training, and as operational software for the documentation of the latent examination process. *ACEware* is intended to

- facilitate classroom tutorials, self-training, and peer evaluation
- increase the automation and standardization of operational casework within the ACE-V process
- provide a standardized approach for applying the ACE method, building upon SWGFAST standards and EFS
- facilitate increased consistency and proficiency in feature selection by latent print examiners
- help standardize presentation formats for data exchange and evidentiary presentations
- provide a basis for improving latent workload management
- enable direct interoperability between Analysis phase annotation and AFIS searching
- facilitate the development of standard datasets for training and evaluation
- provide a basis for metric-based quality assurance
- provide a standard platform for collection of data and performance metrics
- serve as a basis for analytical examiner performance evaluations

ACEware is being developed as an extension to the FBI's ULW, so that the new functionality will be available to a broad range of users. Its use will complement the *EFS Training Tool*, which can be seen as a lightweight, limited functionality version of *ACEware*, and will continue to serve as an introductory tool available via any web browser; *ACEware* is part of a complete system with much more detailed training functionality as well as integration with operational casework.

4.4 Standardizing casework annotation and exchange

In Section 7.1 we will discuss the *Latent Interoperability Transmission Specification (LITS)* as an application profile that defines AFIS transactions for exchange among state and local law enforcement agencies. However, in addition to AFIS transactions, LITS defines transactions for non-AFIS casework exchange and archiving. *LITS* defines standard transactions for documentation, exchange between human examiners, for validation and quality assurance processing, for quantitative analysis, and for archiving information associated with the ACE-V process. LITS casework transactions include:

- The Comparison (COMP) transaction provides a standard format for two or more friction ridge images, feature markup, and determinations. A COMP transaction can serve as a standard basis for documenting and archiving an examiner's comparison/evaluation determination (individualization, exclusion, or inconclusive) with the detailed basis for that determination. COMP transactions can also be used as a standard basis for interchange of images for blind verification, and images with features and determinations for non-blind verification and technical review. A COMP transaction may also be used to annotate decisions comparing a print with AFIS search results.

- The Analysis (ASYS) transaction provides a means to provide detailed markup and annotation for a single impression that is not associated with other prints. An ASYS transaction can serve as a standard basis for documenting and archiving an examiner's analysis (value) determination with the detailed basis for that determination.
- The Casework Exchange (CWE) transaction provides a format for latent examiners to collect all information related to a case within a single transaction. This transaction permits the storage and exchange of fingerprint; palmprint; plantar; facial/mugshot; scar, mark and tattoo; iris; deoxyribonucleic acid; and other biometric sample and forensic information that may be used in the identification or verification process of a subject. A CWE transaction may contain the original source representations of images, e.g., a high-resolution color image containing multiple latent fingerprint images. CWE transactions may also contain associated contextual information, such as an image of the area where latent fingerprints were captured, and text documents such as crime scene reports.

I believe that COMP, ASYS, and CWE transactions provide a basis for long-term standardization of how latent print examinations are documented, exchanged, verified, reported in legal contexts, and made available for quantitative quality assurance.

Chapter 5 Defining and measuring quality, clarity, and distortion

In this chapter we discuss three studies of latent quality, clarity, and distortion:

- The *Latent Quality Study* (Sections 5.2 through 5.6.3) involved conducting a detailed survey of how the quality and clarity of latents and exemplars are assessed within the latent fingerprint community, developing guidelines and metrics for describing the clarity of friction ridge impressions, and developing software tools to provide objective, reproducible methods for assessment of friction ridge impression clarity. The study resulted in two publications, the definition of ridge clarity/confidence used in EFS, guidelines on assessing clarity that were incorporated into [EFSMI], and prototype software with algorithms to automatically calculate clarity metrics.
- The *Latent Quality Metric* project built on the results of the *Latent Quality Study*, and incorporated clarity/quality metrics into the Universal Latent Workstation (ULW).
- The *Distortion Study* (Section 5.9) developed metrics for quantifying and visualizing linear and nonlinear fingerprint deformations, and software tools to assist examiners in accounting for distortion in fingerprint comparisons; the results are in the publication process.

5.1 Terminology: Quality, clarity, and utility

I have found over years of experience in the forensic and biometric sides of the fingerprint disciplines that the term "quality" is used in various ways (e.g. [89,90]). According to SWGFAST, "quality" is synonymous with "clarity."[91] However, the biometrics community follows engineering practice in the use of the term "quality", delineating three distinct components of sample quality:[92,93]

- ***Character*** refers to the intrinsic data content of the inherent physical features of the subject.
- ***Fidelity*** is the degree to which a sample is an accurate representation of the original features.
- ***Utility*** refers to the value of a sample in terms of suitability for a specified purpose.

Note that "fidelity" is used in biometrics usage to refer to the same concept as SWGFAST's "clarity." Because in biometrics usage fidelity is a component of quality, but in SWGFAST usage clarity is synonymous with quality, the term "quality" can be ambiguous when writing for both audiences at once.

There are three sub-categories of fidelity, as defined by ISO-29794-1:[94]

- ***Acquisition fidelity*** refers to the fidelity of the impression — the accuracy (or possible loss of data) of the process by which the physical friction ridge skin results in an impression. For

[89] *[Hicklin02]*

[90] *[Tabassi04]*

[91] *[SWGFAST-Terminology11]*

[92] *[ISO-29794-1], discussed in [Hicklin06]*

[93] *In addition to sample quality, in [Hicklin06] we also discuss "metadata quality," which are those aspects of quality that cannot be determined through analysis of an image, such as database and administrative problems.*

[94] *[ISO-29794-1] Here I have adjusted the original biometric-specific definitions so be more applicable to latent print examination.*

latents, poor acquisition fidelity is caused primarily by the uncontrolled deposition, by also by problems caused by the substrate, matrix, data collection, and processing.

- ***Sample processing fidelity*** refers to the fidelity of the image — the accuracy of the process by which the impression is converted into the image being used, including (for digital images) scanning, compression, formatting, cropping, or image processing.
- ***Extraction fidelity*** refers to the fidelity of the features — the accuracy with which features can be extracted from the image, which in biometric systems relates to the accuracy of the feature extraction algorithms used. In latent print examination, this would be related to the skill of the examiner in accurately and reliably finding and using the features in the image.

In this thesis, I use the terminology in this way:

- ***Quality*** is a general term that includes character, fidelity, and utility.[95] From another perspective, conclusions are based on an examiner's assessments of the quantity of features, their relationships, and their specificity (all aspects of "character", overlapping with "utility"); as well as the clarity and relative distortion of features ("fidelity").[96] Quality includes the concepts of suitability (value) and difficulty.
- ***Clarity*** refers to fidelity: the extent to which the physical features are faithfully represented in an impression/image being used [97] (acquisition and sample processing fidelity), which relates directly to the confidence that the presence, absence, and details of features can be discerned with confidence (extraction fidelity). As clarity decreases, feature uncertainty increases; some features may not be discerned, image artifacts may be erroneously treated as features, and feature details may be misinterpreted. Clarity is unrelated to the quantity of features in an impression: the ability to discern the presence/absence and attributes of features is independent of the number of features present. For example, a high-clarity area may include no features, such as a clear open field of ridges that contains no minutiae.
- The ***utility*** of an impression relates to how effectively it can be used for a specific use, i.e. human or automated comparison. Here, I use "quality" to refer to the general usefulness of an impression, and use "utility" for those cases in which the usefulness of an impression differs depending on the specified purpose. For example: two clear impressions of a finger in which the friction ridges are entirely dissociated (e.g. dysplasia) would be extremely distinctive when compared by a human examiner, but an automated matcher that depended on typical ridge flow and minutia-based models might consider the impressions unusable; a latent fingerprint of an extreme side that does not include a core or delta is unlikely to have a corresponding region in the exemplars in a database, and therefore has lower utility than another print that has the same clarity and quantity of features but includes the core.
- ***Distortion*** is a failure of acquisition fidelity generally caused by the plasticity of the skin in making an impression. Often, the distortion is apparent in a single impression and can be determined by inspection during Analysis. However, frequently there is no apparent distortion when viewing a single impression, but the configurations of the corresponding features differ substantially when two impressions are compared: I refer to this as ***relative distortion***.

[95] *In retrospect I would have used "clarity" instead of "quality" in the names for the EFS Local quality / confidence map and Latent Quality Assessment Software (LQAS).*

[96] *[SWGFAST-Conclusions13, Locard20]*

[97] *[Ashbaugh99]*

5.2 The Latent Quality Survey [98]

Our 2007 survey of latent print examiners identified the techniques and practices used by latent print examiners to assess the clarity/quality of friction ridge impressions. In that study, 86 latent print examiners assessed the quality of latent and exemplar fingerprint images. Out of a total pool of 1,090 fingerprints, each examiner reviewed approximately 70 fingerprint images, resulting in a total of 5,245 image reviews. For each image, each examiner used a custom software application to mark areas within each impression to indicate the degree of confidence in discerning the features in the image; confidence was marked separately for level-1, -2, and -3 features. Note that clarity was explicitly defined with respect to feature-level confidence. In addition, the examiners provided an overall assessment of each image by indicating whether the image was of value for individualization and/or exclusion, and by indicating the overall difficulty anticipated in performing a comparison using the image (assuming sufficient quality and overlapping area in the exemplar print).

For analysis of results, an “Overall Quality” (OQ) measure combined the examiners’ value and difficulty assessments into a 0-6 scale, ranging from no value (OQ=0), value for exclusion only (OQ=1), of value but very difficult (OQ=2), through of value and very easy (OQ=6).

The results showed the extent of interexaminer variability in assessing overall quality (value and difficulty) and local clarity, and how examiners’ overall quality assessments related to the previous “good,” bad,” and “ugly” (GBU) assessments used in the NIST SD27 dataset.[99] There is general concurrence in human assessments of quality, but there is enough variation between examiners that clear and uniform definitions are needed, which then could be used in training in order to effect more consistent usage. Langenburg observed that inter-examiner variation in annotations can be reduced through training. [Langenburg12a]

There is a strong relationship between the examiners’ assessments of overall quality and the size of the area of clear ridge detail within each fingerprint. There is also a strong relationship between the accuracy of examiners’ pattern classification with overall quality and the size of the area of clear ridge detail.

The latent quality survey resulted in the development of the EFS standard for defining and depicting clarity, described in Section 5.3. Analysis of the examiners’ subjective assessments of fingerprint quality revealed information useful for the development of annotation methods, guidelines, metrics, and software tools for assessing fingerprint quality, as described in Sections 5.4-5.7.

5.3 EFS local clarity map [100]

One of the key results of the latent quality survey was the development of a simplified graphical means of defining clarity at each location in an image in terms of examiner confidence with a specified color-coding scheme. These “clarity maps” provide an intuitive visual depiction of friction ridge clarity. The clarity maps that were developed in the *Latent Quality Study* were incorporated into EFS as the “Local quality / confidence map”.

During the Analysis or Comparison phases of ACE-V, latent print examiners generally follow a series of conscious or unconscious steps when assessing each feature. Consequently, the analysis of clarity can be reduced to a series of assessments: of the presence of any friction ridge information, the continuity of overall ridge flow, the continuity of the paths of individual ridges, and whether features within individual ridges can be discerned. Figure 4 shows this decision process as a series

[98] *Derived from [LQSurvey].*

[99] *[NIST-SD27]*

[100] *Derived from [ANSI/NIST, AssessingLC, EFS-MI]*

of yes/no questions, resulting in a color-coded categorization of local clarity. In essence, color-coding with respect to minutiae is a stoplight: green indicates definitive minutiae, yellow indicates debatable minutiae, and red is used for the areas of the impression without usable minutiae. The areas outside of the region of interest are not colored, indicated in black. Unlike the GYRO and Laird approaches, EFS addresses areas with clear level-3 details in blue: I see this as critically important because while quantitative models can be readily based on minutiae, level-3 similarities are often difficult to quantify; areas in blue provide a basis for explaining determinations that are not justified using minutiae alone.

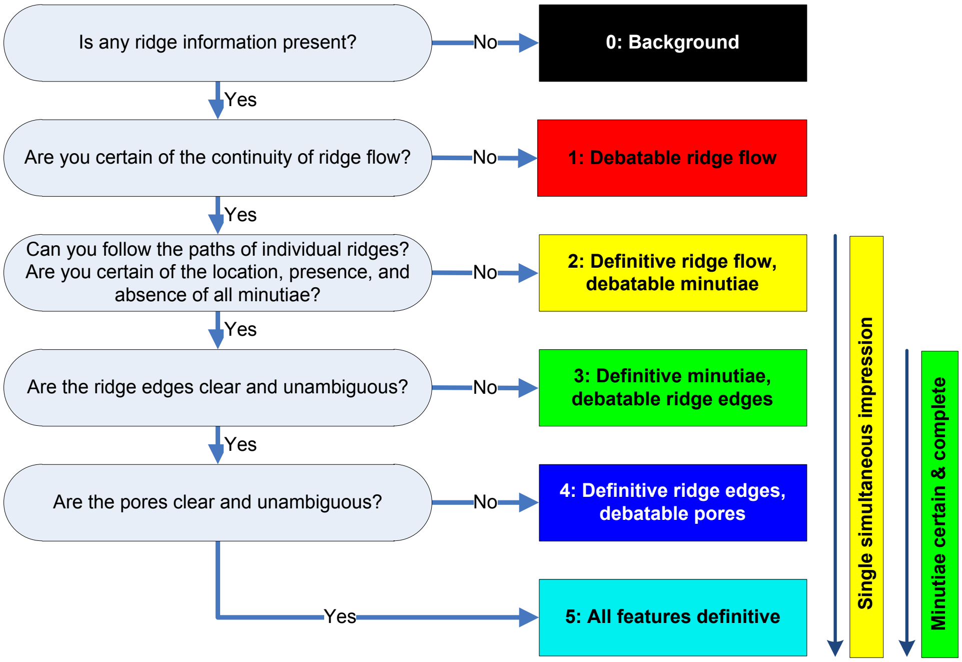

Is any ridge information present?
No
0: Background
Yes
Are you certain of the continuity of ridge flow?
No
1: Debatable ridge flow
Yes
Can you follow the paths of individual ridges? Are you certain of the location, presence, and absence of all minutiae?
No
2: Definitive ridge flow, debatable minutiae
Yes
Are the ridge edges clear and unambiguous?
No
3: Definitive minutiae, debatable ridge edges
Yes
Are the pores clear and unambiguous?
No
4: Definitive ridge edges, debatable pores
Yes
5: All features definitive
Single simultaneous impression
Minutiae certain & complete

Figure 4: Decision process for the assessment of local clarity as used in EFS local quality/confidence maps.[101]

Clarity maps provide a person (or software program) reviewing the image a standard, straightforward means of assessing the size and degree of clarity within various portions of the image. While the exemplar in Figure 5 can be described in words, the clarity map immediately conveys that the examiner found two areas (colored red) without continuity of ridge flow, and larger areas (colored yellow) in which minutiae and individual ridge paths are debatable and may therefore potentially contain false or missed features. Different examiners may differ in their assessments of images: this approach provides a means of indicating what a given examiner sees in an image; comparison of maps between multiple examiners can be used to depict the extent of (dis)agreement in their assessments.

[101] *Although the EFS local quality/confidence map definition differentiates between Blue and Aqua, in several years of working with examiners on clarity markup, I have found that the additional Aqua category adds complexity without any additional value. Clear ridge edges are highly useful; the utility of pores is more debatable.*

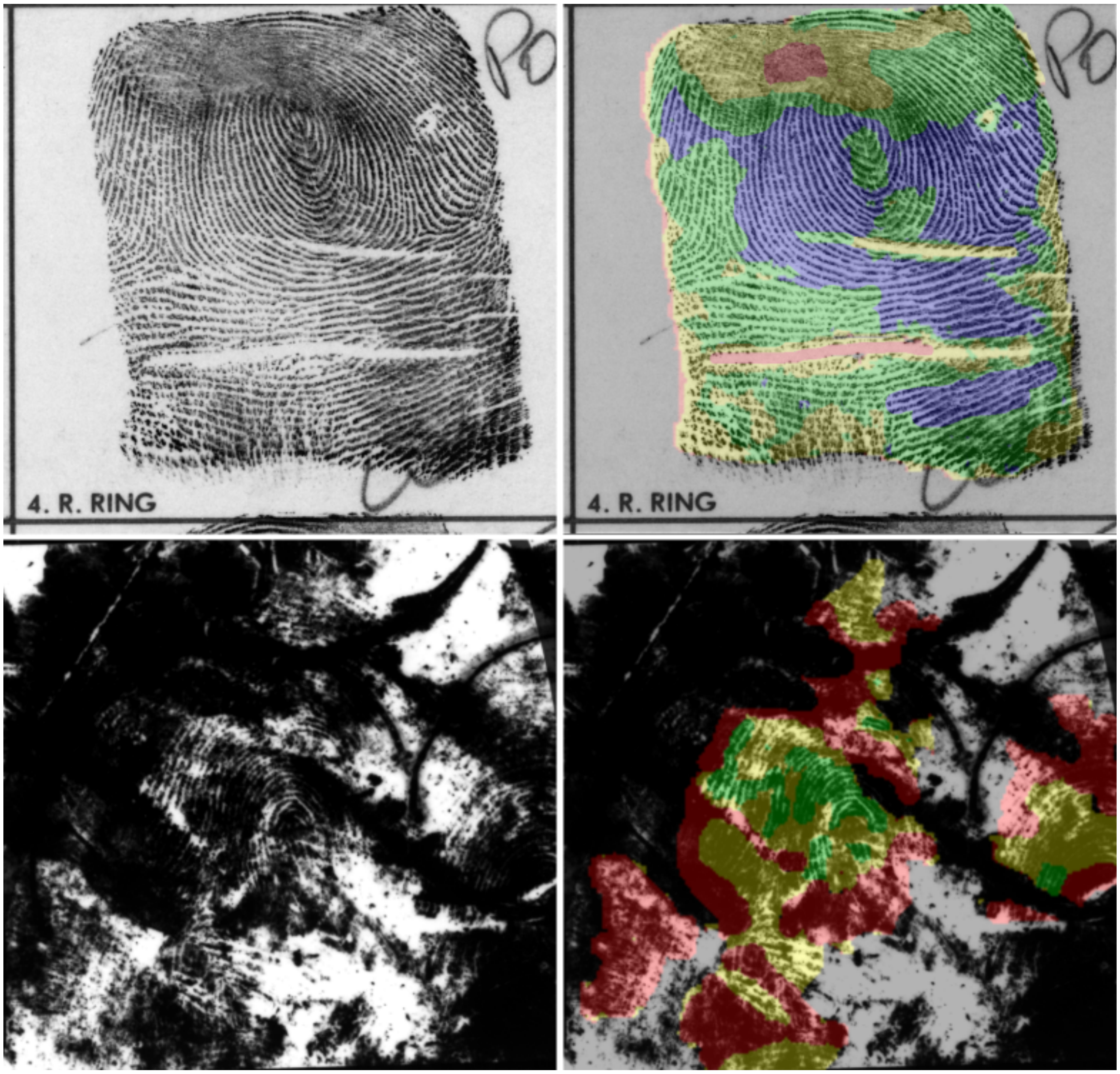


Figure 5: Examples of clarity maps: (top) for an inked exemplar; (bottom) for a latent image with multiple discontinuous areas.[102]

The value of an image depends upon the size and continuity of the clarity map areas. Clarity maps are particularly important for images with extensive discontinuities: the small separations of debatable ridge flow (red) are key, because those define the problem areas that can cast doubt on comparison decisions. The latent print in Figure 5 is complex, containing multiple impressions, slippage, and double taps; the associated clarity map indicates an examiner's assessment of the areas that contain continuous ridge flow, and literally depicting the areas that should be treated with caution in performing comparisons (often referred to as "red flags").

Analysis of clarity maps can be rapid and effective: when viewed at thumbnail size, dozens of images can be reviewed at a glance, with anomalies becoming immediately apparent, as shown in Figure 6. Much of the assessment of the overall utility of the image can be reduced to analyses of these clarity maps: ideal images have large blue or green areas, whereas poor images have little green or blue, and notable gaps, discontinuities, holes, or concavities.

[102] *Figures from [AssessingLC]. Images originally from [NIST-SD27].*

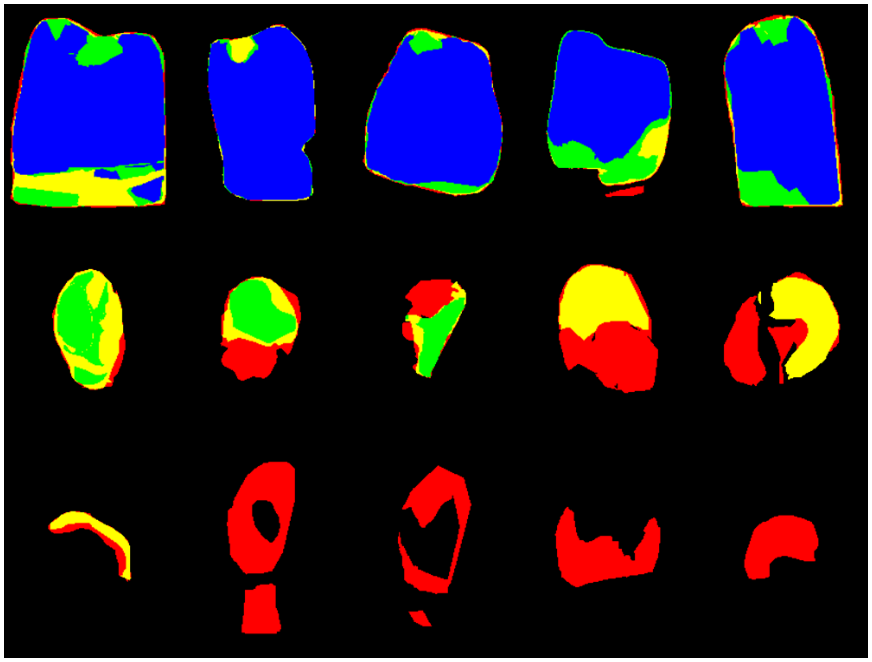


Figure 6: Clarity maps shown at thumbnail size; all images are the same scale. The top row includes exemplars rated "very easy" by examiners; the second row are latents rated "difficult" or "very difficult" by examiners; the third row are latents rated "of no value" by examiners.[103]

It is critical to note that the clarity maps are not contingent on minutiae or other features. A clarity map can be used in conjunction with marked features to indicate degrees of confidence in specific features. For example, minutiae in a yellow area are not definitive; minutiae in a green area are definitive but with little or no associated ridge edge detail; and minutiae and ridge edges in a blue area are definitive but with little or no associated pore detail (level-3 detail). Clarity maps can indicate distinctions between the definitive absence of features and the lack of discernible features: a green area without any marked minutiae indicates an open field of ridges (definitive absence of minutiae), whereas a yellow area without any marked minutiae indicates an ambiguous area that may contain undetected minutiae.

5.3.1 Uses for clarity maps

Clarity maps provide a reliable, commonly defined means for interchange of assessments of clarity and confidence in features made during the analysis or comparison stages of friction ridge examination. Clarity maps may be used in documentation, communication among examiners, resolution of conflicts between examiners, and as a means of rapid visual assessment of impressions.

Clarity maps may also be used as an aid in automated fingerprint matching. Clarity maps provide a means for examiners and automated systems to communicate confidence levels associated with feature annotation. Human-marked clarity maps included with minutiae in searches of an AFIS can be used by the AFIS to determine which minutiae are definitive, as well as to determine which unannotated areas are open fields of ridges: feature-by-feature confidence information provides the means for an AFIS to make exclusions based on contradictory features.

[103] *Figure from [AssessingLC]*

5.3.2 Related color-coding annotation methods

Other systems propose complementary methods for the annotation of latent casework using color coding to indicate clarity or confidence in features: GYRO[104] color-codes minutiae, Laird[105] color-codes both minutiae and ridge tracings, and PiAnoS[106] color-codes areas. LQSurvey, GYRO, and Laird were all published in *JFI* in 2011, and all were apparently in development for some time previously; the EFS approach was first presented publicly in 2008 and incorporated in ULW in 2009. All the methods are variations of a red/yellow/green stoplight, but differ in the details:

- Laird uses green, yellow, and red for Analysis markup with equivalent meanings to EFS: green means certain, yellow uncertain, and red is an area of no value. However, Laird uses blue to mark "scars, creases, scratches, and other physical distortion" (EFS uses blue to indicate clear level-3 information). For comparisons, Laird changes the meaning of red to indicate discrepancies.
- GYRO uses three levels of confidence for minutiae (green means high, yellow medium, and red low), as opposed to the two levels used in EFS and Laird. GYRO uses orange to indicate features that differ between analysis and comparison.
- Laird and GYRO mark minutiae and individual ridge paths, whereas EFS marks areas. (Laird marks no value areas in red).
- PiAnoS is most similar to EFS in that it color-codes the area (not the minutiae), and the definitions for the colors are similar to EFS: green (high quality) if Level 1,2, and 3 are distinct; orange (medium quality) if Level 1 is distinct, most of the Level 2 details are distinct, and there are minimal distinct Level 3 details; red (low quality) if Level 1 may not be distinct, most of the Level 2 details are indistinct, there are no distinct Level 3 details; areas without any ridges are not marked.
- The area outside the image does not have a color or is black in all the methods.

One implication of color-coded markup is that for the first time, color-blind latent print examiners are at a disadvantage. The frequency of color blindness varies by ethnicity and is much more prevalent among males: about 7 percent of the US male population are red-green colorblind.

5.4 Latent Quality Assessment Software (LQAS)

The data and findings from the *Latent Quality Study* led to the development of prototype Latent Quality Assessment Software (LQAS), which was designed as a proof-of-concept interactive tool for the evaluation of clarity. LQAS included functionality for the manual definition of clarity maps using a painting interface, and automated definition of clarity maps (Section 5.5). LQAS incorporated a variety of functions to process clarity maps, resulting in aggregate clarity measures and calculation of an overall clarity (OC) metric (Section 5.6.2). LQAS also included functionality for the marking of corresponding points, providing a method for overlapping print areas, and calculation of clarity metrics in the overlapping areas (Section 5.6.3). Most of the LQAS functionality has been incorporated into ULW (versions 6.5 and later; see LQMetric, Section 5.7). LQAS itself was never released publicly.

5.5 Variability in examiners' assessments of local clarity

In the *Latent Quality Survey* and in the subsequent *Sufficiency for Value* and *Sufficiency for Individualization* studies, we observed notable variation among examiners. The inter-examiner variation in markup seen in Figure 7 was typical: although all of the examiners marked the same

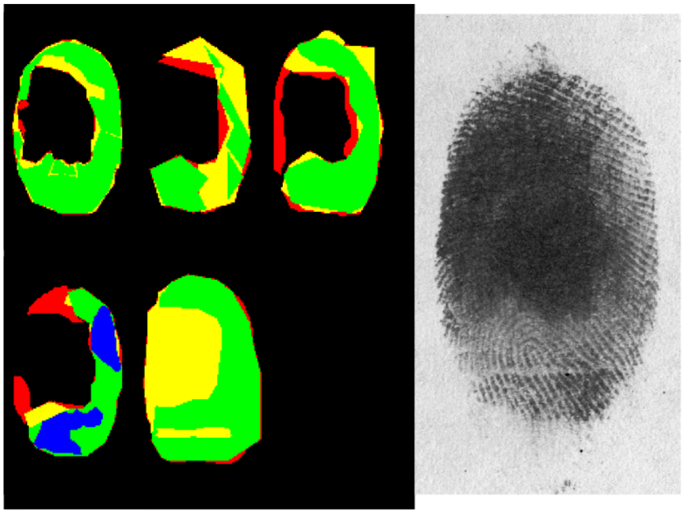
[104] *[Langenburg11]*

[105] *[Laird11]*

[106] *[PiAnoS15]*

basic areas, they frequently assigned different degrees of confidence to the features found in the area.

The purpose of documentation is not to avoid subjectivity, but almost the opposite: documentation illustrates what that specific examiner saw in an impression, and therefore comparison of documentation among examiners allows us to see the extent of subjectivity. The differences in markup among examiners correlated to the examiners' assessment of the value and difficulty of the images: examiners who assessed a given impression as easy were more likely to indicate higher confidence or larger areas of confidence than examiners who assessed the impression as difficult.



🟦 Confidence in both Level-2 and -3 detail

🟩 Confidence in Level-2 detail; low or none in Level-3

🟨 Confidence in Level-1 detail; low or none in Level-2

🟥 Low confidence in Level-1 detail

⬛ No confidence at any level

Figure 7: Clarity maps indicating examiner confidence from <LQSurvey>, showing typical variation among five different examiners. The color-coding definitions here predate those used in EFS.[107]

5.6 Automatic assessment of quality

Automatic assessment of the quality of an impression can be performed at the local clarity level (resulting in a clarity map) or at the overall level (resulting in a value assessing the utility of the impression). This provides an overview of the approach used in LQAS and discussed in [AssessingLC]; a subsequent publication will discuss the topic with respect to the *Latent Quality Metric* project, which uses a more sophisticated approach and builds on subsequent external research.[108] [ISO-29794-4] provides a detailed discussion of fingerprint quality metrics.

5.6.1 Automatic assessment of local clarity

The process of automatically generating clarity maps involves developing (or collecting) a variety of image processing algorithms that measure attributes associated with impression clarity, and using machine learning methods to select and combine the results from those algorithms based on a set

[107] *Figure from [LQSurvey]*

[108] *e.g. [Yoon12,Yoon13]*

of training data. For clarity maps, the best training datasets currently available are based on the median of the local clarity maps from multiple examiners.

Various low-level image processing algorithms assess attributes useful for discerning local clarity within the image. Feature extraction algorithms (such as HO39/MINDTCT[109] and FBI RFES[110]) produce a variety of intermediate representations that are useful in assessing low-level clarity. For the purpose of generating clarity maps, the intermediate representations must not be contingent on the presence of features, because clarity maps are used to assess whether the presence or absence of features can be determined with confidence. Low-level clarity algorithms are useful if individually or in combination with other algorithms they can serve to predict the target local clarity values, which in this case were the median values across examiners. In the simplest case, the result of a low-level clarity algorithm is monotonic with respect to the target, which means that it can serve as a predictor across its full range. Other such low-level metrics can be partial predictors: several metrics are effective at differentiating poor clarity areas but ineffective with respect to good and very good clarity areas.

Types of low-level algorithms we found useful as the basis for assessing clarity include

- Grayscale distribution algorithms — Some generic image processing algorithms are useful as low-level clarity algorithms, most notably those related to grayscale distribution (e.g. grayscale mean and standard deviation). Most of these measures are more effective for exemplars than for latents because of the relatively consistent contrast between ridges and background; for latents, these are generally most useful in the extremes (e.g. in identifying very low contrast, very dark, and very light images), but are less useful through the rest of their ranges.
- Ridge direction variation algorithms — Most fingerprint processing algorithms include one or methods to determine the direction of ridge flow at various sampling points through the image; these can be implemented using various techniques including comb filters, Gabor filters, or Fourier transforms (DFTs or FFTs). Once the directions are determined, anomalies can be detected by assessing the difference between the directions at neighboring sampling points. These approaches are very effective at identifying very poor or background areas. They are much less effective in differentiating within good quality areas for the interesting reason that these same algorithms are also used to identify cores and deltas, which are points of inflection within ridge flow, but should not be flagged as poor quality. These algorithms can be derived from HO39/MINDTCT and in RFES. This is used in NIST NFIQ,[111] and is the only local metric used in the DOD's FIQM.[112]
- Frequency and power spectrum measures — Because of the relative regularity of ridges, fingerprints naturally lend themselves to frequency approaches. Feature detectors use frequency analysis in identifying ridge flow direction and strength, and are therefore implicitly used in ridge flow strength or direction measures, or binarized images. Intermediate representations of frequency magnitude can be derived from the Fourier transforms used in MINDTCT and RFES. Frequency or power spectrum approaches have been proposed in multiple approaches to assessing fingerprint quality.[113] Frequency and power spectrum measures can be very effective at separating good ridge detail from clear backgrounds. However, for the areas of

[109] *[NIST-MINDTCT] The NIST MINDTCT minutia extractor was derived from the UK Home Office HO39 minutia extractor, which was developed in 1989.*

[110] *The FBI Remote Fingerprint Editing Software (RFES) was latent fingerprint client software c. 2000-2007, replaced by ULW. The RFES source code was made available at the time with the software.*

[111] *[Tabassi04]*

[112] *FIQM: Fingerprint Image Quality Measurement [Yen06]*

[113] *E.g. [Fierrez-Aguilar05, Nill07]*

latent fingerprints with unclear ridge detail and noisy or complex backgrounds, they are of value but less effective.

- MINDTCT quality — In 1998 I combined intermediate representations of algorithms already present in the HO39 minutiae extractor then used by ULW into a 4-level scale: curvature (ridge direction variation), low contrast (grayscale range), grayscale reliability (grayscale mean and standard deviation), and ridge flow strength (Fourier frequency magnitude derivative). This quality algorithm was retained when HO39 was recoded as NIST MINDTCT, and is used by the NIST Fingerprint Image Quality metric (NFIQ) to define the quality of regions and minutiae in exemplar fingerprints.[114]
- Difference of binarizations — Minutiae extraction algorithms often use binary images to represent the detected ridge flow. The binarization process is not trivial: binarization involves a complex process of filtering and processing the image based on ridge direction, ridge strength, and grayscale factors, combined with logic to address discontinuities. By taking the difference of two different binarization algorithms (HO39/MINDTCT and RFES), the result shows those areas where the different algorithms agreed on the locations of ridges. The approach works for these two algorithms because they are sufficiently different in approach: the algorithms, implementation details, parameters, and decision points are substantively different. In well-defined areas, the resulting ridges are nearly identical, with increasing disparity as quality drops. This approach differs from the other raw local quality metrics in that those methods attempt to predict whether ridge detail can be detected; difference of binarizations is based on the actual effect, on whether two methods of identifying ridge structures have the same (or similar) results.

A cautionary note on how such metrics can be misused: edge distance, a simple metric of measuring the distance from the edge of the image, can be used to eliminate boundary conditions at the edges of images. Interestingly (or amusingly), initial machine learning results showed that distance from the edge was an effective predictor of quality throughout its range, which is the obvious effect of training on a set of images in which the region of interest is centered in the image – but obviously not a useful metric! This is why machine learning approaches need careful review: a black box machine learning approach that does not allow visibility into how the inputs are used can base its functionality on accidental characteristics of the data such as this. Edge distance was found to be effective – and appropriate – at the extreme edges of the image.

Automated assessment of local clarity has a notable issue with many latent fingerprint images in that the backgrounds of latent prints frequently contain multiple impressions or complex patterns that the current metrics cannot readily differentiate from the impression of interest. Such issues are much less common with exemplars, although written and printed text can result in similar issues for exemplars scanned from paper fingerprint cards. Ideally, human examiners would explicitly mark the regions of interest, especially for prints containing multiple impressions or with complex backgrounds. When this is not possible, it is necessary to estimate the region of interest: the region of interest is assumed to be the largest contiguous area with definitive ridge flow (yellow or better).

The automated local clarity algorithm in LQAS used a combination of difference of binarizations, ridge direction variation, grayscale distribution, and ridge flow strength (Fourier frequency magnitude derivative). Recursive partitioning was used as the machine learning approach.

5.6.2 Aggregating local clarity into an overall clarity metric[115]

Clarity maps can be aggregated into a single overall clarity metric to represent the overall clarity for an image. Such an overall clarity metric could be appropriate for use in cases in which fidelity is to

[114] *[Tabassi04]*

[115] *Derived from [AssessingLC]*

be assessed independently of utility, such as in evaluating latent processing methods. For our purposes, we wanted to separate overall clarity from quantity measures (e.g. minutia count) to determine how they could be most effectively combined.

Assessing the overall clarity of an image requires the aggregation of local clarity data over the image. While the size of the area for each local clarity value is correlated to overall assessments of an image,[116] both visual assessments and machine learning analysis showed that area alone was ineffective as an overall assessment of a fingerprint image. Aggregation methods need to address not just size, but also the consistency of the data, accounting for factors such as gaps, discontinuities, or concavities. Such methods of aggregating local clarity values are necessary not just for overall assessments of clarity, but also for automated region of interest estimation, which we based on the largest contiguous area of ridge flow.

The goal in deriving an overall clarity metric was to develop a repeatable monotonic value that corresponded to human examiner assessments of the value and difficulty of an image, given a clarity map created by a human examiner or by software. We determined that in order for an overall clarity metric to be useful to latent print examiners, it needed to employ a single 0-100 scale for both latents and exemplars representing the value and difficulty of the impression.

Ideally, impressions would have large areas of high clarity that are generally convex and without gaps, so that the area in between any two minutiae could readily be interpreted. In [AssessingLC] we describe the derivation of an overall clarity metric based on the sizes of the areas within a clarity map, but in which locations are weighted more heavily if they are in large continuous areas, and away from gaps and edges. For example, a clarity map with a single large elliptical area of green would have a much higher overall clarity value than a clarity map with the same total amount of green in discontinuous, irregular areas. The scale was primarily based on the size and consistency of the areas of definitive minutiae (green or better), and the highest clarity values were limited to impressions with large areas of both definitive minutiae and clear ridge edges (blue or better).

The Latent Quality Survey results provided a useful but imperfect basis for training an overall clarity algorithm. The value and difficulty assessments from the survey were assessments of the overall quality, not just overall clarity, and therefore affected by factors such as whether the image was a latent or exemplar print, or the number of minutiae visible in the image. Since these assessments did not directly correspond to our goals for an overall clarity metric, they could not be used in a standard machine learning process as training and test data. Instead, these assessments were used to define a heuristic algorithm as part of a feedback loop using analysis with recursive partitioning, analysis of the images, and development or enhancement of aggregation algorithms. We evaluated the effectiveness of our overall clarity metric by comparison with human examiner assessments of the value and difficulty of latent and exemplar fingerprints.

The result is a scale in which Overall Clarity generally ranges from 1 to 10 for “no value” latents, 5 to 20 for “value for exclusion only” latents, 10 to 50 for very difficult or difficult latents, and 40 to 80 for easy or very easy latents. When compared against the informal “good, bad, ugly” (GBU) scale used in the NIST SD-27 dataset,[117] the median Overall Clarity was 14 for “ugly” prints, 35 for “bad” prints, and 49 for “good” prints. While the Overall Clarity metric correlates to the examiners’ informal, subjective assessment of difficulty, the Overall Clarity metric is more repeatable and reproducible; it is precise, more amenable to analysis, and provides a standard means of communicating assessments of clarity.

[116] *[LQSurvey]*

[117] *[NIST-SD27]*

5.6.3 Overall quality metrics

There are two general issue to be considered in the development of a quality metric that I feel should be discussed here:

- A quality metric needs to distinguish between what I consider general quality (overall usefulness for a range of potential purposes) and utility for a specific purpose (e.g. AFIS matching). Most uses respond similarly to the same attributes, such as large contiguous high-clarity areas and a quantity of high-confidence features. However, a latent quality metric should consider instances in which purposes do not overlap. For example, a metric used to help examiners make value/no value decisions is focused on distinctions among the lowest-quality images, whereas a metric assessing the probability of AFIS matching will generally focus on distinctions within latents of value; combining and weighting these disparate purposes in a single metric is a non-trivial task.
- The likelihood that a comparison will result in an individualization or exclusion is certainly correlated to the quality of the latent being compared; however, any quality metric that is attempting to predict the effectiveness of comparison based on the quality of the latent alone will necessarily be imperfect. Comparison, whether performed by an examiner or by an AFIS, depends on not just the quality of the latent, but the quality of the exemplar, the extent of overlap, the corresponding quality between the two impressions (see Section 5.7), and the relative distortion between the two impressions (see Section 5.9).

A variety of previous work has been focused on the development of exemplar fingerprint quality metrics used to predict automated fingerprint identification system (AFIS) matcher scores (e.g. NIST NFIQ, DoD IQM; AFIS often include proprietary metrics). Such metrics have been shown to be operationally effective in establishing criteria for the acceptance/rejection of submitted images and in system modeling. Quality metrics for latent prints have been discussed in the literature,[118] but have not been widely implemented.

5.7 LQMetric[119]

The FBI's Latent Quality Metric (LQMetric) software automatically assesses the quality of latent fingerprint images. LQMetric is an estimate of the probability that an NGI image-only (LFIS) search would hit at rank 1, assuming the mate is in NGI: for example, an LQMetric value of 80 means that the latent is 80% likely to hit at rank 1 ***if*** the mate is present in the database. Note that the probability of a hit depends not only on latent quality, but is also determined by factors such as the quality of the exemplar(s), the extent of overlap between the latent and exemplar(s), and whether the subject is in the database — any of which could prevent a high-quality latent from hitting.

LQMetric is incorporated into the Universal Latent Workstation (Version 6.5 and later), for either interactive or commandline/script use.

The description of the LQMetric algorithm has not yet been publicly released.

LQMetric was developed to predict whether a latent would match on an automated system; this ability to match is similar but not always the same as how an examiner would assess the quality or value of a latent. LQMetric agrees more with examiner assessments of high quality than low quality: high LQMetric latents are almost always of value and usually "good", but latents with low LQMetric values may or may not be considered to be of value by examiners. Some of the low correlation between LQMetric and examiner assessments for poor quality latents is due to disagreements among examiners: examiners were much more likely to disagree with each other for the latents where LQMetric was less than 50. We also compared LQMetric against informal assessments of

[118] *e.g. [Yoon12, Yoon13]*

[119] *[Hicklin15]*

"good", "bad", and "ugly", and found that most latents that were informally assessed by examiners as "good" resulted in LQMetric values from 65 to 90, "bad" from 45 to 65, and "ugly" from 20 to 45.

This evaluation was conducted to determine how well LQMetric agrees with latent print examiners' assessments of the quality of latent prints for non-AFIS casework. Latent print examiners rated the quality of latent prints, and were given side-by-side pairs of latent prints to indicate which was of better quality. We compared these human examiners' assessments to LQMetric values.

There is general agreement between examiners' ratings of latent images and LQMetric scores. The agreement increases as the LQMetric score increases, especially when the LQMetric score is 50 or larger. There is general agreement between examiners and LQMetric as to which image in a pair is better. The agreement increases as the difference in LQMetric scores between the two images in a pair increases. When LQMetric indicates that a latent is high quality, examiners overwhelmingly agree. However, examiners often consider latents to be of value when LQMetric indicates that a latent is low quality. We feel that LQMetric could be useful as an objective, automated measure of latent quality for non-AFIS casework.

5.8 Corresponding clarity[120]

Local and overall clarity measures for a single impression do not directly address how clarity affects the comparison of two impressions. A clear area in one impression is irrelevant if there is no corresponding area available in the other impression, or if the clarity of a corresponding area is substantially lower. When comparing corresponding areas in two impressions, the area of lower clarity limits the comparison. For example, in Figure 8, there are large areas in each image that cannot be used for comparison because there is no corresponding area available; a comparison cannot take full advantage of the incipient ridges in the blue area in the center image, because of the lower clarity of the corresponding area in the left image. The area and clarity of corresponding regions can be depicted in a corresponding clarity map that combines the clarity maps for each of the individual impressions: in Figure 8, the corresponding clarity map is the result of transforming and superimposing the clarity maps for the two impressions, and selecting the lower clarity value at each sampling point.

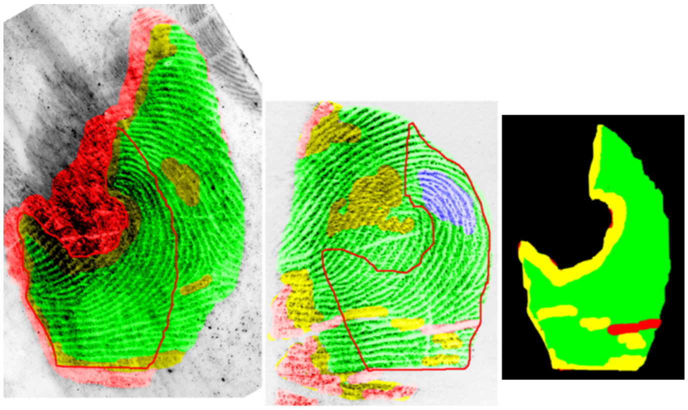


Figure 8: Example of the effect of clarity in a comparison. The outlines indicate the corresponding regions of interest in the two fingerprints. The corresponding clarity map on the right combines the clarity maps for the two fingerprints at each sampling point.

The process of calculating a corresponding clarity map requires a transformation of the two constituent clarity maps so that they are in the same Cartesian space. The clarity map for one

[120] Derived from [AssessingLC]

impression (generally the latent print) must be transformed so that it can overlay the clarity map for the other impression. This process requires the marking of registration points for the two images. The transformation of the clarity maps so that they can be superimposed can be accomplished through various warping methods, as discussed in Section 5.9. Affine or projective transformations can be used for impressions that have minimal relative distortion; greater levels of distortion require more sophisticated approaches, such as thin-plate spline transformations. After the transformation, the clarity maps can be superimposed, and a corresponding clarity map created by taking the lesser value from each sampling point of the two clarity maps. Once a corresponding clarity map is defined, it can be processed as any other clarity map, resulting in corresponding overall clarity or quality metrics.

Corresponding clarity maps may be of operational interest for documenting comparisons; corresponding clarity metrics may be appropriate for use in quality assurance processes, such as in flagging complex comparisons that require additional review.

5.9 Quantifying distortion[121]

When fingerprints are deposited, variations in pressure in conjunction with the inherent elasticity of friction ridge skin often cause linear and nonlinear distortions in the resulting ridge and valley structure. Distortions in the fingerprint can be caused by the substrate (e.g. curved or flexible objects), matrix (e.g. viscous substances), development medium (e.g. powder buildup), and the pressure and direction of deposition.[122] Deposition pressure (downward pressure) can change the width of ridges and valleys, as well as the appearance of minutiae, ridge edge details, and pores.[123] Shearing (lateral pressure in a single direction) will cause elongation or compression of the print, resulting in linear differences in the location of minutiae or other features. The most complex non-linear distortions are caused by torque (twisting pressure), which can cause apparent changes in the overall pattern as well as substantial differences in the relative locations of features.[124] Latent fingerprints can be highly distorted due to the combination of some or all of these factors. Exemplars also can be distorted, particularly in the upper corners of rolled prints.

Anatomical constraints affect how the finger pad reacts to pressure. Areas in the center of the finger can distort more than the less flexible tips or edges. The ability of skin to stretch or compress is affected by the direction of ridge flow: skin is less flexible in the direction of ridge flow than perpendicular to flow. Therefore cores and deltas respond differently to pressure than open fields of parallel ridges, and different pattern types react differently to pressure.[125] The effects of pressure and finger deformation may or may not be apparent in the analysis of a single impression. However, even when individual prints do not appear to be distorted, the relative distortion between two prints can have a serious impact on comparison.

The distortion model described in the *Distortion Study* models the linear and nonlinear relative distortion between pairs of latent and exemplar prints, based on correspondences annotated using EFS. We first globally align the correspondences through an affine (linear) transform, and then the thin-plate spline (TPS) algorithm is applied to model the non-linear deformation between the minutia correspondences. An example of the effects of linear and nonlinear transformations is shown in Figure 9.

[121] *Derived from [Distortion]*

[122] *[Ashbaugh99]*

[123] *[Richmond04]*

[124] *[Maceo09]*

[125] *[Maceo09]*

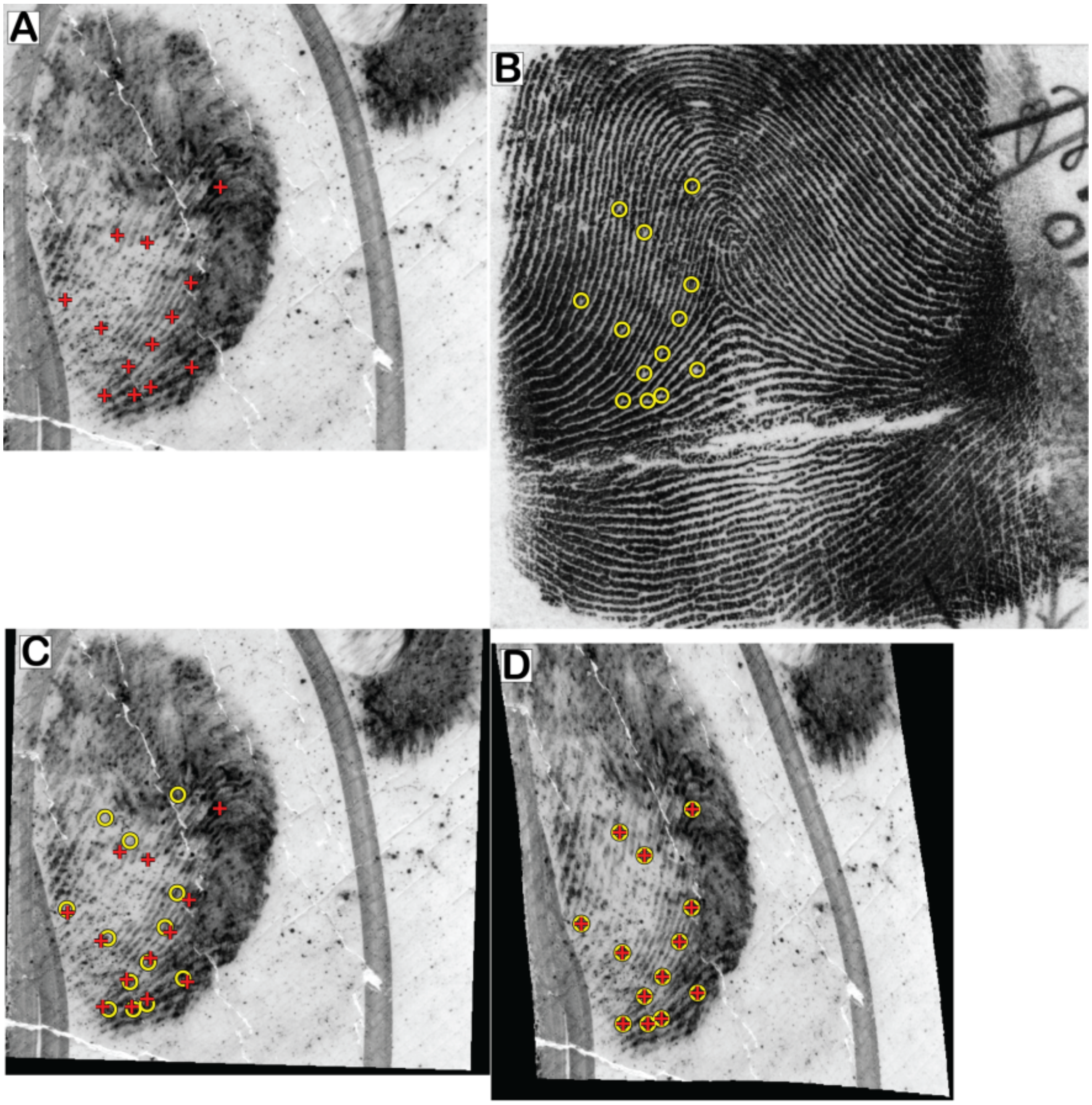


Figure 9: Effect of linear and non-linear distortion models. (A) Latent overlaid with minutiae (crosses). (B) Exemplar overlaid with minutiae (circles). (C) Latent after linear (affine) transformation; alignment is poor for some minutiae. (D) Latent after nonlinear (TPS) transformation; all latent minutiae (crosses) overlay precisely the exemplar minutiae (circles). Images A-D are aligned vertically and horizontally with respect to the top right minutia.

We introduce grid warps and heat maps for visualizing the relative deformation between two impressions, as shown in Figure 10. The grid warp is a straightforward method of visualizing relative deformation. The transformation function is used to warp a 2D grid of vertical and horizontal lines; the resulting warped grid provides for visualization of local deformation within the impression. The heat map is another method for visualizing relative deformation between impressions, which makes use of the residual distance metric. Instead of computing the distance between minutiae, we compute the Euclidean distance between the original and transformed features, and the magnitude of the distance is used as a visual cue for relative deformation. Higher intensity values in the map indicate higher levels of deformation while the opposite is true for lower values. Figure 10 provides an illustration of the grid warps and heat maps.

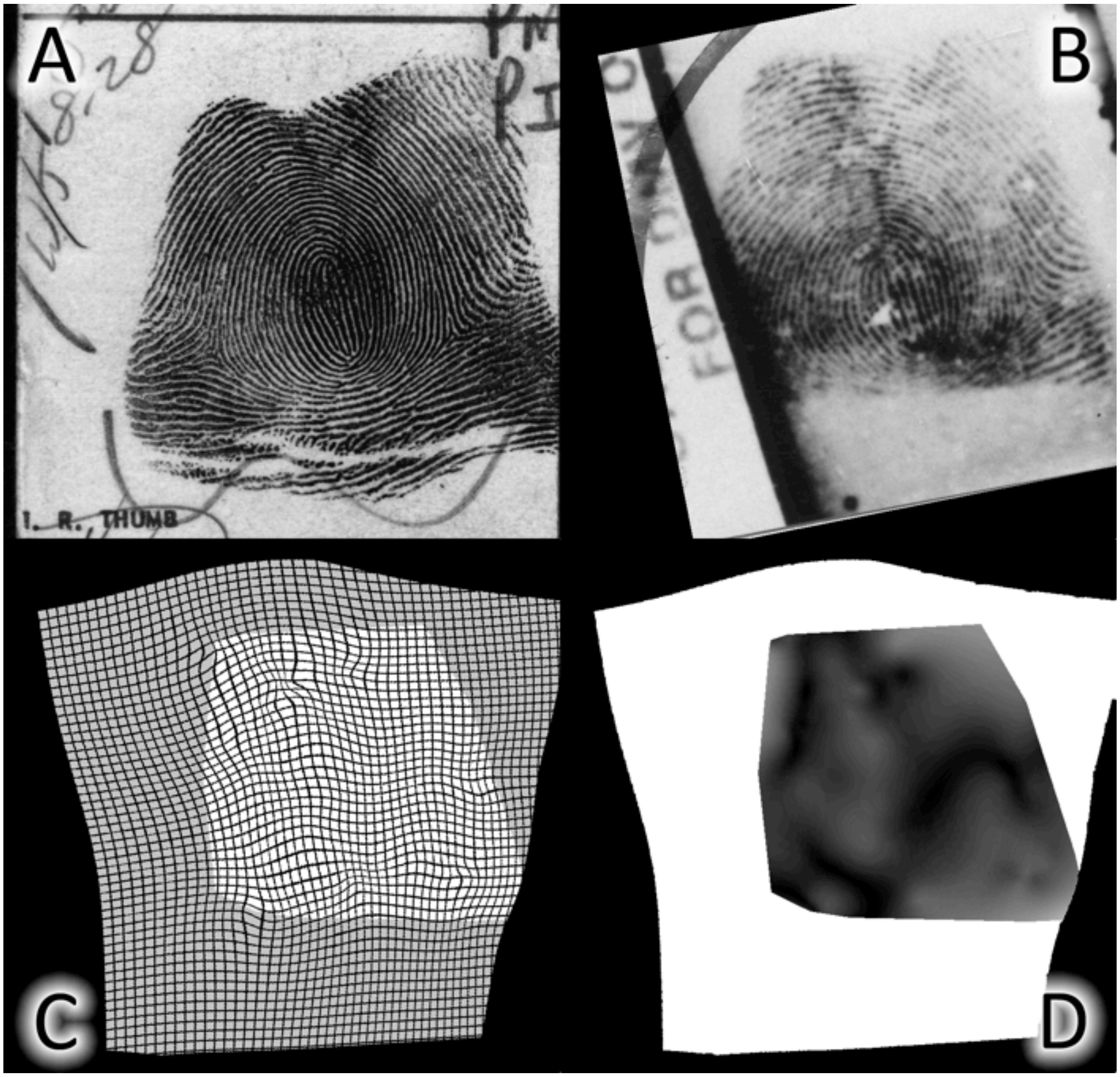
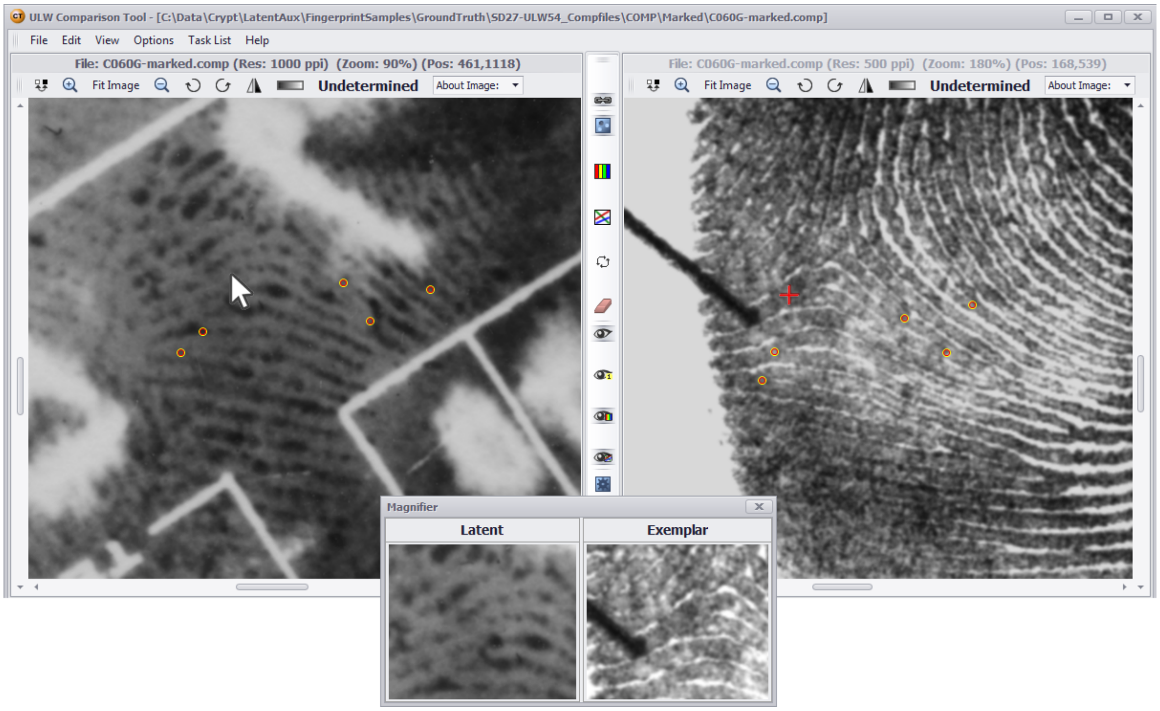


Figure 10: Methods of visualizing relative deformation: (A) exemplar; (B) latent (rotated but not otherwise transformed); (C) grid warp of the deformation (with convex hull of the minutiae highlighted); (D) heat map of the deformation.

We describe two metrics that may be used to characterize the relative deformation between a set of impressions: a Euclidean metric that captures the residual distance between corresponding minutiae points, and the bending energy metric which is provided through the TPS model. Residual distance accounts for both linear and nonlinear distortion, whereas bending energy accounts only for non-linear distortion. In casework, such metrics could be used to flag comparisons that are especially distorted, which then could be required to have additional review or other quality assurance procedures. When examiners mark corresponding points during comparison, the distortion metrics discussed here could be used as integrity checks during comparison, by identifying potentially erroneous corresponding points for which very high values of either residual distance or bending energy indicate amounts of distortion that would be improbable or impossible in correctly-annotated minutia correspondences. For evaluation, we deliberately created latent-exemplar pairs with erroneous correspondences: errors were introduced by selecting a correspondence at random, and then swapping the latent points from the nearest neighboring correspondence; erroneous markups were created in this way with two, four, six, and eight incorrect correspondences.[126] Both residual distance and bending energy were effective at separating the incorrect from correct correspondences, but bending energy was notably more effective.

During fingerprint comparison, one usability issue often encountered during fingerprint comparison is that when the examiner's eyes are moving between two images, it is easy to lose track of the specific locations being compared. In the ULW Comparison Tool, the warping technique described here is used in the implementation of a "ghost cursor," which serves as a reference point when making comparisons. Using EFS corresponding points marked by the examiner, the software defines a distortion model to map projected correspondences between locations in the two images; a minimum of three points is required. Wherever the user places the cursor, the software will display a ghost cursor at the estimated corresponding location, as shown in Figure 11. Since the ghost cursor is displayed in real time, the examiner can use it while moving the cursor to follow ridges and count ridges. In comparing prints, the distance between the areas being compared can be problematic: it is much easier to perform a detailed comparison when the areas being compared are immediately next to each other. ULW addresses this problem with "magnifier" functionality (Figure 11): when the user chooses to display the magnifier, the areas immediately around the cursor and ghost cursor are displayed side by side, with the latent rotated to the local relative orientation. The magnifier is not static, but tracks cursor movement about the image, allowing detailed comparison when following the sequence of ridges. The ghost cursor works well in areas near corresponding points, but becomes less effective as the cursor moves farther away from corresponding features. Feedback from examiners has indicated that the ghost cursor and magnifier have been found to be useful as optional tools to assist in performing comparisons; they are easily hidden when not desired.

[126] *One swapped pair = two incorrect correspondences.*



ULW Comparison Tool - [C:\Data\Crypt\LatentAux\FingerprintSamples\GroundTruth\SD27-ULW54_Compfiles\COMP\Marked\C060G-marked.comp]

File: C060G-marked.comp (Res: 1000 ppi) (Zoom: 90%) (Pos: 461,1118)

File: C060G-marked.comp (Res: 500 ppi) (Zoom: 180%) (Pos: 168,539)

Figure 11: Example of ghost cursor in ULW Comparison Tool. The cursor (arrow in left image) and corresponding points (circles), are used by the software to display a ghost cursor (red cross in right image) at the estimated corresponding location. The magnifier is shown at the bottom.

Chapter 6 Evaluating latent print examiners

Transparency requires having an accurate understanding of the latent print examination process, and measuring what examiners do. Here I discuss several studies we conducted to provide insight into the examination process by evaluating the accuracy, reproducibility, and repeatability of examiners' determinations, and by evaluating the bases for those determinations.

6.1 Evaluating Examiners' Determinations: the Black Box and Black Box Repeatability studies

Our *Black Box Study* was a large-scale study of the accuracy and reproducibility of latent print examiners' Analysis and Comparison determinations. The *Black Box Repeatability Study* retested examiners to evaluate the repeatability of their determinations. The *Black Box Study* in particular has been influential: it was introduced in court (Minnesota v Terrell Dixon, 2011) the day after publication, and has been frequently ever since.

The key value of these studies was to provide a general understanding of examiners' determination rates, and the implications of these determinations. These studies need to be seen as part of a larger research effort to understand the accuracy of examiner conclusions, the level of consensus among examiners on decisions, and how the quantity and quality of image features relate to these outcomes. We designed the study to enable additional exploratory analyses and gain insight in support of the larger research effort.

These studies were conducted to provide assessments of a variety of measures, to provide data and lessons learned so that later targeted studies can be conducted. There is substantial variability in the attributes of latent prints, in the training and capabilities of latent print examiners, in the types of casework received by agencies, and the procedures used among agencies.[127] Average measures of performance across this heterogeneous population are of limited value[128] — but do provide insight necessary to understand the problem and scope future work. Furthermore, there are currently no means by which all latent print examiners in the US could be enumerated or used as the basis for sampling: a representative sample of latent print examiners or casework is impracticable.

To reduce the problem of heterogeneity, we limited our scope to a study of performance under a single, operationally common scenario that would yield relevant results. This study evaluated examiners at the key decision points during analysis and evaluation. Operational latent print examination processes may include additional steps, such as examination of original evidence or paper fingerprint cards (as opposed to making conclusions on electronic images), review of multiple sets of exemplars from a subject when available, consultation with other examiners, revisiting difficult comparisons, verification by another examiner, and quality assurance review. These steps are implemented to reduce the possibility of error, and therefore the error rates for conclusions reported by agencies should be lower than the individual examiners' rates.

It would be highly desirable for studies to be conducted in which participants were not aware that they were being tested (eliminating the "Hawthorne effect"). However, conducting a study in which participants do not know they are being tested is much more complex than some have suggested.[129]

[127] *The Black Box study's survey of participants (included in the report's supplemental material) provides insight into latent print examiners' operating procedures across the community.*

[128] *[Budowle09]*

[129] *e.g. [Haber09, Haber14a]*

The practicality of such an approach even within a single organization would depend on the type of casework. Agencies that conduct fully electronic casework (i.e. in which examiners only see electronic images, not physical evidence) could allow insertion of test data into actual casework, and I believe that it would be practical to conduct Black Box studies specific to those agencies. However, this may be complex to the point of infeasibility for agencies in which most examinations involve physical evidence, especially when chain-of-custody issues are considered. Combining results among multiple agencies with heterogeneous procedures and types of casework would be problematic. (I suggest possible alternatives in Section 8.1.7.)

6.1.1 The Black Box Study[130]

In this study, 169 latent print examiners each compared approximately 100 pairs of latent and exemplar fingerprints from a pool of 744 pairs. The fingerprints were selected to include a range of attributes and quality encountered in forensic casework, and to be comparable to searches of an automated fingerprint identification system (AFIS) containing more than 58 million subjects. This study evaluated examiners on key determination points in the fingerprint examination process: the point in Analysis when an examiner decides whether a latent is value for individualization (VID), of value for exclusion only (VEO), or of no value (NV); and the point in Comparison/Evaluation when an examiner compares a latent and an exemplar and makes a determination of individualization, exclusion, or inconclusive.

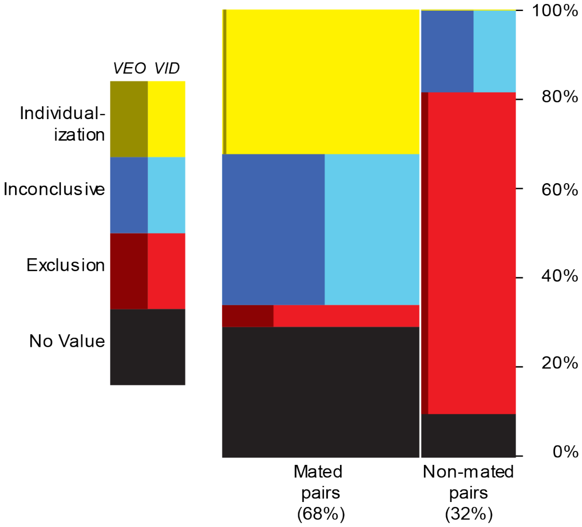


Figure 12: Distribution of determinations in the Black Box study.

Figure 12 shows the distribution of the 17,121 determinations in the Black Box Study. Comparison determinations were limited to VID and VEO latents; 23% of all determinations resulted in no value determinations (no comparison was performed).

Six erroneous individualizations occurred (false positive rate = 0.1%). Two of the false positive errors involved a single latent, but with exemplars from different subjects; two of the false positives were made by a single examiner. None of the erroneous individualizations was reproduced by any other examiners, indicating that verification (if fully independent of the original examination, e.g. blind) would be expected to have detected the errors.

Erroneous exclusions were much more prevalent than erroneous individualizations (false negative rate = 7.5%). Eighty-five percent of examiners made at least one false negative error, despite the fact that 65% of participants said that they were unaware of ever having made an erroneous

[130] *Derived from [BB]*

exclusion after training. False negatives were distributed across half of the image pairs that were compared. Verification of exclusions (especially blind verification) is not standard practice in many organizations, in part due to the large number of exclusions encountered in casework. We estimate (based on a statistical procedure detailed in the paper) that ***if*** blind verification were routinely conducted on exclusions, the false negative rate would drop to 0.85%.

6.1.1a Posterior probabilities

False positive and false negative rates are important accuracy measures, but assume *a priori* knowledge of true mating relationships, which of course are not known in forensic casework. In practice, the recipient of a determination generally wants to know the probability that it is correct: it would be desirable to know what proportion of individualization determinations are correct (Positive predictive value, PPV),[131] and what proportion of exclusion determinations are correct (Negative predictive value, NPV). The problem with PPV and NPV is that they are functions of the proportions of mated and nonmated data. The proportion of mated pairs in casework is very difficult to estimate, as it varies substantially among organizations, by case type, and by how candidates are selected — and because in operational casework it is not possible to know definitively whether an image pair is mated or not (other than by examiner determinations, which are what we are trying to evaluate). Mated comparisons are far more prevalent in cases where the candidates are suspects determined by non-fingerprint means than in cases where candidates were selected by an AFIS. Figure 13 depicts the effect of the proportion of mated data on PPV and NPV. As the proportion of mated pairs decreases, PPV decreases and NPV increases: note that if the test mix changed to 1% mates and 99% nonmates, only 80% of individualization determinations would be expected to be correct — based on this model, which only provides a rough estimate.

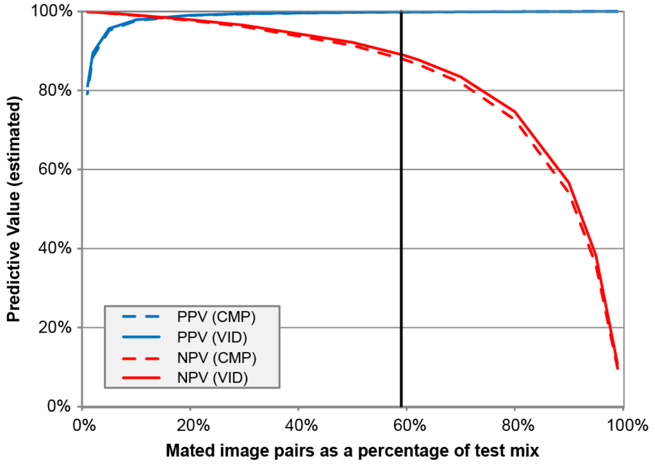


Figure 13: PPV and NPV as a function of the proportion of mates in the test mix. The observed predictive values (PPV=99.8% and NPV=88.9% for VID comparisons) correspond to the actual test mix (indicated) where 59% of VID comparisons were mated pairs; all other values are estimated.

6.1.1b Reproducibility of determinations

Each image pair was examined by an average of 23 participants. Their determinations can be regarded as votes in a decision space (Figure 14). Examiners frequently differed on determinations: of mated pair determinations, 10% were unanimous true positives, 38% unanimous inconclusives; of non-mated pair determinations, 25% were unanimous true negatives, 9% were unanimous

[131] *The same concept is sometimes described in reverse as the False Discovery Rate (1-PPV), which is the percentage of individualization decisions that are false positives. [Kohler08]*

inconclusives. The prevalence of false negatives is evident in the vertical spread of mated pairs; the few false positives are evident in the limited horizontal spread of the non-mated pairs. The points along the diagonal represent pairs on which all examiners reached conclusions (as opposed to inconclusive or no value): note that for some image pairs, all examiners reached a conclusion, but up to 30% made erroneous exclusions.

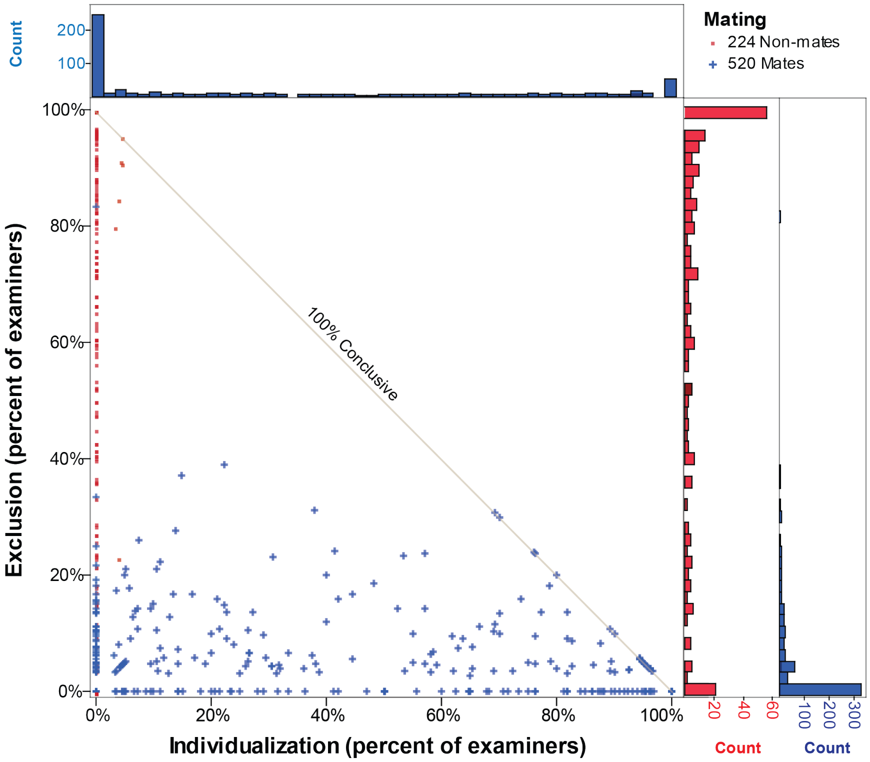


Figure 14: Determination rates on each image pair. Percentage of examiners making an individualization determination (x-axis) vs. exclusion determination (y-axis) on each image pair; mean 23 presentations per pair. Image pairs in which the latent is VEO or NV are treated as inconclusive.

The charts in Figure 15 show the percentage of examiners reaching consensus (y-axis) on each image/image pair (x-axis) for three types of determinations. Areas of unanimous (100%), decile (10%, 90%), and quartile (25%, 75%) consensus are marked. For example, in (A), at a 90% level of consensus (y-axes), examiners agreed that 40% of the latents were VID (interval from 60% to 100% indicated by a horizontal line in upper right).

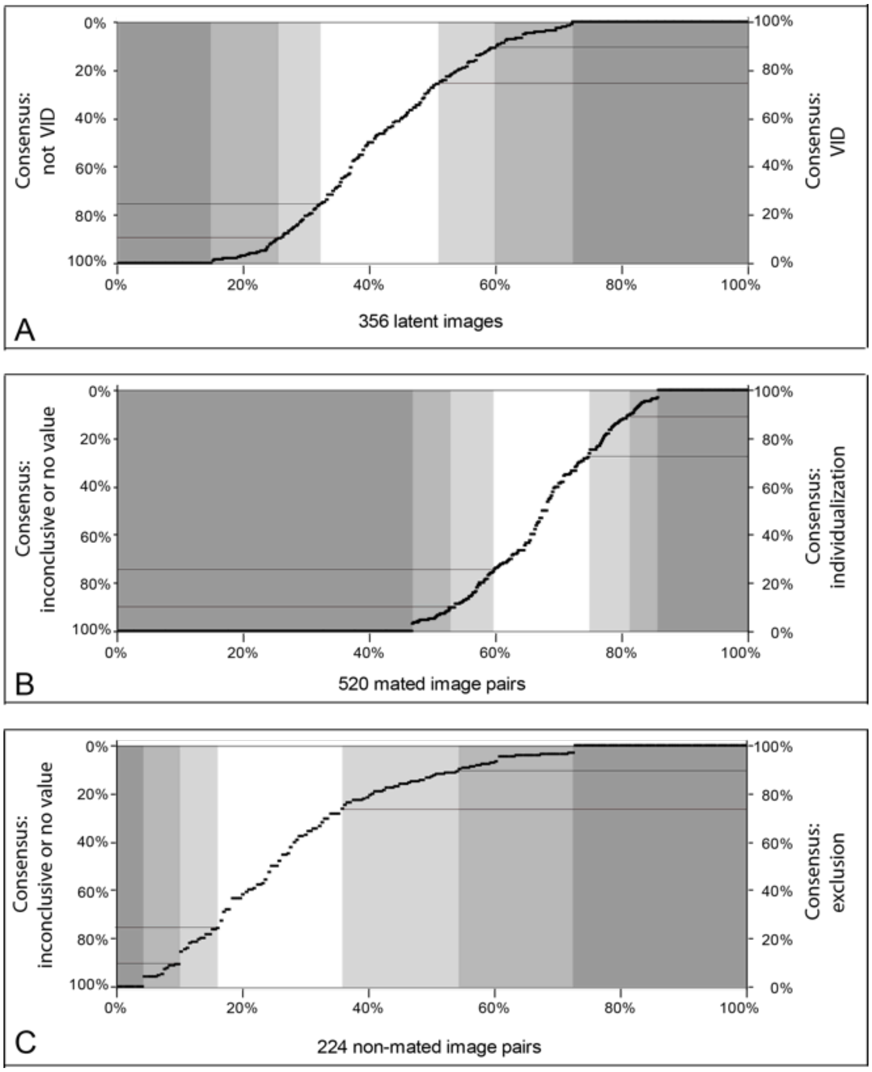


Figure 15: Examiner consensus on determinations: (A) value for individualization, (B) individualization of mated pairs (false negatives are omitted), (C) exclusion of non-mated pairs (false positives are omitted).

It was not unusual for one examiner to render an inconclusive determination while another made an individualization determination on the same comparison; this is known as a "missed ID" in some organizations. Among all determinations based on mated pairs, 23.0% resulted in determinations other than individualization even though at least one other examiner made a true positive on the same image pair; 4.8% were not individualization determinations even though the majority of other examiners made true positives.

Missed IDs have operational implications in that some potential individualizations are not being made, and contradictory decisions are to be expected. Such disagreements come to the fore in conflict resolution (when the initial examiner and verifier disagree) — but for laboratories that only verify individualizations, any missed IDs made by the initial examiner will not be detected.

6.1.1c Examiner skill

The skill of latent print examiners is multidimensional and is not limited to error rates, but also includes the rates of successful determinations, such as conclusion rate, the percentage of individualization or exclusion conclusions as opposed to no value or inconclusive determinations.

These other aspects of examiner skill are important not just to the examiner's organization, but to the criminal justice system as well: an examiner who is frequently inconclusive is ineffective and thereby fails to serve justice. We found that examiners differed substantially along these dimensions of skill: examiners' conclusion rates varied from 15% to 64% (mean 37%) on mated pairs, and from 7% to 96% (mean 71%) on non-mated pairs. The observed range in conclusion rates may be explained by a higher level of skill (ability to reach more conclusions at the same level of accuracy), or it may imply a higher risk tolerance (more conclusions reached at the expense of making more errors). Tests could be designed to measure examiner skill along the multiple dimensions discussed here. Such tests could be valuable not just as traditional proficiency tests with pass/fail thresholds, but as a means for examiners or their organizations to understand skills for specific training, or for tasking based on skills (such as selecting examiners for verification based on complementary skill sets). Certified examiners had higher conclusion rates than non-certified examiners without a significant change in accuracy. Length of experience as a latent print examiner did not show a significant correlation with conclusion rates.

6.1.2 The Black Box Repeatability Study[132]

In the *Black Box Repeatability Study*, we retested 72 examiners on comparisons they had performed in the original *Black Box*, after an interval of approximately seven months; each examiner was reassigned 25 image pairs for comparison. We also had 900 cases in which an examiner saw the same latent twice, within hours or days (described here as "within-test" repeatability, limited to Value determinations). We compared these repeatability (intraexaminer) results with the reproducibility (interexaminer) results derived from our previous study.

We used percentage agreement ($\overline{P}$) to describe both intra-examiner agreement (repeatability) and inter-examiner agreement (reproducibility). This commonly used statistic simply describes the proportion of times paired responses are in agreement — either multiple raters on the same test item in the case of reproducibility, or the same rater in the case of repeatability. Percentage agreement ($\overline{P}$) is defined as follows. Let P_i represent the extent of agreement on the i^{th} image (or image pair):

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij}-1)$$

where n is the number of determinations, k is the number of determination categories, and n_{ij} is the number of determinations assigning the i^{th} image (or image pair) to the j^{th} category. P_i is a proportion and can take on values from 0 to 1. When calculating reproducibility, n represents the number of examiners deciding on the i^{th} image (or image pair). When calculating repeatability, $n=2$, representing the test and retest determinations made by one examiner.

$\overline{P}$ is simply the mean agreement over a set of N test questions (images or image pairs):

$$\overline{P} = \frac{1}{N} \sum_{i=1}^{N} P_i$$

6.1.2a Repeatability of value determinations

On the question of whether a latent was of value for individualization (2-way decision: {VID, not VID}), repeatability was $\overline{P}$ = 89.7% (Figure 16A). When examiners were required to further differentiate NV from VEO (3-way decision: {VID, VEO, NV}), repeatability dropped to $\overline{P}$ = 84.6% (Figure 16B). Complete reversals (between NV and VID) occurred at the rate of 1%. The *Within-test* repeatability data showed very similar results: repeatability was $\overline{P}$ = 92.2% (2-way) and $\overline{P}$ =

[132] *Derived from [BBRR]*

88.8% (3-way); complete reversals (between NV and VID) also occurred at the rate of 1%. Reproducibility of VID determinations was unanimous on 42% of the latents. The extent of unanimity reflects the data selection: this test was designed to focus on difficult image pairs; if the test had included more latents that were obviously of value or obviously of no value, the overall reproducibility of value determinations would have been higher.

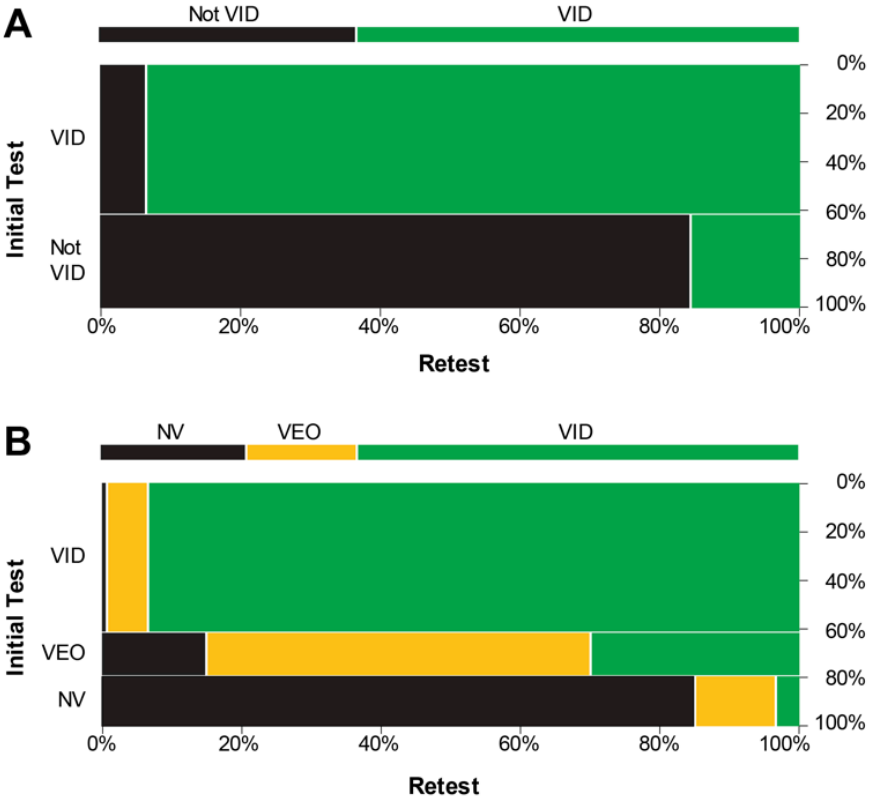


Figure 16: Repeatability of latent value determinations: (A) 2-way {VID, Not VID} latent value decisions; (B) 3-way latent value decisions {NV, VEO, VID} including category "value for exclusion only".

Low repeatability of value determinations was almost entirely restricted to latents on which reproducibility was also low (Figure 17). On the retest, changed decisions occurred on nearly half of the latents on which there was not unanimous agreement among examiners (mean of 5.0 retest decisions per latent). Among the 197 images on which there was not unanimous reproducibility, repeatability was $\overline{P}$ = 83.3%; on these same 197 images, reproducibility was $\overline{P}$ = 75.2%. This association demonstrates that in almost all cases, the specific images on which examiners individually were not consistent in their own decisions also resulted in disagreement among examiners.

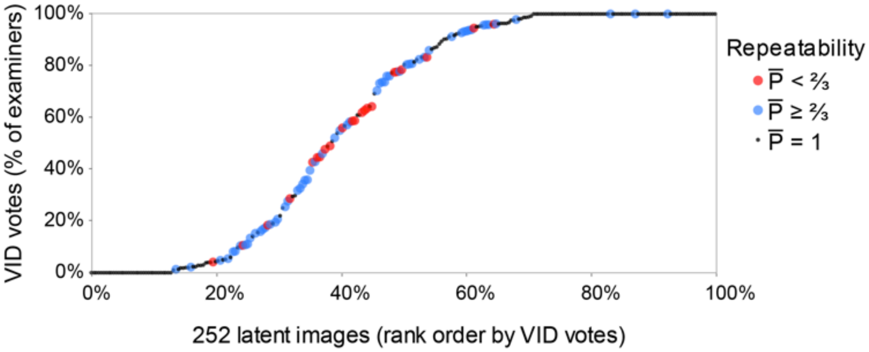


Figure 17: Repeatability and reproducibility of 2-way latent value decisions {VID vs. Not VID}. Percentage of examiners rating each latent VID (y-axis), in rank order (x-axis), color-coded by repeatability; n=252 latents on which at least 3 examiners were retested. Examiners were initially unanimous on 107 of these 252 latents; value determinations changed on 3 of these. Reproducibility rates were based on 53.2 mean examiners per latent (s.d. 21.7); repeatability rates were based on 5.0 mean examiners per latent (s.d. 2.3).

6.1.2b Repeatability of comparison determinations

For nonmates, repeatability based on three decision categories {VID individualization, exclusion, no conclusion} was $\overline{P}$ = 85.9%: 90.6% of (true) exclusion determinations were repeated; 73.1% of no conclusion determinations were repeated. For mates, repeatability (based on the same three decision categories) was $\overline{P}$ = 90.3%; 89.1% of VID individualization determinations were repeated; 90.9% of no conclusion determinations were repeated. Most of the difference in the repeatability of no conclusion determinations between the mated and nonmated sample populations may be explained by the fact that the mated dataset included a much higher proportion of poor-quality images.

As we saw with value determinations (Figure 17), low repeatability of comparison determinations was almost entirely restricted to latents on which reproducibility was also low (Figure 18). The majority of determinations that were not repeated changed to or from inconclusive or VEO determinations: most of the intra-examiner inconsistency was with respect to sufficiency to make a conclusion.

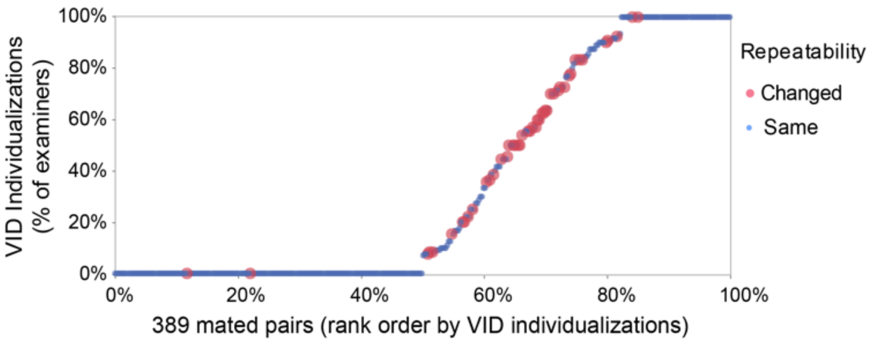


Figure 18: Repeatability and reproducibility of 2-way individualization decisions {VID individualization, other}. Percentage of examiners individualizing mated image pairs (y-axis), in rank order by VID individualization (x-axis), color-coded by repeatability. Y-axis is based on 4,006 initial determinations (excludes false negative responses; 10.3 mean examiners per image pair; s.d. 2.6). Color-coding is based on 792 retest determinations on 389 mated image pairs (2.0 mean examiners per image pair; s.d. 1.1). Non-repeated determinations occurred on 46 of the 389 image pairs. Reproducibility was unanimous on 257 of the 389; determinations were not repeated on 2 of these.

Examiners repeated 89.1% of their individualization determinations, and 90.1% of their exclusion determinations; most of the changed determinations resulted in inconclusive determinations. Repeatability of comparison determinations (individualization, exclusion, inconclusive) was 90.0% for mated pairs, and 85.9% for nonmated pairs. Repeatability and reproducibility were notably lower for comparisons assessed by the examiners as "difficult" than for "easy" or "moderate" comparisons, indicating that examiners' assessments of difficulty may be useful for quality assurance.

Six false positives were committed by five examiners on the initial test: none of these errors were reproduced in the initial test, and none were repeated in the retest (n=4; the examiner who made two errors in the original test did not retake the retest). No new false positive errors were committed during the retest, which is consistent with the false positive rate of 0.1% on the initial test.

The retest participants committed false negative errors at the rate of 8.8% (FNR_{CMP}) on the initial test,[133] of which 30.1% were repeated. We estimate the probability that another examiner would reproduce one of these errors to be 19%. We understand these comparative results as follows: "self-verification" (several months later) detected 69.9% of the false negative errors, whereas independent examination of the same images by another examiner (analogous to blind verification) would have detected an estimated 81%. We estimate that if every exclusion determination were verified, the resulting rate of erroneously corroborated false negatives would be 2.7% ("self-verified") and 1.7% (blind-verified). Interestingly, much of the relative benefit of blind verification over this type of self-verification relates to the wide variability in FNR by examiner: false negative errors are produced disproportionately by those examiners with high FNRs, so another examiner selected at random to perform verification is likely to have a lower FNR. Difficulty was not predictive of whether false negative errors would be repeated; the data suggest that greater difficulty is weakly associated with lower reproducibility for false negative errors. Although most errors were not repeated on the retest, examiners did introduce new false negative errors. After correcting for the difference in test mix between the initial test and the retest, no significant net change in false negative error rate was observed.

[133] *Slightly higher than the 7.5% FNR in the overall Black Box Study.*

Most of the inter- and intra-examiner variability concerned whether the examiners considered the information available to be sufficient to reach a conclusion (e.g. individualization vs. inconclusive). This variability was concentrated on specific image pairs such that repeatability and reproducibility were very high on some comparisons and very low on others. Much of the variability appears to be due to discretization error, making categorical decisions in borderline cases. Lack of repeatability or reproducibility is much more understandable if we consider, for example, that an examiner may (consciously or unconsciously) be 51% convinced that individualization is more appropriate than inconclusive in the comparison at hand. I feel that the s-curves shown in Figure 17 and Figure 18 are as effective a means as I have encountered of showing why probabilistic models are needed to replace or augment examiners' determinations. The s-curves show that the decision space for value and individualization determinations is really a continuum of how certain the examiners are that a determination is warranted; the instances on the slope of the s-curve show that the categorical responses that examiners are required to make are not well-suited to the data. Given this, I do not see the 85-90% repeatability and reproducibility rates in this test as a criticism of the examiners, but a criticism of the system: about 10-15% of the test data had no obvious answer, and therefore it is unreasonable to expect the answers to be consistent in those instances; for some comparisons the answer is clear, and therefore repeatability and reproducibility are high.

6.2 The Sufficiency for Value Study[134]

In the Black Box study we showed that there is often uncertainty as to whether a given determination is warranted, as shown by imperfect repeatability and reproducibility. The issue has to do with sufficiency: whether examiners (individually or collectively) consider that the information available in the impressions is sufficient to make a given determination. The objective of the *Sufficiency for Value Study* was to describe how image clarity and feature content are associated with the assessment of latent value by latent print examiners. Our motivations for studying the associations between latent markup and value determinations were to understand the variability in latent print examiners' value determinations, to determine if there is a basis for defining value based on the clarity and quantity of features, and to develop a basis for understanding sufficiency for comparison determinations.

6.2.1a Materials and Methods

This study analyzed extended friction ridge feature markup and value determinations for 1850 latent fingerprints; the latents, markups and value determinations were the products of the ELFT-EFS, Quality, and Black Box studies. International Association for Identification (IAI) Certified Latent Print Examiners marked extended friction ridge features and made latent value determinations in compliance with EFS. Latents were marked without reference to exemplars. The examiners used ULW to manually mark the images and record their value determinations. Twenty-one examiners performed the markup, but six examiners produced about half of the markups, so these results cannot be assumed to be broadly representative of all examiners. In general, individual examiners performed the markup; however, for 17% of the latents, groups of four examiners collaborated in an attempt to produce the best possible markup of each latent.

6.2.1b Results

Value assessments were associated with the number of minutiae marked, as shown in Figure 19: notice that there are no sharp decision thresholds (specific minutia counts that divide one type of determination from another). Budowle et al.[135] discussed an informal threshold of seven or more

[134] *Derived from [SuffValue]*

[135] *[Budowle06]*

minutiae used by some examiners to "proceed with an analysis." In our data, while there is clearly no threshold at seven minutiae, 1% of VID (and 50% of Not VID) determinations were made on latent prints with fewer than seven minutiae.

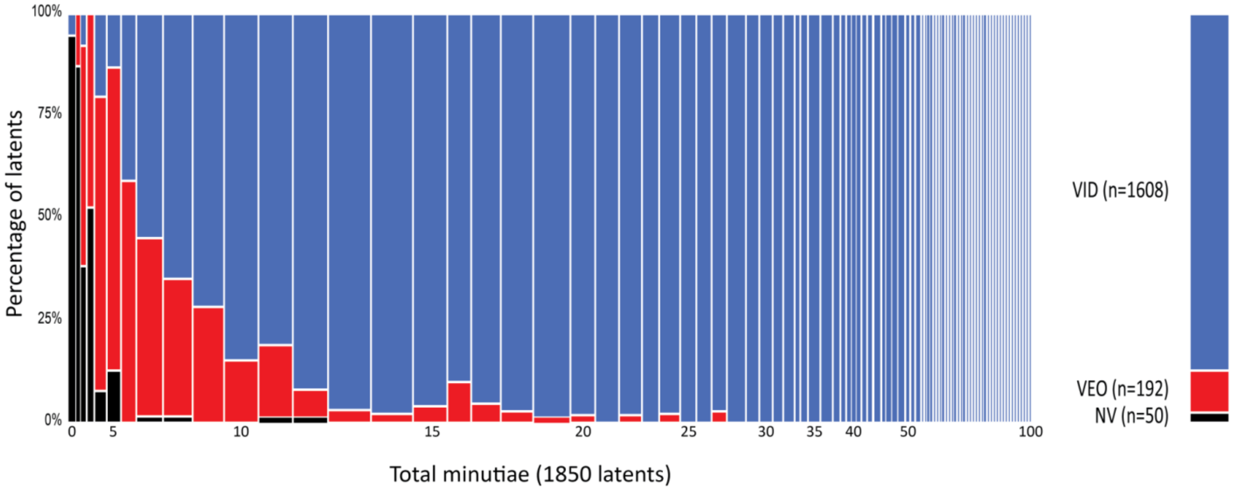


Figure 19: Association of minutia count and value assessments. Heights represent the percentage of each type of determination associated with each count; widths are proportional to the counts in each bin.

Figure 20 uses receiver operating characteristic (ROC) curves to show the strength of associations between an examiner's marked features and that examiner's VID determinations. The left chart describes the data presented in Figure 19, but in terms of the error trade-offs that result when predicting VID determinations by varying a minutia count decision threshold. At a threshold of 12 or more minutiae (the national standard in many countries), 84% of the VID determinations would have been correctly predicted, but 12% of the NV and VEO determinations would have been incorrectly predicted to be VID.[136]

[136] *The ROCs presented here provide meaningful comparisons of models on our data, but should not be extrapolated beyond this data: a different distribution of data could substantively change the rates shown in these graphs.*

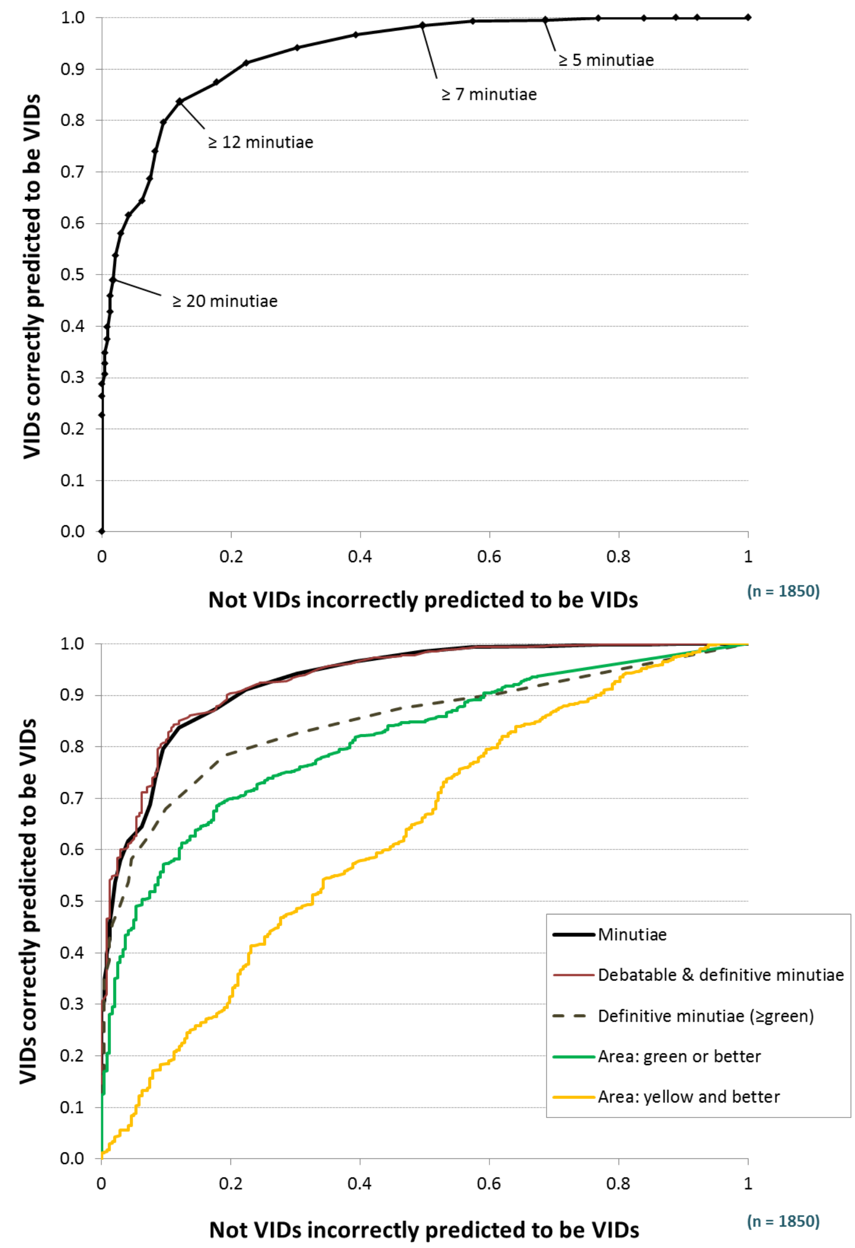


Figure 20: (Left) Receiver operating characteristic (ROC) curve showing the accuracy of a simple model using minutia count as a predictor of value for individualizaton. The x-axis shows the false positive rate (1-specificity); the y-axis shows the true positive rate (sensitivity). Minutia counts are indicated as dots; four counts are labeled as examples. (Right) Comparison of various logistic regression models predicting VID determinations from examiner markup; the minutiae curve is the same in both charts.

Figure 20 (right) shows results from several models predicting VID determinations from metrics derived from latent markup. When comparing alternative models using ROCs, stronger associations result in operating points closer to the upper left corner; a lack of any association would result in a diagonal line from the top upper right corner to the bottom lower left corner. While most of the metrics had some predictive capability, none of the individual metrics approached minutia count as a predictor of VID determinations. We used logistic regression and recursive partitioning to explore each pairwise combination of metrics with respect to VID determinations; none of these combinations provided significant discriminatory power beyond minutia count alone. For example, adding terms such as total area marked as green or higher clarity, or counts of cores and deltas did not improve the model; likewise separately weighting debatable minutiae (those in areas of yellow clarity) and definitive minutiae (those in areas of green or higher clarity) did not improve upon the minutia count model.

As we saw in the results of the *Black Box Study*, one examiner's value determinations are not always reproduced by other examiners. Figure 21 shows the extent to which agreement on value determinations is associated with minutia count. No prints with one or more minutiae marked were unanimously rated NV; only one latent print with more than nine minutiae marked was rated NV by a majority of examiners. No prints were unanimously rated VEO; only one print was rated VEO by more than 75% of examiners. No prints with fewer than ten minutiae marked were unanimously rated VID; only two prints with fewer than ten minutiae marked were rated VID by a majority of examiners.

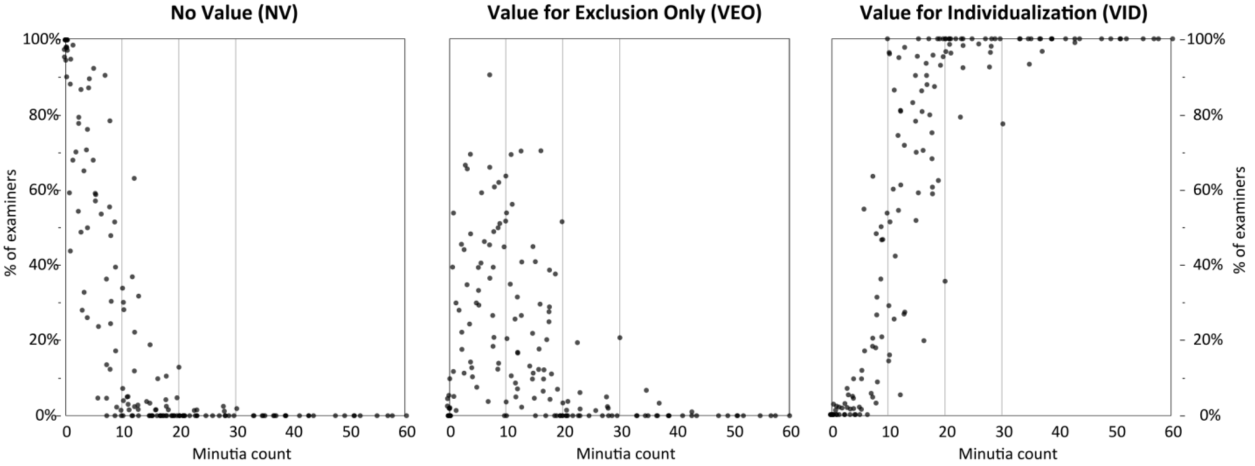


Figure 21: Reproducibility of value determinations by minutia count; n=166 latents with one markup each, but a mean of 56 examiners providing values determinations for each.

Figure 22 summarizes the strength of association of one examiner's markup and the value determinations of other examiners. The "value" ROC shows the effectiveness of using examiners' value determinations to predict the value determinations of other examiners; for example, the marking examiners' VID determinations successfully predicted 95% of the Black Box VID determinations, but also predicted 31% of the Black Box Not VID determinations. The informal GBU scale is as effective as minutia count in predicting VID determinations (although less continuous).

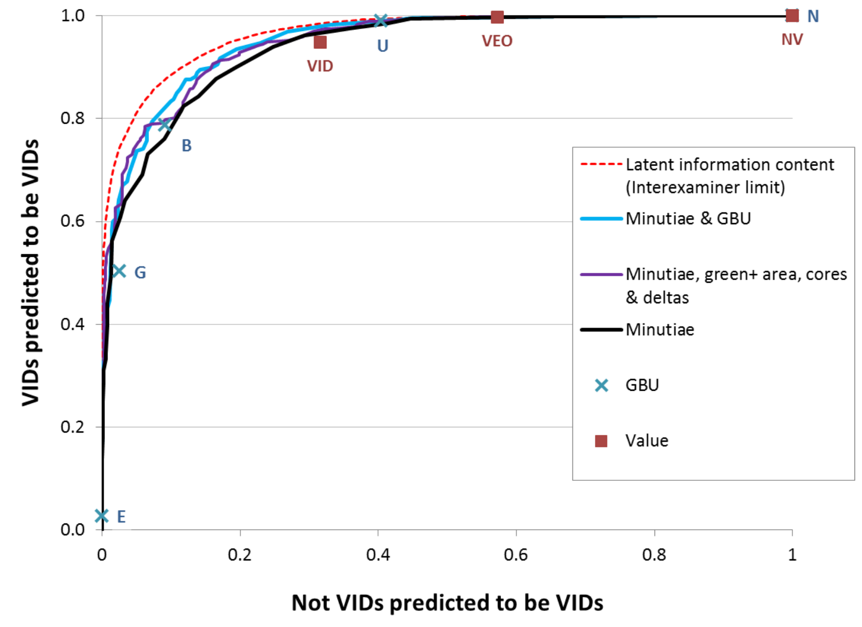


Figure 22: Predicting independent value assessments for VID determinations. These ROCs compare models for predicting a second (independent) value assessment {Not VID, VID} from an initial markup and value assessment (n=166 latent prints with markup and value assessments, predicting 9,322 independent value assessments). The "inter-examiner limit" describes a logistic model having one parameter for each of the 166 latent prints; this shows the limit imposed by inter-examiner variation on value determinations. The “value” ROC uses the initial examiner’s value determination to predict the independent value assessments. (n=166)

Since examiners do not always agree on value determinations, it is not possible to have any model that always predicts value: the disagreement sets a ceiling on how well any model could perform. This is shown in Figure 22 as the inter-examiner limit curve (dashed red), which describes a logistic regression model with 166 parameters, one for each latent print. This model accounts for all of the variability in value determinations that can be attributed to the prints themselves; the remaining variability arises from examiner disagreements on their value determinations. This model represents an upper limit to what might be achieved in any model derived from the markup of these prints. Therefore, even models that account for the specificity or relationships of features, or additional feature or clarity-based metrics would not exceed this limit. The minutia model has an equal error rate (where FPR equals 1-TPR) of 15%, whereas the red curve shows that the equal error rate cannot be less than 10% for any model that is based on latent print characteristics alone: two-thirds of the residual error resulting from the minutia model is explained by examiner variability on value determinations.

Examiners vary substantially in their minutia counts. Figure 23 shows associations between the value determinations and minutia counts for 387 latents; two examiners independently annotated each print. Examiners often disagreed substantially in their minutia counts: the standard deviation for the difference in minutia counts is 4.0 among latent prints with a mean minutia count of 5 to 15. The dispersion (disagreement on minutia counts) increases as the number of minutiae increases, roughly in proportion to the square root of the mean minutia count. Dispersion is substantially higher among those latent prints on which examiners disagreed on the latent value than among those where examiners agreed. When examiners disagreed on latent value, the examiner making the higher value assessment usually counted more minutiae (n=37; p=0.001, one-sided): the mean difference in minutia count was 2.5 (s.d. 4.7); the mean minutia count was 8.1 (s.d. 5.2). The intra-examiner associations between minutia count and VID determinations are not substantially stronger than the inter-examiner associations.

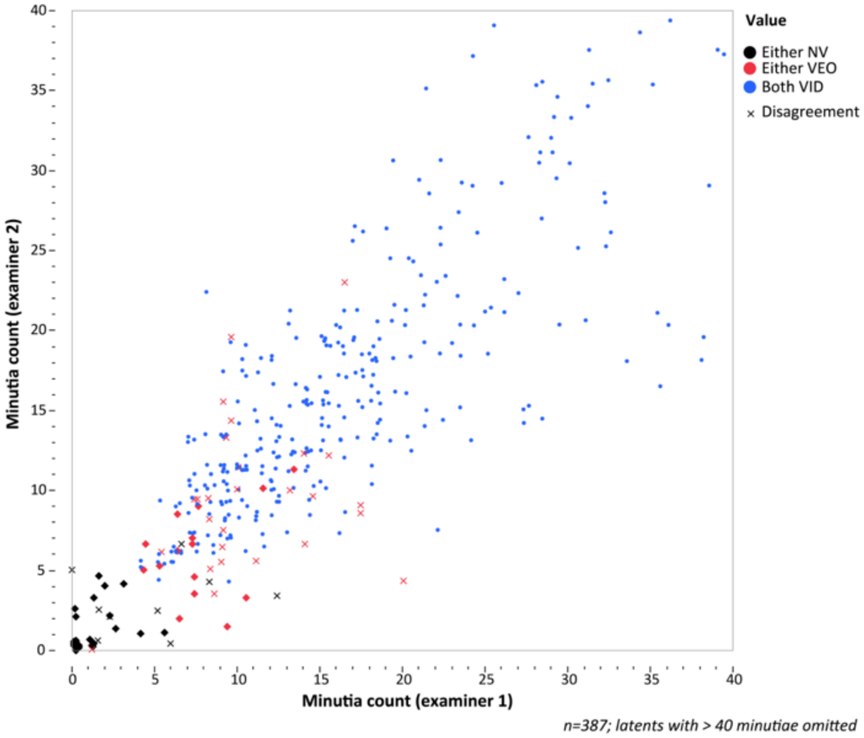


Figure 23: Variation in value determination between two examiners assessing the same latent (only 387 of 421 latent prints shown due to truncating axes at 40 minutiae)

6.2.1c Discussion

None of the metrics other than minutia count provided significant additional discriminatory power for VID determinations: after accounting for minutia count and examiner disagreements, the theoretical potential for other metrics to contribute is relatively small. For discriminating between NV and VEO determinations, pattern classifiability and counts of cores and deltas were effective in combination with minutia count.

A surprising result of the study was the failure of image clarity metrics to improve substantially on the minutia count model. The decision space for analysis and comparison determinations has been described as having two dimensions: quality and quantity of features.[137,138] We would therefore expect latent value determinations to depend on both dimensions, so that VID determinations could be associated with high clarity, low minutia count prints as well as low clarity, high minutia count prints. In our data, we found a strong association between clarity (quality) and minutia count (quantity): although print clarity was strongly associated with value determinations, clarity metrics provided little additional discriminatory power beyond minutia count alone. For example, areas of clear level-3 detail (e.g. ridge edge details) were only present in prints with high minutia count. Several factors might account for the failure of clarity metrics to substantially complement minutia

[137] *[Vanderkolk09]*

[138] *[SWGFAST-Conclusions13]*

count: clarity and quantity of features were highly associated in our data; the metrics used may not fully capture clarity in the ways that examiners use that information; examiners may not have used the clarity categories consistently in their markup; our data included a relatively small sample size of NV and VEO determinations, limiting our ability to observe small effects; and the general lack of repeatability and reproducibility of both markup and value determinations limits the potential for improving models beyond minutiae alone.

This study does not address whether examiners' value determinations are correct: it only reveals patterns of association. In theory, the correctness of value determinations could be based on whether examiners could subsequently make correct comparison conclusions given a suitable exemplar; however, this is unrealistic given the variability in examiners' comparison determinations, and the impracticality of determining the correctness of the comparison conclusions. Instead of correctness, we evaluated the appropriateness of individual value determinations based on consensus among latent print examiners. There may be situations where examiner subjectivity in value determinations is acceptable: for example, a skilled examiner may make a VID determination for a print that would be far beyond the expertise of a junior examiner to compare. Should examiners make value judgments based on their own skill levels, or based on their expectation of other examiners' skills? If a forensic laboratory wished to report reproducible value determinations regardless of the examiner assigned to the case, then examiners would have to predict the value determinations that would be made by other examiners.

As we discussed previously, the value of latent prints is a continuum that is not well described by binary (value vs. no value) determinations. Additional means of describing value (such as an examiner's assessment that a print is "complex",[139] or a quality metric such as discussed in Chapter 5) may be useful in flagging prints whose value determinations are likely to be debatable. Such means of expressing value could be used in establishing business processes to manage risk and optimize workload based on value: for example, a quality assurance process could have different procedures for low-quality (or ugly or complex) latents, such as review of value determinations, directing such prints to highly qualified examiners, or requiring additional verification when such prints are used in comparison. Agencies should not expect that value determinations on complex prints would be reproducible. Frequently, NV determinations are not verified; although inappropriate VID determinations will often be detected by verification of the subsequent comparison determinations, inappropriate NV determinations will not be detected, potentially resulting in missed conclusions.

6.3 The White Box Study

The Black Box studies were named such because they treat each examiner as a black box: the evaluations provides input (images), gets output (determinations), but there was no attempt to assess how or why the examiners made a given determination. This was not due to lack of interest, but was because we needed a detailed understanding of black box results as a baseline before we addressed the complex problem of delving into the basis for determinations.

Our *White Box Study* was conducted as a counterpoint to the *Black Box Study*. This experiment was designed to investigate the relationship between examiners' annotations and their determinations: examiners annotated features, clarity, and correspondences between latent and exemplar fingerprints to document what they saw when performing examinations. This is termed a "white box" approach because it attempts to understand the bases for the examiners' determinations, as opposed to the "black box" approach, which only evaluates the determinations. The White Box Study was addressed in three reports: *Sufficiency for Individualization* (Section 6.3.2), *Analysis to Comparison* (Section 6.3.3), and *Interexaminer Variation in Minutia Markup* (Section 6.3.4).

[139] *[SWGFAST-StdExam13]*

6.3.1 White Box Methods and Materials[140]

The test procedure was designed to correspond to that part of ACE casework in which a single latent is compared to a single exemplar print (latent-exemplar image pair). During the Analysis phase, only the latent was presented; the examiner annotated clarity and features and recorded a value determination: value for individualization (VID), value for exclusion only (VEO), or no value (NV). If VID or VEO, the examiner proceeded to the Comparison/Evaluation phase,[141] in which the exemplar was presented for side-by-side comparison with the latent: the examiner annotated clarity and recorded a value determination for the exemplar; compared the two images and further annotated the features to indicate correspondences and discrepancies; recorded a comparison determination (individualization, exclusion, or inconclusive); and indicated the difficulty of the comparison. The Verification phase was not addressed in this study. Examiners could review and revise their work prior to submitting their results. Examiners were free to modify the annotation 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.

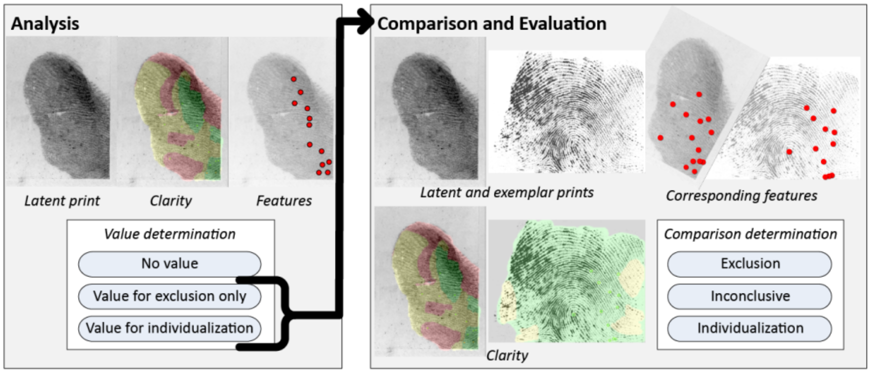


Figure 24: Test workflow. Each examiner was assigned a distinct, randomized sequence of latent-exemplar image pairs. For each pair, the latent was presented first for a value determination. If the latent was determined to be no value, the test proceeded directly to the latent from the next image pair; otherwise, an exemplar was presented for comparison and evaluation.

The software application used for our experiment was a variant of the ULW Comparison Tool. It included tools for annotating the fingerprints, simple image processing, and recording the examiners' determinations. Fingerprint annotations complied with EFS; the test instructions were derived in part from [EFSMI]. In the Analysis phase, the examiners provided the following annotations pertaining to the latent: local clarity map (produced by "painting" the images using the six EFS color-coded levels of clarity); 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 (VID, VEO, or NV). If the latent print was determined to be VEO or VID, the examiner provided the following annotations during the Comparison/Evaluation phase: latent and exemplar clarity; latent and exemplar features, as well as correspondences (definitive and debatable) and discrepancies; latent and exemplar value determinations; comparison determination (individualization, exclusion, or inconclusive); and comparison difficulty (very easy/obvious, easy, moderate, difficult, very difficult).

[140] *Derived from [SuffID]*

[141] *I capitalize Analysis and Comparison to indicate that I am referring to the ACE phases. For brevity, I use Comparison to refer to the Comparison/Evaluation phase.*

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: 90% were certified (or qualified by their employers) as latent print examiners; 82% were from the U.S.

6.3.1a Fingerprint selection and assignments

Fingerprints were collected under controlled conditions for research and selected from operational casework. In the Black Box study, in which much of the focus was on the correctness (accuracy) of the determinations, we only used images collected under controlled conditions because it was critical that the mating be known definitively. In this study, it was less critical that the mating be known with certainty because the objective was to investigate the bases for examiners' determinations, not their correctness. Here, in order to increase the variety of attributes (such as substrates, matrices, and processing methods), we included prints from operational casework. Mating of casework prints was established through the use of multiple additional corroborating latents and exemplars that were available in these cases; mating was not established solely through the use of the latents presented in the test.

Nonmated pairs were selected to result in challenging comparisons. They were prepared by down-selecting among exemplar prints returned by searches of over 58 million subjects (580 million distinct fingers) in the FBI's Integrated AFIS (IAFIS), and among neighboring fingers from the same subject; neighboring index, middle, or ring fingers from a subject often have similar fingerprint pattern classifications and therefore are more likely to be similar than two random fingerprints.

Although the fingerprints actually came from casework or were collected to resemble examples from casework, the sampling strategy was not designed to yield a mix of prints that would be representative of typical casework. Instead, 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). These four dimensions were selected to evaluate their effects on individualization determinations. The sampling method emphasizes pairs with low counts of corresponding minutiae in order to focus on the threshold between individualization and inconclusive, with the implication that our results would show lower interexaminer reproducibility than would be typical in casework.

We assigned 22 image pairs to each examiner: in order to concentrate the test design on sufficiency for individualization, each examiner was assigned 17 mated pairs and 5 nonmated pairs. The emphasis on mated pairs was not revealed to participants; the true proportions would have been obscured through NV determinations, inconclusive determinations, and erroneous exclusions.

The test yielded 3,730 valid responses from the Analysis phase (170 examiners, mean 12.4 examiners per latent). Among these were 2,796 responses on mated pairs with valid responses from both phases (165 examiners, mean 12.1 examiners per image pair). Different analyses used different subsets of these responses.

For comparisons that resulted in three or more corresponding features, each examiner's clarity maps for the latent and exemplar were superimposed using a thin-plate spline deformation model (method detailed in the Distortion Study); a "corresponding clarity" map was then defined as the minimum clarity at each location of the two superimposed maps, as described in [AssessingLC]. Also, for each image and each image pair, the clarity maps from all examiners who were assigned that pair were combined to produce median clarity maps representing a group consensus, reducing the impact of outlier opinions and imprecision. Clarity measures, including various area measures and the "Overall Clarity" metric,[142] were derived from each of the clarity maps (original, corresponding, and median).

6.3.2 Sufficiency for Individualization (White Box 1)[143]

This paper focused on the question of sufficiency for individualization: how much information do examiners require in order to make an individualization rather than inconclusive determination? What constitutes sufficiency for an examiner to reach an individualization determination is a critical question that has been the subject of extensive discussion and debate for many years. As part of our investigation, we sought to determine what information must be accounted for when describing the decision threshold, how the reproducibility of individualizations is associated with annotations, and to what extent disagreements among examiners arise from differing criteria as to what constitutes sufficiency vs. differing interpretations of the prints.

What constitutes sufficiency for individualization as opposed to inconclusive determinations? Here we explore the following aspects of that question: What is the association between examiners' annotations and their own determinations? What is the association between one examiner's annotation and another examiner's determination? What are the factors explaining the reproducibility of annotations and determinations among multiple examiners?

6.3.2a Materials and Methods

> *For a discussion of the materials and methods for the overall White Box study, see section 6.3.1 White Box Methods and Materials. This section discusses those aspects of materials and methods specific to the Sufficiency for Individualization study.*

In order to describe the decision boundary between individualization and inconclusive, we often restrict our attention to the 2671 mated pairs with inconclusive (including no value) or individualization determinations (i.e., omitting erroneous exclusions from the 2796 total White Box responses on mated pairs). We omit the exclusions because the decision criteria for exclusions and individualizations are distinct: an increase of corresponding information provides support for an individualization vs. an inconclusive determination, whereas an increase of discrepant or contradictory information provides support for an exclusion vs. an inconclusive determination. Exclusions may be based on pattern class information when there is insufficient information for individualization, or they may result from an examiner's determination that a single feature was discrepant despite otherwise having sufficient information for individualization.

6.3.2b Associations between examiners' annotations and their determinations

The number of minutiae annotated by examiners is strongly associated with their own value and comparison determinations (Figure 25). Value is a preemptive sufficiency decision: NV indicates that any comparison would be inconclusive. For both value (Figure 25A) and comparison (Figure 25B) determinations, a count of seven minutiae is a tipping point between determinations: for any minutia count greater than seven, the majority of ***value*** determinations were VID, and for any ***corresponding*** minutia count greater than seven, the majority of ***comparison*** determinations were individualization. Only sixteen individualization determinations (1% of all individualizations) had fewer than seven corresponding minutiae marked; most of these can be explained as having additional corresponding features (either nonminutia features or "debatable" correspondences) or as invalid annotation (features were marked in both images but not the correspondences). These

[142] *[AssessingLC]*

[143] *Derived from [SuffID]*

results are consistent with our previous findings on the sufficiency for value determinations,[144] as well as those of other researchers: Budowle et al.[145] discussed an informal minimum threshold of seven minutiae for value determinations; Langenburg[146] observed that examiners were more likely to make VID determinations than not VID starting at about seven to eight minutiae, and the crossover point for individualization was about eight to nine corresponding minutiae.

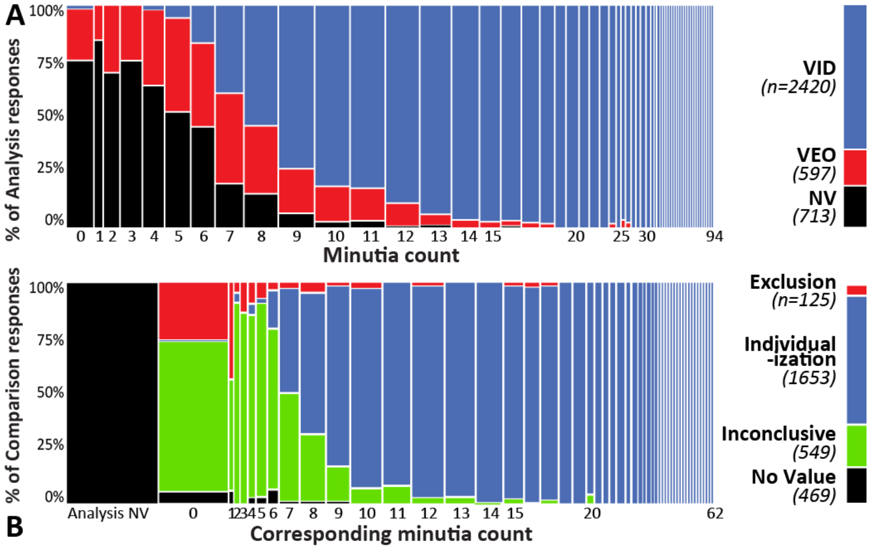


Figure 25: Associations of (A) minutia count and value determinations from analysis of the latent (n=3730); (B) corresponding minutia count and determinations from comparison of latent and exemplar prints on mated data (n=2796). In (B), 1.6% of determinations with 12 or more corresponding minutiae marked were not individualized. A few responses in (B) indicate NV with corresponding minutiae due to examiners changing their value determinations during Comparison.

High minutia counts are not limited to VIDs and individualizations: there are high-count VEO determinations (ranging up to 27 minutiae) and high-count inconclusive determinations (up to 20 corresponding minutiae); the majority of these determinations are on prints with discontinuous areas or low-clarity minutiae. Figure 25B also shows erroneous exclusions (red): these occurred at a lower rate (5.5%) than in the Black Box study.

Among nonmated image pairs, 89% had no corresponding minutiae marked, and few had more than seven corresponding minutiae marked. The single erroneous individualization (false positive) had 14 corresponding minutiae marked (the highest count among 582 comparisons of nonmated pairs, shown in Figure 31). In Figure 25B we see that when 14 corresponding minutiae are marked, individualization is the typical determination for mated image pairs, and therefore the minutiae count for the false positive does not stand out as an anomaly.

We evaluated a variety of models relating the probability that an examiner would individualize to factors derived from that examiner's annotations. For example, we use the following logistic regression model to relate the probability of individualization to corresponding minutia count (*CMin)*:

$$logit(\pi) = \beta_0 + \beta_{CMin} * CMin, \qquad (Eq\ 1a)$$

[144] *[SuffValue]*

[145] *[Budowle06]*

[146] *[Langenburg12a]*

where π is the probability of individualization for an examiner given *CMin* as marked by that examiner. This can also be expressed as

$$probability(ID) = \frac{1}{\left(1+e^{\left(-\beta_0+\beta_{CMin}*CMin\right)}\right)} \qquad \textit{(Eq 1b)}$$

We use misclassification rate as a summary statistic when comparing the models. Misclassification rates are calculated by treating the models as classifiers, where the model is interpreted as having predicted an individualization if and only if the estimated probability is greater than 0.5. As reported in Table 1, the fitted model from *Eq 1a* predicts that an examiner who marks eight or more corresponding minutiae will individualize, resulting in a misclassification rate of 6.0% (2.4% of mated pairs were individualized with CMin≤7; 3.6% were not individualized with CMin≥8).

To assess the effectiveness of this model, we can compare this 6.0% misclassification rate to the base misclassification rate for this dataset, which results from a (trivial) model with no independent variables that always predicts the most prevalent examiner response. In this case, the base rate model predicts that examiners would always individualize mated pairs, and therefore it misclassifies responses whenever examiners actually determined NV or inconclusive (38.1%). Misclassification rate describes the effectiveness of our models in explaining examiner determinations; it is ***not*** referring to whether the determinations made by examiners are "correct" or "incorrect." The misclassification rates reported here are specific to this dataset (the 2671 mated pairs, omitting erroneous exclusions) and are not estimates of operational rates.

Table 1 summarizes the performance of several models. Including additional modeling terms based on nonminutia annotations (clarity; cores, deltas, or other features; difficulty) did not markedly improve on the *CMin* model; this is a notable result given that we designed the study to measure the effect of these dimensions. This finding is consistent with our previous results regarding value determinations,[147] and those of Neumann et al.[148]

Predictors	Description	Misclass
None	*(base rate)*	38.1%
CD>0	*whether any cores or deltas were marked*	38.1%
Difficulty	*very easy to very difficult*	24.1%
OverallClarity	*area metric derived from corresponding clarity map*	17.1%
CMin>2	*whether corresponding clarity map could be created*	13.6%
CMin>0	*whether any corresponding minutiae were marked*	12.6%
CMin	*count of corresponding minutiae*	6.0%
CMin_yellow; CMin_green	*CMin in areas of debatable and definitive clarity*	6.0%
CMin; OverallClarity		5.8%
CMin; PtStd	*whether examiner followed a 12-point standard*	5.7%
CMin; Examiner	*Which examiner; 166 degrees of freedom*	3.0%

Table 1: Misclassification rates for models describing associations between annotations and individualization determinations by the same examiner (n=2671 responses by 165 examiners on 231 mated pairs).

We conducted analyses using analogous models associating annotations with latent value determinations; those findings generally parallel our findings for comparison determinations, and confirm and expand upon our previous findings reported in the *Sufficiency for Value* study.

The consistency with which participants annotated the image pairs had an impact on the strength of associations revealed by these models. For example, some examiners never marked cores or deltas, and the majority never marked "other" features (level-3 details). While markup of minutiae would be familiar to most examiners from AFIS searches and markup of cores and deltas from pattern classification, annotation of clarity and level-3 features would be novel to most participants.

[147] *[SuffValue]*

[148] *[Neumann13b]*

Corresponding clarity had a strong influence on sufficiency decisions, but that influence is subsumed by the count of corresponding minutiae: we presume that clarity is an important determinant of the selection of minutiae, but it has minimal additional effect after the minutiae are selected. Table 1 shows that most of the association captured by *OverallClarity* derives simply from whether or not the examiner marked corresponding minutiae: the *CMin>0* and *CMin>2* models explain much of the association; note that corresponding clarity maps can only be constructed if at least three corresponding points are marked.

The *CMin + Examiner* model includes a term indicating which examiner made the determination, resulting in a 3.0% misclassification rate. Specifically, the model becomes:

$$logit(\pi[i,j]) = \beta_0 + \beta_{CMin} * Cmin[i] + Examiner[j] \qquad (Eq\ 2)$$

where π_{ij} is the probability that image pair i is individualized by examiner j. β_0 and β_{Cmin} are fixed effects corresponding to an intercept and the number of corresponding minutiae (Cmin) marked on image pair i. $Examiner_j$ is a random effect due to examiner j.

The *Examiner* terms model each examiner's individual individualization rate. The remaining 3.0% could be explained by lack of repeatability of the examiner's association between *CMin* and determinations, inconsistent usage of annotations among examiners, other interaction effects between examiners and image attributes, or limitations of the metrics used.

6.3.2c Associations between corresponding minutiae and determinations

Figure 26 shows the association between corresponding minutia counts and determinations, as well as the reproducibility of counts and determinations among examiners. The strong association between examiners' minutia counts and their own determinations shown in Figure 25B is seen here as a color change in the vertical dimension. Figure 27 shows a subset of this data to more clearly reveal the interexaminer variability on each image pair. For most image pairs (x-axis), we see substantial interexaminer variability in both the corresponding minutia counts (vertical spread) and determinations (color). This extensive variability means that we must treat any individual examiner's minutia counts as interpretations of the (unknowable) information content of the prints: saying "the prints had N corresponding minutiae marked" is not the same as "the prints had N corresponding minutiae." The variability also implies that one examiner's minutia count is a weak predictor of another examiner's determination: for example, while we might have assumed that having one examiner mark 13 or more corresponding minutiae and individualize would guarantee that any other examiner would also individualize, that is not true; most of the mated image pairs had one or more examiners mark 13 or more corresponding minutiae pairs.

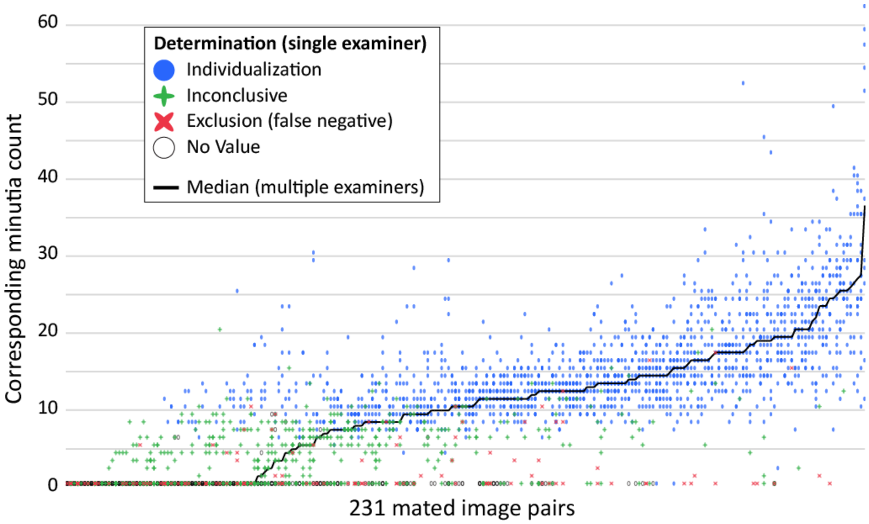


Figure 26: Corresponding minutia count (y-axis) and determination (color) by image pair (x-axis). Each column of points contains the set of all responses for a given image pair. Some points are superimposed, indicated through color blending. X-axis is sorted by median, then by mean corresponding minutia count. Latents that were determined NV and not compared are shown as having zero corresponding minutiae. NV responses with one or more corresponding minutiae are due to examiners changing their value determinations during Comparison. (n=2796 responses by 165 examiners to 231 mated image pairs.)

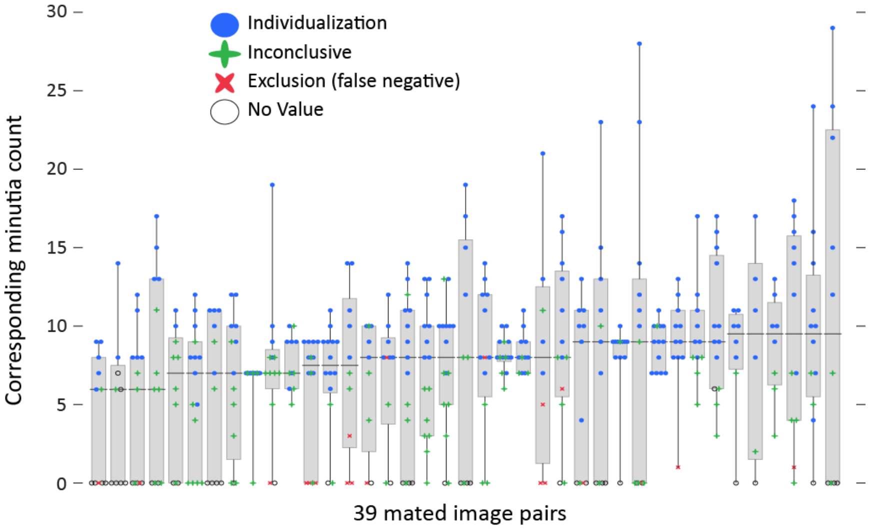


Figure 27: Detail of Figure 26 for the 39 image pairs that had median corresponding minutia counts between 6 and 9.5, with the addition of box plots showing interquartile range, minima, and maxima. (n=452 responses; 6 to 16 responses per image pair.)

In [BB] and [BBRR], we observed that variability in determinations was concentrated on certain image pairs, but did not characterize the attributes of those prints. In Figure 26 and Figure 27, we see that the reproducibility of determinations is associated with the median corresponding minutia count and is lowest on image pairs with a median corresponding minutia count between about six to nine. Interexaminer variability in corresponding minutia counts is seen across all image pairs, except where there is unanimous agreement on zero corresponding minutiae. Disagreements on sufficiency for individualization tend to be associated with substantial disagreements on corresponding minutiae; similar observations have been made previously.[149] When examiners made an inconclusive determination, they typically reported fewer than 12 corresponding minutiae; these counts were independent of the median count reported by those who individualized. The individual examiners' determinations generally transition from inconclusive to individualization between about six to nine corresponding minutiae, which is relatively independent of the other examiners' counts. An increasing median corresponding minutia count is associated with fewer examiners making inconclusive determinations. The variation in the counts remains even when examiners agree on individualization. However, the critical instances occur when annotation disagreements are associated with differing determinations. Failure to see correspondence is a notable cause for variation in the counts: on 42% of inconclusive determinations on mated pairs, examiners marked no corresponding minutiae. "Corresponding features" is only a particularly meaningful concept when the examiner is at least leaning toward individualization: if the examiner cannot find any areas of possible correspondence or "anchor points", marking no corresponding points would be the expected response. Individualization disagreements arose on 61% of mated pairs. When an examiner fails to individualize a mated pair that is individualized by another examiner, it is considered in some agencies as a "missed ID": 10% of responses were missed IDs on mated pairs that were individualized by the majority of examiners.

Differences in minutia counts understate the variability among examiners: annotations may have similar minutia counts but differ greatly in which specific minutiae were marked. Some differences relate to lack of concurrence in what constitutes minutiae, especially within cores and deltas. Some of the variability in minutia selection may be due to the examiners themselves not being consistent in their minutia selection: in this study, a small number of latents were presented to examiners twice, and substantial variability of the analysis minutiae that were marked was observed.

An individual examiner's corresponding minutia counts are not highly consistent descriptions of how well the image pairs correspond: given an image pair as a stimulus, the minutia counts are subjective responses with limited reproducibility among examiners. Based on our inspection of the annotated images, we notice several factors that contribute to interexaminer differences in which minutiae were marked. These include whether to mark minutiae that are not clear or are difficult to interpret; what constitutes a minutia close to cores and deltas; the extent of the region of interest, such as when marking discontinuous impressions; and how to mark features such as incipient ridges and dots, which some examiners marked as minutiae.

To quantify the variability in corresponding minutia counts and attribute it to specific sources, we use an Analysis of Variance main effects model with minutia counts as responses to the image pairs and the examiners to whom they were assigned:

$$CMin[i,j] = \beta_0 + \beta_{ImagePair}[i] + \beta_{Examiner}[j] + \varepsilon[i,j], \qquad (Eq\ 3)$$

where the betas are unknown parameters for an intercept, each image pair, and each examiner.

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). If examiners always agreed on the corresponding minutia count for each image pair, all of the variance would be attributed to image pair effects. We find that 65% of the variance can be attributed to image pair effects, 11% to examiner effects, and 24% is residual (Table 2). These examiner effects represent a tendency by some examiners to mark more minutiae than other examiners. This results in a

[149] *[Evett96, Langenburg09b, Langenburg12b, Neumann13b]*

standard deviation of 2.8 corresponding minutiae, after controlling for image pair effects; this value is large in relation to the critical range of about six to nine corresponding minutiae in which examiner determinations generally transition from inconclusive to individualization (Figure 25B). Some of the residual variance is likely to be associated with limited repeatability of minutia counts by individual examiners.

Random Effect	*St. Dev.*	*Variance*	*(95% bounds)*	*% of Total Variance*
Examiner	***2.8***	***8.1***	***(6.4 - 10.5)***	***11.0%***
ImagePair	***6.9***	***47.6***	***(39.7 - 58.1)***	***64.6%***
Residual	***4.2***	***18.0***	***(17.0 - 19.0)***	***24.4%***

Table 2: Image pair and examiner effects on corresponding minutia counts, showing restricted maximum likelihood estimates. (n=2796 responses by 165 examiners to 231 mated image pairs)

6.3.2d Predicting another examiner's individualization determination

From the Black Box study we saw that reproducibility of individualization determinations is much higher on some image pairs than others, but that study did not provide any data for predicting for a given image pair whether agreement would be high or low. The only current method to assess whether an individualization or inconclusive determination is appropriate in a particular case is by consensus among examiners. Therefore, it is of great interest to estimate the probability that one examiner's determination of sufficiency would be reproduced by other examiners, taking into account that examiner's expressed basis for the determination.

We evaluated several logistic regression models predicting individualization determinations by one examiner from the responses (annotation and determination) of another examiner to the same image pair (Table 3). As we saw when modeling associations between annotations and determinations by the same examiner, accounting for factors such as clarity or the examiner's rating of comparison difficulty does not substantially improve upon predictions based on *CMin* alone.

Predictors	**Misclass.**
None (base rate)	39.8%
Difficulty	26.3%
OverallClarity	23.7%
OverallClarity; CMin	20.9%
Determination ***{Individualization, Insufficient}***	20.5%
CMin	20.4%
CMin_green; CMin_yellow	20.0%
Determination; CMin	20.0%
CMin; Difficulty	19.9%

Table 3: Misclassification rates for models using one examiner's annotations and determinations to predict a second examiner's individualization determinations. (14,608 paired responses by 165 examiners, reweighted to n=231 mated image pairs) See Table 2 for definitions of predictor variables.

Comparing the paired-examiner models of Table 3 with the same-examiner models of Table 1 shows that although examiners' associations and determinations are strongly associated, these same annotations are not as strongly associated with other examiners' determinations; for example, the misclassification rate for paired-examiner models based on corresponding minutia count is 20.4% versus 6.0% same-examiner models. The reason for this difference is the substantial interexaminer variability in both corresponding minutia counts and determinations, both of which negatively affect this prediction. If annotations from multiple examiners are available (not typical in operations), we can predict determinations using voted metrics for each image pair, such as median *CMin*, which are less affected by the interexaminer variability in corresponding minutia count.

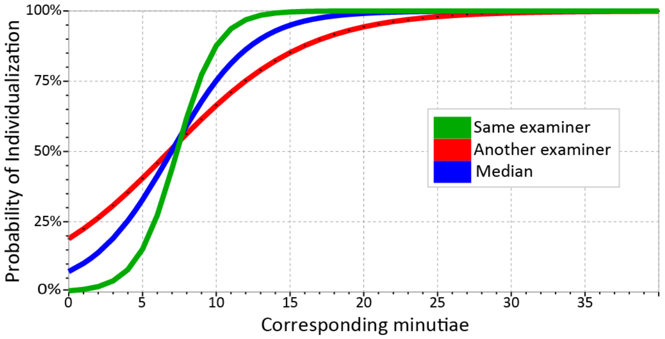


Figure 28: Logistic models estimating the probability of individualization based on corresponding minutia counts, on mated image pairs: (green) probability that an examiner would individualize based on the ***same*** examiner's corresponding minutia counts (6.0% misclassification, see Table 1); (red) probability that ***another*** examiner would individualize based on this examiner's minutia counts (20.4% misclassification, Table 3); (blue) probability that an examiner would individualize based on the ***median*** of all examiners' corresponding minutia counts (13.6% misclassification, Table 4).

Figure 28 shows the substantial differences in predictive ability among the same-examiner *CMin* model, the paired-examiner *CMin* model, and a model based on the median(*CMin*) across multiple examiners. All three models estimate approximately 50% probability of individualization at seven corresponding minutiae. However, the models differ on where they estimate 90% probability of individualization: when the ***same*** examiner marked 10 corresponding minutiae (green), when the ***median*** count was 13 (blue, median), or when ***another*** examiner marked 17 (red). Examiners' determinations are much more closely aligned with their own *CMin* than with others' *CMin*, limiting the effectiveness of using one examiner's annotations to predict other examiners' determinations.

6.3.2e Factors explaining agreement on sufficiency

Whether a given image pair would be individualized by an examiner can be seen as a function of that examiner's tendency to make individualization determinations and the tendency of all examiners to individualize that image pair. By modeling examiner determinations as dependent responses to which image pair was presented to which examiner, we can establish how much of the observed variation in examiner responses is associated with these two factors and the extent to which these two factors fall short of a full explanation. Letting $\pi[i,j] = Probability(Individualization)[i,j]$, for image pair *i* and examiner *j*, we can fit a logistic regression model such as

$$logit(\pi[i,j]) = \beta_0 + \beta_{ImagePair}[i] + \beta_{Examiner}[j], \qquad (Eq\ 4)$$

which has separate parameters for each image pair and each examiner (394 degrees of freedom). The relative contributions of examiner effects and image pair effects are summarized in Table 4A. Predicting individualizations based on which image pair was compared reduces misclassification from a base rate of 38.1% to 13.0% (Table 4A, *ImagePair*); this is equivalent to predicting the determination for an image pair based on majority vote (13% of determinations were in the minority). This 13.0% misclassification rate defines a limit for any model of this data based only on image attributes, as a necessary consequence of examiner disagreements on the determinations; if examiners were always unanimous on their individualization determinations, the misclassification rate for the *ImagePair* model would be zero. The *Examiner* model (32.8%) reduces misclassification from the base rate due to differences among examiners' individualization rates.

Having thus evaluated the overall magnitude of the image effects, we then fit simple models based on specific measures derived from the annotations (Table 4B). By comparing the models of Table

4B with those of Table 4A, we can assess how well those simple models explain the basis for sufficiency decisions. Note that the models describe image pairs using predictors that are fixed for each image pair (indexed by *[i]*, not by *[i,j]*) in order to model the effects of the image pairs on determinations. For this purpose, we use voted metrics derived from the annotations of multiple examiners to produce our best estimate of each attribute. The 13.6% misclassification of the *Median(CMin)* model is nearly as low as the rate for the *ImagePair* model (13.0%), and therefore accounts for nearly all of the observed variation in the examiner responses that could be explained by attributes derived from the image pair; paired-examiner models accounting for attributes such as clarity, complexity, or nonminutia features ***cannot*** introduce much additional predictive information, as they are bounded by the 13.0% misclassification due to the reproducibility of determinations. Just as we saw for same-examiner predictions, corresponding minutia count is the dominant factor in determinations.

	Predictors	DF	Misclass
	None (base rate)	0	38.1%
A	Examiner	164	32.8%
	ImagePair	230	13.0%
	ImagePair; Examiner	394	6.3%
B	CD_rate	1	31.6%
	MedianOverallClarity	1	24.3%
	CD_rate; MedianOverallClarity	2	23.1%
	Median(CMin); MedianOverallClarity	2	13.6%
	Median(CMin)	1	13.6%
	Median(CMin); Examiner	165	9.5%

Table 4: Misclassification rates for models describing individualization as a dependent response to (A) image pairs and examiners and (B) attributes of the image pairs as estimated by median statistics (derived from all examiner responses). n=2671 responses. CD_rate: proportion of examiners who marked a core or delta. MedianOverallClarity: Overall Clarity from the median corresponding clarity map. DF = degrees of freedom.

When we model individualization determinations as responses to both *Median(CMin)* and *Examiner*, the misclassification rate drops to 9.5% (vs. 13.6% for *Median(CMin)* alone); much of the further reduction to 6.3% in the *ImagePair + Examiner* model may be due to overfitting. We know from our previous research that a substantial proportion of determinations are not repeated on retest,[150] and we estimate that more than half of the 9.5% misclassification rate can be attributed to this lack of repeatability. The remainder of the misclassification is due to _ImagePair*Examiner_ interaction effects.

Comparing the models in Table 4 with those in Table 1 reveals that examiners' determinations are much more strongly associated with their own corresponding minutia counts than with the median estimates, as we saw in Figure 28. This implies that individual annotations are a good description of the basis for examiner determinations, as opposed to suggesting that examiners all tend to see and rely upon the same features, yet describe them inconsistently. The limited reproducibility of corresponding minutia counts demonstrates that the subjective annotations of these examiners do not consistently describe intrinsic attributes of the images themselves. By comparing the *ImagePair* model (misclassification rate 13.0%, Table 4) to the same-examiner *CMin* model (misclassification rate 6.0%, Table 1), we see that the individual examiner's minutia counts are part of a combined

[150] *[BBRR]*

response to the images that reflects the subjective outcome of the ACE process and goes beyond the consensus response to the images reflected in the *ImagePair* model.

6.3.2f Effect of point standard

Ten of the participants who indicated in the questionnaire that their agency or country has a 12-point standard conformed to that standard in their responses. Although one might expect that a high point count threshold would be associated with a lower individualization rate, participants following a 12-point standard were no less likely to individualize than those without a point standard. The individualization rate was 69% among those examiners following a 12-point standard (n=10) and 62% among the remainder (n=155); the difference is not statistically significant.

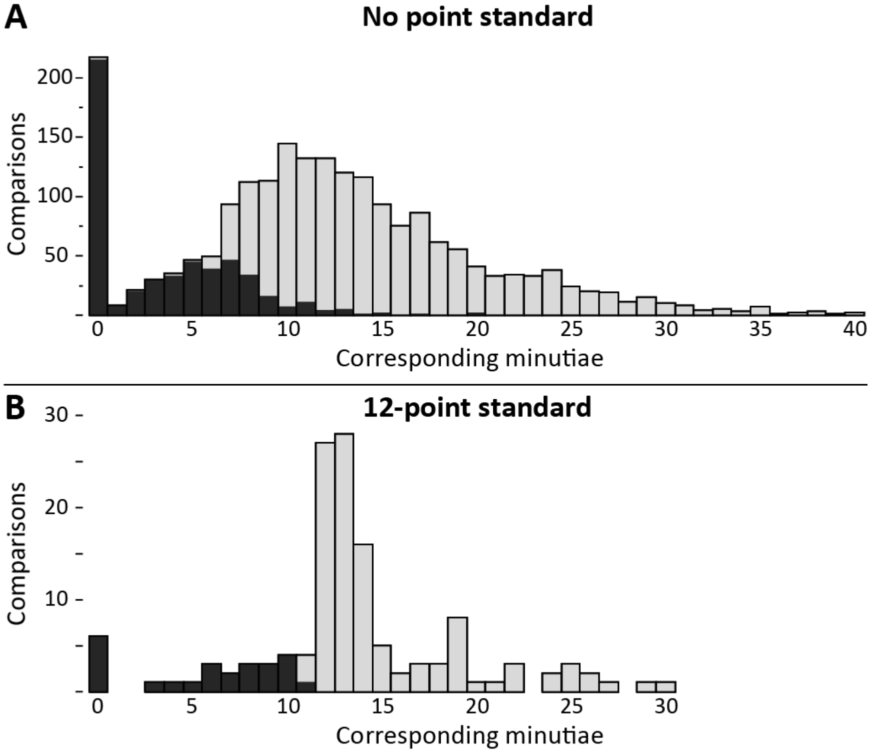


Figure 29: Distribution of corresponding minutia counts by (A) the majority of participants (n=2062 comparisons of mated pairs by 155 examiners) and (B) those participants following a 12-point standard (n=135 comparisons of mated pairs by 10 examiners). Colored by determination: inconclusive (black), individualization (gray); NV not included.

As shown in Figure 29, the number of corresponding minutiae examiners marked differed greatly between those following a 12-point standard and the remainder of participants. Given the balanced assignments, we would expect no substantial difference in these two distributions: we would expect a smooth distribution in the number of corresponding minutiae that examiners marked based on how the prints were selected. Instead we see abrupt steps in both distributions: those examiners following a 12-point standard were much more likely to mark 12 corresponding minutiae than 11, and those without a point standard were much more likely to mark seven corresponding minutiae than six. Evett and Williams[151] made a similar observation, noting that examiners following a 16-point standard avoided counting 15 points. These abrupt steps indicate

[151] *[Evett96]*

that examiners' counting appears to be influenced by their determinations. Conceptually ACE separates examination into different phases, so that corresponding features are defined in Comparison prior to the determination being made in Evaluation. However, these results indicate that we cannot assume causality between minutia counts and determinations. We might hypothesize that examiners subconsciously reach a preliminary determination quickly and this influences their behavior during Comparison (e.g., level of effort expended, how to treat ambiguous features); Stoney discusses this concept as a "leap of faith." [Stoney91] Similar results were found in the Analysis phase. The sample of participants following a 12-point standard is very small and not necessarily broadly representative of examiners who follow point standards.

6.3.2g Minutia thresholds

We have seen that across multiple examiners there is a gradual transition from inconclusive to individualization that can be described in terms of minutia counts. We might expect individual examiners to each have their own thresholds, and that these would vary from examiner to examiner with the consequence that some examiners individualize more often than others. The minimum number of corresponding minutiae that each examiner reported when individualizing varied among examiners. More than one-third of examiners individualized with eight or fewer minutiae, but others had a minimum count as high as 14. While some examiners based individualizations on fewer than seven minutiae, on review, all of the outliers with fewer than five corresponding minutiae can be explained as improper annotation, and most of the outliers with five or six corresponding minutiae rely on nonminutia features. After discounting the outliers that we believe were due to improper annotation, we did find examples of individualizations with as few as six corresponding minutiae, or five minutiae and two level-3 features.

Our analyses indicated that most of the dispersion in minimum minutia count is a consequence of the limited number of measurements obtained per examiner (i.e., small sample size: 17 mated comparisons per examiner). The minimum is an extreme statistic and biased upwards: if each examiner had been assigned many more comparisons, more opportunities would have lowered the observed minimum for many examiners. In particular, on a larger test, we would expect the proportion of examiners who individualized with seven or eight corresponding minutiae to increase. However, the small samples do not account for all of the variation in minimum minutia counts. As we showed above (Table 4A, *Examiner*), there are real differences among examiners' individualization rates, more than can be explained by the random test assignments. Our simulations demonstrate that these differences in individualization rates contribute very little to the dispersion in minimum minutia count. Nevertheless, we do observe some differences among examiners in the minimum number of corresponding minutiae marked when individualizing (beyond imprecision and chance): notably, some examiners only individualize when they mark nine or more corresponding minutiae. The simulations show that, apart from sampling limitations, the primary significance of a higher minutia count threshold appears to relate to differences in examiner judgment as to which features to mark (i.e., a higher minimum count means some examiners mark more minutiae than others on the same prints), not to differences in judgment as to which prints to individualize (i.e., a higher minimum count does not mean that they are less likely to individualize). Differences in individual minimum minutia count thresholds do not appear to be an important factor contributing to differing individualization rates.

6.3.2h Discussion

In a controlled study designed to ascertain the factors that explain examiners' determinations of sufficiency for individualization, latent print examiners recorded the bases for their determinations by providing detailed, standardized annotations of the fingerprints. The fingerprints used in this study were selected to test the boundaries of sufficiency for individualization determinations, and we deliberately limited the proportion of image pairs on which we expected examiners to have unanimous determinations; therefore, the reproducibility and error rates reported in this study should not be assumed to represent latent print examination in general.

While erroneous individualizations and exclusions are obvious concerns, differences in examiners' assessments of sufficiency also have serious operational implications. Such differences may result in conflict between examiners at the time of verification or in court. Disagreements among examiners on whether there is sufficient information to make an individualization does not imply that the determinations are erroneous (i.e., false positives or false negatives), as discussed in the Black Box study.

The study was designed to assess the associations between annotations and determinations, not to assess whether examiners' decisions to make individualization vs. inconclusive determinations were "correct" in an absolute sense. From the Black Box Repeatability study, we expected variability among examiners with respect to individualization determinations: we reported that two examiners agreed whether or not to individualize 86.6% of the time; in other words, 13.4% of the time a second examiner in that study would disagree whether the information content was sufficient to make an individualization determination. Disagreements on borderline decisions are expected, and requiring categorical decisions exaggerates examiner differences. Two examiners may both agree that a given decision is borderline, but reach different determinations in part because the discrete categories force them to make a choice.

The study revealed substantial differences among examiners' annotations. We cannot tell whether this is due to differences in how examiners see and interpret the data or merely to differences in how they document their interpretations. Differences in interpretation may arise at several points during examination: an examiner analyzing an unclear print must decide whether there is sufficient continuity when determining the limits of the region of interest to be used; an examiner analyzing a ridge within an unclear region must determine whether or not features are present; and an examiner must decide during Comparison whether potentially corresponding features are within a reasonable tolerance for differences in appearance. Each of these decisions may contribute to differences in interpretations and thus to differences in annotations. Additionally, there were many cases in which examiners made inconclusive determinations on mated pairs because those examiners failed to find any correspondences between the prints.

In addition to differences in interpretation, a lack of clear criteria in the latent print discipline specifying when and how to mark features may have contributed to much of the observed variability in annotations.[152] The lack of generally accepted and detailed standards for defining and recording the bases for conclusions limits the effectiveness of studies such as this, as well as the effectiveness of reviews of operational casework. Courts are now more frequently requiring that examiners demonstrate their bases for conclusions (during discovery, admissibility, and trial). Examiners are rarely trained specifically on how to interpret, select, and record features (other than for AFIS searches) in a standard, reproducible manner. Consistently applied and rigorously defined methods of performing and documenting ACE-V would result in a more transparent process, which could be more readily validated in research or in operations. Standardized annotation, such as the EFS markup used here, may be of operational benefit as a means of documenting and communicating the bases for examiners' determinations, especially for complex or disputed prints. Although the annotations collected in this study were based on recent standards, we recognize that the software and instructions were unfamiliar to many participants, and this may have contributed to the variability in annotations.

We found examiners' individualization determinations to be closely related to the number of corresponding minutiae marked. Other factors describing the fingerprints, such as clarity and level-

[152] *[Anthonioz08, Langenburg09a, Langenburg12b, Neumann13b]*

3 details, were not as strongly associated, and only a small proportion of the variability in determinations remains unexplained by corresponding minutia count. This finding is consistent with the Sufficiency for Value study and the findings of Neumann et al.[153] — although Neumann concluded that sufficiency is driven not just by the number of minutiae but also by the spatial relationships between the minutiae.

We designed our experiment to allow us to measure the extent to which various factors played a role in determining sufficiency for individualization, following the publication by SWGFAST of a conceptual Sufficiency Graph that depicts a complementary role between quality (an assessment of the overall clarity of the impression) and the quantity of minutiae for sufficiency for individualization.[154] We found, contrary to the SWGFAST proposition, that models accounting for clarity and minutia count performed no better than models that only accounted for minutiae count: we assume clarity influences which minutiae are marked rather than providing additional complementary information.

ACE distinguishes between the Comparison phase (assessment of features) and Evaluation phase (determination), implying that determinations are based on the assessment of features. However, our results suggest that this is not a simple causal relation: examiners' markups are also influenced by their determinations. How this reverse influence occurs is not obvious. Examiners may subconsciously reach a preliminary determination quickly and this influences their behavior during Comparison (e.g., level of effort expended, how to treat ambiguous features). After making a determination, examiners may then revise their annotations to help document that determination, and examiners may be more motivated to provide thorough and careful markup in support of individualizations than other determinations. As evidence in support of our conjecture, we note in particular the distributions of minutia counts, which show a step increase associated with decision thresholds: this step occurred at about seven minutiae for most examiners, but at 12 for those examiners following a 12-point standard. An interesting question for future research is to what extent examiners' latent value and comparison determinations may influence their use (and markup) of minutia and other features.

Although we expected variability in minutia counts, we did not expect the counts to vary as much as they did, especially in those critical cases in which examiners do not agree on their determinations and precise counting might be pivotal. The differences in minutia count understate the variability because the annotations not only differ substantially in total minutia counts, but also in which specific minutiae were selected. The limited reproducibility of minutia markup may be expected to have an operational effect on AFIS latent print searches, which are predominantly based on examiners' markup of minutiae; variability of annotations among examiners implies that search results would vary among examiners. Similarly, proposed models for probabilistic conclusions[155] based on examiners' minutia markup would result in different probability estimates for different examiners or even for the same examiner on different occasions.

Examiners' annotations are much more strongly associated with their own determinations than with those of other examiners. Neumann et al. observed the same result, noting that examiners are internally coherent, but consistency among examiners is low.[156] The observation that different determinations are often associated with substantially different annotations suggests that disagreements over sufficiency arise not only from differences in judgment about what constitutes sufficiency, but also from basic differences in interpretation of the prints.

[153] *[Neumann13b]*

[154] *[SWGFAST-Conclusions13]*

[155] *e.g. [Neumann12, Neumann13a, Abraham13]*

[156] *[Neumann13b]*

Whereas our previous Black Box study design was well-suited to estimating overall rates for errors and the reproducibility of determinations, one anticipated benefit of the white box approach used here was that the markups would reveal which determinations would be likely to result in disagreements related to the marginal sufficiency of the information. For quality assurance, it would be operationally desirable to flag sufficiency decisions that may be unreliable so that extra action could be taken: for example, flagging determinations that may not be highly reproducible, or flagging instances in which an examiner's determinations do not follow from that examiner's own markup. However, because of the limited reproducibility of minutia counts and determinations, one examiner's annotation and determination are often unreliable predictors of another examiner's determination. More consistency in annotations, which could be achieved through standardization and training, should lead to process improvements and provide greater transparency in casework.

6.3.3 Analysis to Comparison (White Box 2)[157]

In *Analysis to Comparison* we describe 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,[158] 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."[159] Others argue that a recurring, reversible and blending ACE model is preferable.[160] The rationale for linear ACE is based on concerns regarding circular reasoning,[161] in which the examiner's interpretation of features in a latent are influenced during Comparison by "reasoning 'backward' from features visible in the [...] exemplar."[162] A notable example of the problem of bias from the exemplar resulting in circular reasoning occurred in the Madrid misidentification:[163] the initial examiner reinterpreted five of the original seven Analysis points to be more consistent with the (incorrect) exemplar. Evett and Williams[164] 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.,[165] in a discussion of interexaminer reproducibility, provide additional examples showing changes to minutiae markup made during Comparison.

In this study we assessed 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?

[157] *Derived from [A-C]*

[158] *e.g. [Mayfield11, HumanFactors12, Kassin13]*

[159] *[HumanFactors12]*

[160] *[Vanderkolk04]*

[161] *[IEEGFI04, Langenburg12a]*

[162] *[Mayfield06], p. 139*

[163] *[Mayfield06]*

[164] *[Evett96]*

[165] *[Neumann13], p. 80*

- How were changes in latent value assessments associated with changes in markup?

6.3.3a Materials and Methods

> *For a discussion of the materials and methods for the overall White Box study, see section 6.3.1 White Box Methods and Materials. This section discusses those aspects of materials and methods specific to the Analysis To Comparison study.*

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 generally report clarity results by aggregating the six EFS clarity levels specified by the examiners 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.

Our analyses of changes in value determinations are limited to a subset of 3,709 responses (out of the 3,730 total White Box responses, this omits 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).

6.3.3b Results

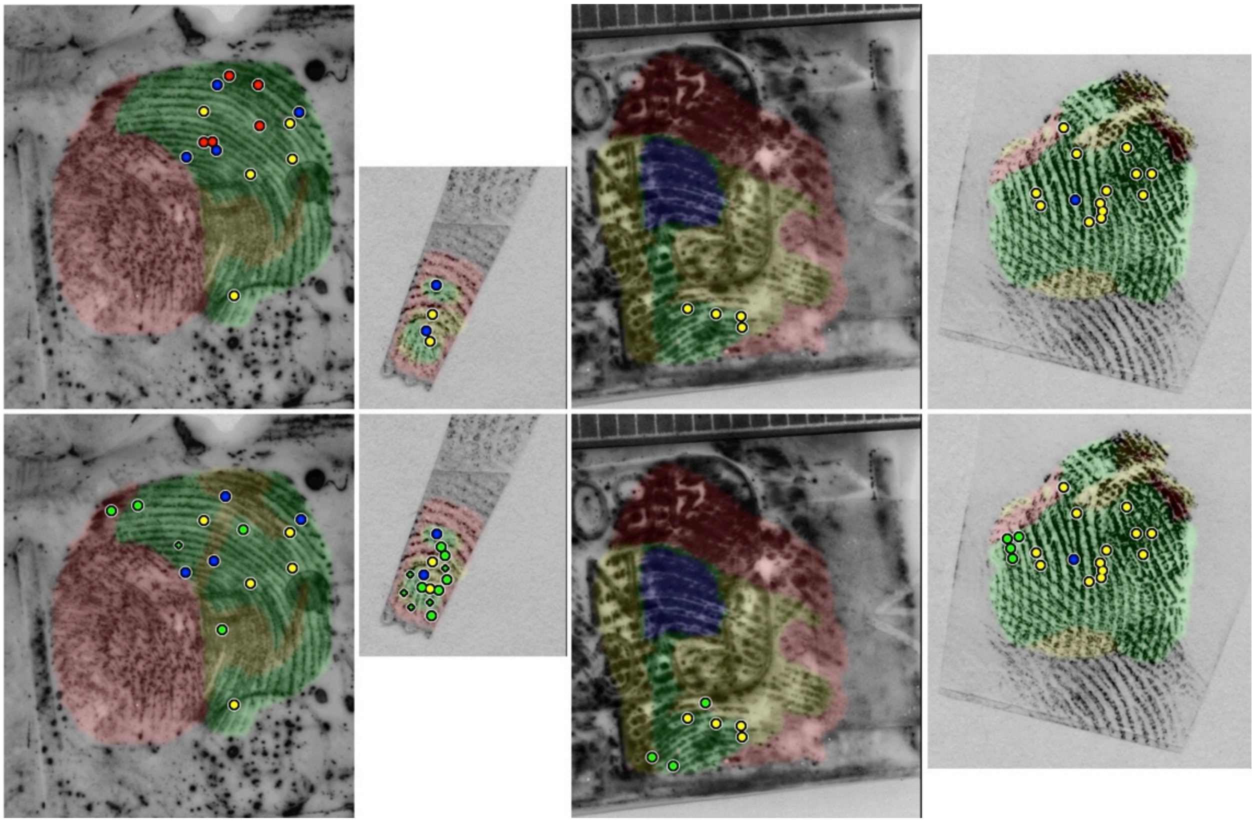


Figure 30: Example markups of four latents from Analysis (top) and from Comparison (bottom). Retained features are in yellow, moved in blue, deleted in red, and added in green. Other (non-minutiae) features are shown as crosses.

Figure 30 shows examples of changes between Analysis and Comparison. Table 5 shows an overview of the changes in markup by feature type. 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, 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 5: 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. 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 6). When examiners individualized, they almost always added or deleted minutiae (90.3% of individualizations[166]). 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. 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.[167] 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.

		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 6: 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);

[166] *93.2% of individualizations had moved, deleted, or added minutiae.*

[167] *Although distinct procedures were used to select latent fingerprints for use in mated vs. nonmated pairs, 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.*

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, examiners rated exclusions and inconclusives as substantially more difficult when markup was required (in this study) than when no markup was required (in the Black Box study).

Most minutiae were marked in Clear areas. The rates of changed minutiae were higher in low-clarity areas, especially for added minutiae (Table 7). 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. The concentration of changes in Unclear areas is even more pronounced when analyzed by the ***median*** clarity across multiple examiners: 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 7: Minutia changes by clarity.

Examiners modified 13% of the latent clarity maps during Comparison, with a higher rate of change for mated data. 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.

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.

Comparisons tended to result in a net increase in total minutiae. We see a strong subjective component to these changes. Based on a model of the change in minutia count as a response to the image pair and examiner, 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. 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. 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 31). 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. 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[168] and in the Neumann et al. study.[169] 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.

[168] *[Mayfield06]*

[169] *[Neumann13b]*

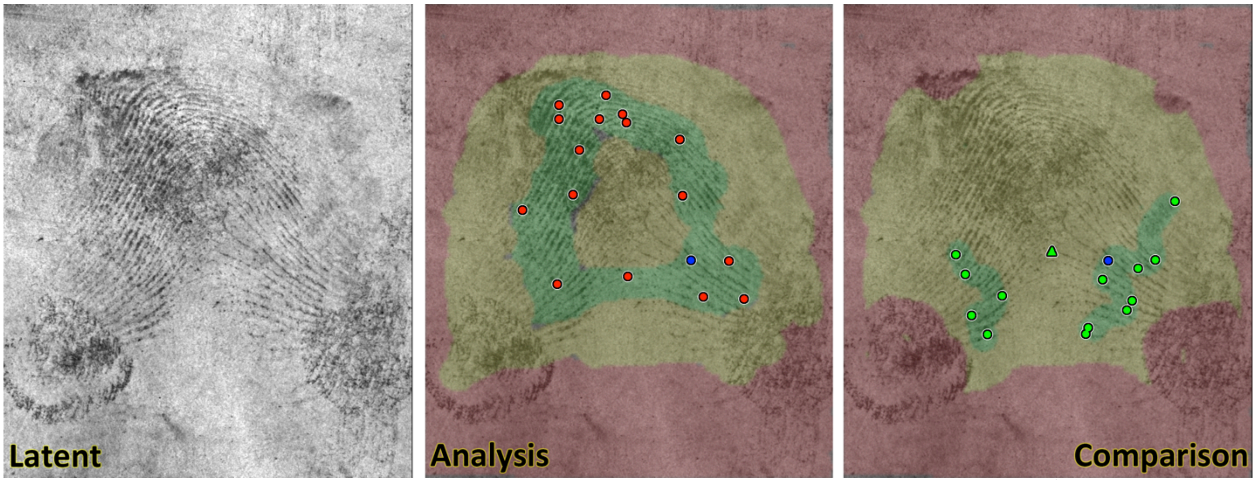


Figure 31: Image pair that resulted in the sole erroneous individualization.[170] Minutiae are shown as circles, deltas as triangles, cores as squares. 16 minutiae were deleted (red), 13 added (green), 1 moved (blue), and 0 retained; 1 delta was added. The examiner rated this comparison Easy.

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. 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.

On our Black Box test, 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.

As summarized in Table 8 and Table 9, 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 9 continue to hold when tested on that subset.

[170] *The exemplar involved in the erroneous individualization cannot be shown for privacy reasons.*

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 8: 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.

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 9: Latent value in Analysis and Comparison: percentages corresponding to Table 8. Changed value determinations are highlighted. (n=3,006)

In the Black Box study, 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 9).

6.3.3c 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 in the Sufficiency for Value study.

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[171] 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

[171] *[SWGFAST-Conclusions13]*

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 the *Sufficiency for Value* study, "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[172] 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.

6.3.4 Interexaminer Variation in Minutia Markup (White Box 3)[173]

Differences in minutia counts understate the variability among examiners: examiners' markups may have similar minutia counts but differ greatly in which specific minutiae were marked. The *Interexaminer Variation in Minutia Markup* study evaluated how the markup of minutiae varies among examiners, during both the analysis of a latent, and the comparison with an exemplar.

Why does it matter if examiners mark different minutiae? The conventional wisdom has been that it doesn't matter which features examiners use for their conclusions as long as they reach the same conclusion. However, because there is substantial interexaminer variation in determinations (as we saw in the Black Box and Black Box Repeatability studies), there is reason for scrutiny of which features examiners use. In some legal cases,[174] different conclusions among examiners have hinged on different interpretations regarding the presence or correspondence of features. Even if differences in interpretations of features do not result in differing conclusions, differences in the interpretation or markup of features underscore the subjectivity of the latent print examination process.

6.3.4a Interpretation and documentation of minutiae

Given the importance of minutiae in the examination process (as the primary feature used as the basis for almost all value and individualization determinations), documentation of minutiae warrants discussion.

The simple definition is that a minutia is the location where a ridge ends or bifurcates (forks), as shown in Figure 32.[175]

[172] *e.g. [Mayfield06, Langenburg12b, Neumann13b, Evett96, Swofford13, Haber09]*

[173] *Derived from [IEVMM]*

[174] *e.g. [Mayfield06, McKie11, Canen02, Wright97, German14, Jackson14]*

[175] *Some older definitions include dots as a third type of minutia, but terminology has shifted because dots are not readily detected by Automated Fingerprint Identification Systems (AFIS).*

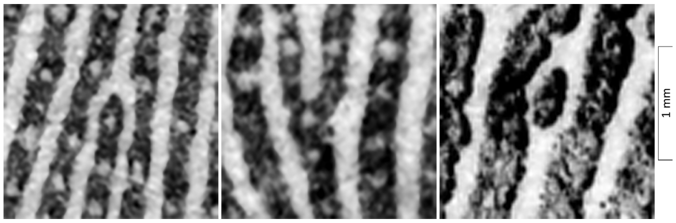


Figure 32: Examples of minutiae: (Left) ridge ending, (Middle) bifurcation, (Right) dot. Ridges are shown in black, and valleys are shown in white.

However, not all ridge features are as readily classified as those shown in Figure 32. Disagreements among examiners may be due to differences in interpretation (e.g., due to ambiguous features, low clarity, or disagreements regarding the boundaries of the region of interest), or merely to differences in how examiners document their interpretations (e.g., due to human error, or unfamiliarity with instructions and tools).

Figure 33 shows examples where examiners might disagree due to ambiguous features. Some features in friction ridge patterns are not easy to classify (even given very high clarity images) due to the shape and configurations of ridge patterns. In these instances, differences in markup may not imply actual differences in interpretation among examiners, but disagreements regarding the definition of a minutia and which features should be documented. For example, in Figure 33d, the notable “feature” is the scar, which is not readily reduced to specific point locations of ridge endings and bifurcations, and one may expect examiners’ minutia markup will vary in the area of the scar.

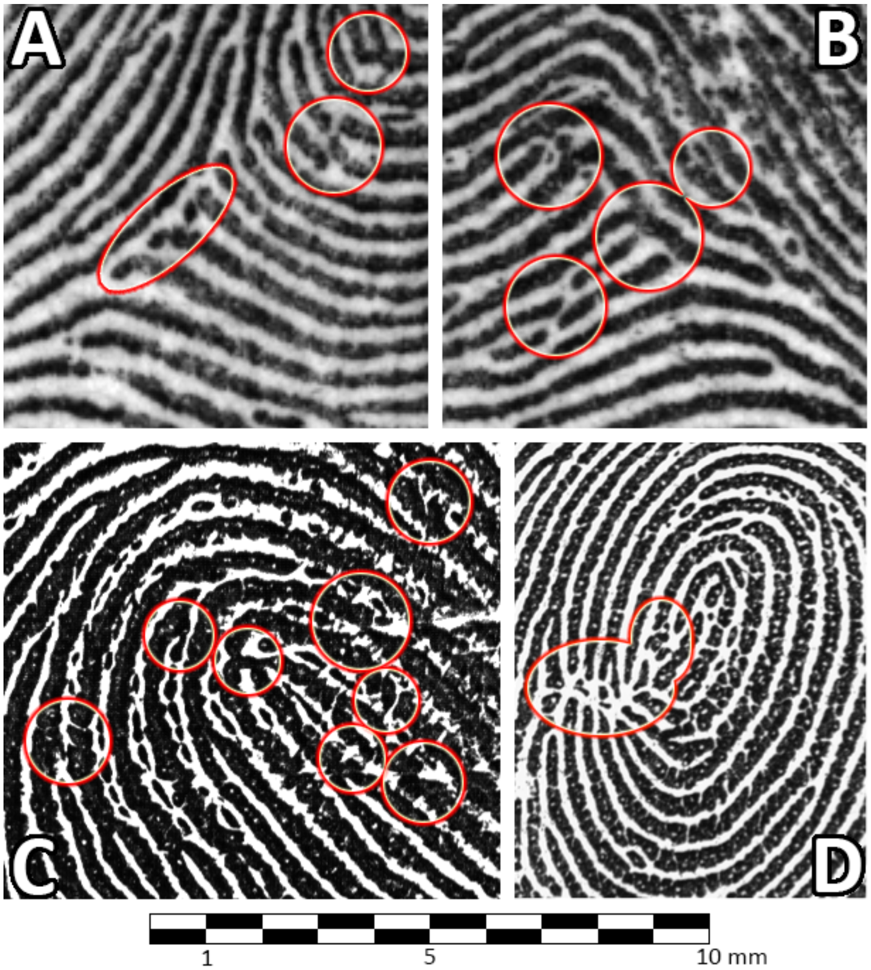


Figure 33: Examples of features that are intrinsically difficult to classify. Each of the images shows examples of ridge flow that do not fit into simple definitions of minutiae, such as unusual ridge shape, unusual interactions between ridges, interactions with incipient ridges, and interactions with a scar.

Latents are often poor quality (e.g., Figure 34), due to factors such as uncontrolled deposition (e.g., distortion, smearing, superimposed prints), substrate (surface on which the print is deposited), matrix (substance transferred to the surface), and development (physical or chemical process used to visualize the print). In practice, examiners often differ in their assessments of whether the information in an unclear area is sufficient to determine that a minutia is present, and therefore we can expect that markup in unclear areas will be less reproducible than in clear areas. Differences in reproducibility by clarity are to be expected: examiners should generally agree on minutiae in Clear areas, but may or may not agree in areas they consider Unclear.

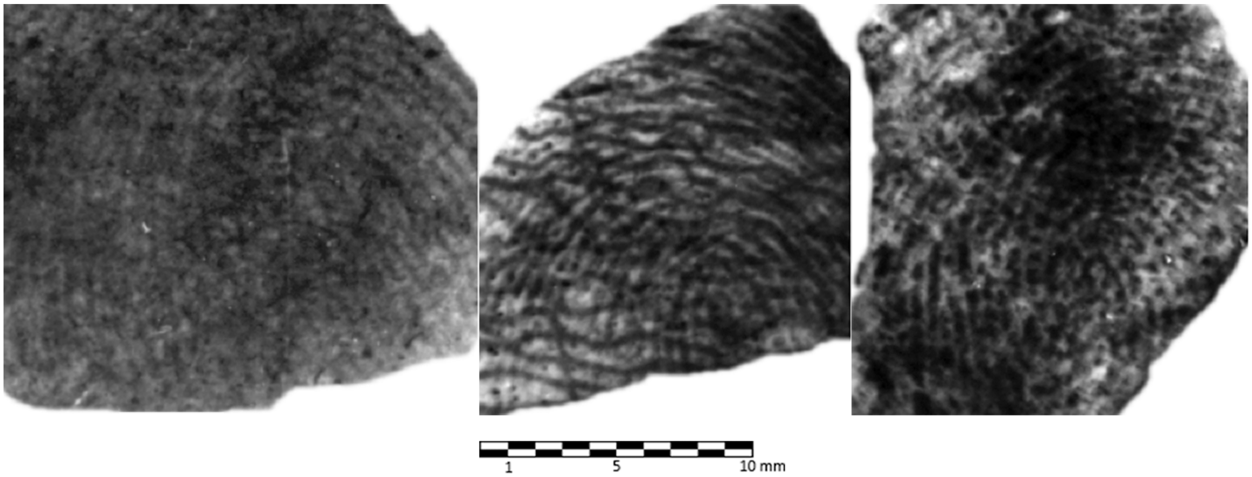


Figure 34: Low-clarity examples where the ***presence or absence*** of minutiae is ambiguous.

Even when examiners agree that a minutia is present and should be marked, clarity may affect their assessments of the exact locations and types of minutiae (e.g., Figure 35).

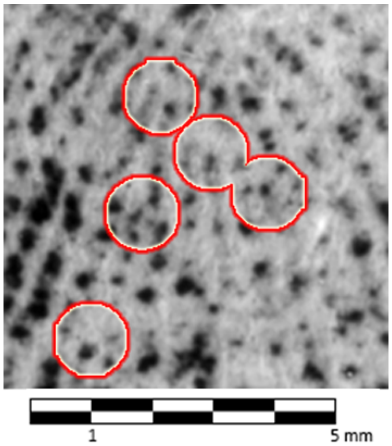


Figure 35: Ambiguous minutia ***locations and types***. Each circle indicates an area where three ridges converge to two ridges, so a minutia must be present, but cannot be located precisely, and the type (whether it is a ridge ending or bifurcation) is ambiguous.

Another source of disagreement in minutia markup stems from disagreements regarding the boundaries of the impression being considered. Generally, examiners are looking to compare a single contiguous impression, in which they can assess the relative positions and topological relationships of minutiae and other features. However, it is not apparent whether some images (e.g., Figure 36) contain a single impression or multiple superimposed impressions, and therefore examiners may disagree on whether specific minutiae are part of the impression of interest. Some of the disagreements regarding minutiae in the Madrid misidentification[176] were based on differing assessments of whether the image contained a single impression, a double touch (partially superimposed impressions from the same finger), or impressions from two fingers. A similar

[176] *[Mayfield06]*

situation occurs even in clear impressions when examiners may differ in whether to consider the area below the crease (i.e., in the medial segment of the finger) as the same impression.

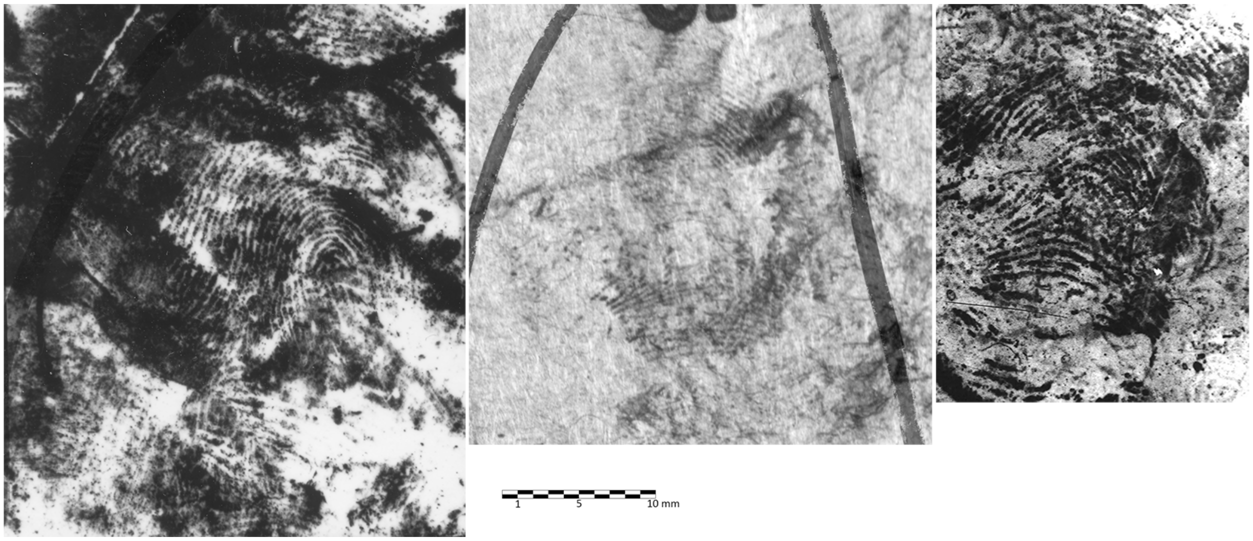


Figure 36: Problematic examples where the ***area*** to be marked is ambiguous because it is ambiguous which areas are from a single continuous impression. The example on the right is the latent from the Madrid misidentification.[177]

Some of the variation in markup can be attributed to a lack of clear criteria specifying when and how to mark minutiae, and to a lack of standardization. While 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 the examination process, the details of how to document minutiae are mostly unspecified. Because documentation of minutiae is not standardized in practice, it is difficult to ascertain the extent to which variation among examiners can be attributed to actual differences in interpretation, as opposed to differences in how examiners choose to document their work. Few agencies train examiners specifically on how to interpret, select, and record minutiae in a standard, reproducible manner, other than for AFIS searches, which generally require following proprietary rules. Those agencies that do require markup vary substantially on how that markup is effected, including pinpricks in physical photographs, color-coding approaches [GYRO], software-based solutions specific to fingerprints [ULW, Mideo Latentworks®, PiAnOS], and generic image processing software. Several authors[178] have stressed the need for standardization of minutia markup. In this study we use the Extended Feature Set (EFS) format as defined in [ANSI/NIST] and supporting guidelines for examiners.[179] 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 of the various factors we have discussed that may result in interexaminer variation in minutia interpretation or markup, there is currently no means of defining a definitive minutia markup for any given latent: either in tests or in operational casework, we can compare examiners' markups against each other, or against a group consensus, but cannot judge whether or not they are correct in an absolute sense.

[177] *[Mayfield06]; image from [German14]*

[178] *[Neumann13b, Dror11a, Langenburg12a, Swofford13, SuffID, A-C]*

[179] *[EFSMI]*

As discussed in Section 3.5.4, EFS provides multiple means for an examiner (or automated feature extractor) to define the attributes of minutiae, and (at least as important) the uncertainty of these assessments:

- Minutia type is defined as bifurcation, ridge ending, or unknown. Complex types are marked as combinations of these types.
- Minutiae location is an (x,y) coordinate, with a radius of uncertainty when the precise location cannot be determined. There are a variety of circumstances in which the radius of uncertainty should be defined, such as low clarity, when the minutia type cannot be determined, or when a ridge ending tapers into a long incipient.
- The minutia direction (theta) provides for angle uncertainty, which is appropriate when ridges curve strongly, for minutiae on very short ridges, and for "delta-type" minutiae in which three angles are approximately equal.
- Uncertainty regarding the presence or absence of minutiae is indicated using the clarity map: green or blue areas indicate high confidence in all marked minutiae, and confidence that there are no unmarked minutiae in the area; yellow areas indicate that marked minutiae are debatable and unmarked minutiae may exist.

EFS does not define or suggest symbols for minutiae. In 2013 SWGFAST issued a position statement encouraging AFIS manufacturers to standardize symbols.[180]

PiAnoS[181] uses an approach different from EFS, defining symbols for ridge endings, bifurcations, and unknown type, and an additional symbol for unknown location.

[180] *"SWGFAST encourages all manufacturers of AFIS technology to implement the use of standardized feature symbols. As part of the National Institute of Standards and Technology's Law Enforcement Standards Office's work to improve latent AFIS interoperability, one issue that has been raised is the variation in how fingerprint features are displayed among AFIS products. Having a standard set of symbols would benefit examiners utilizing multiple AFIS systems on a regular basis. Additionally, standardized feature symbols would aid in user acceptance and training from product to product." [SWGFAST-Symbols13]*

[181] *[PiAnoS15]*

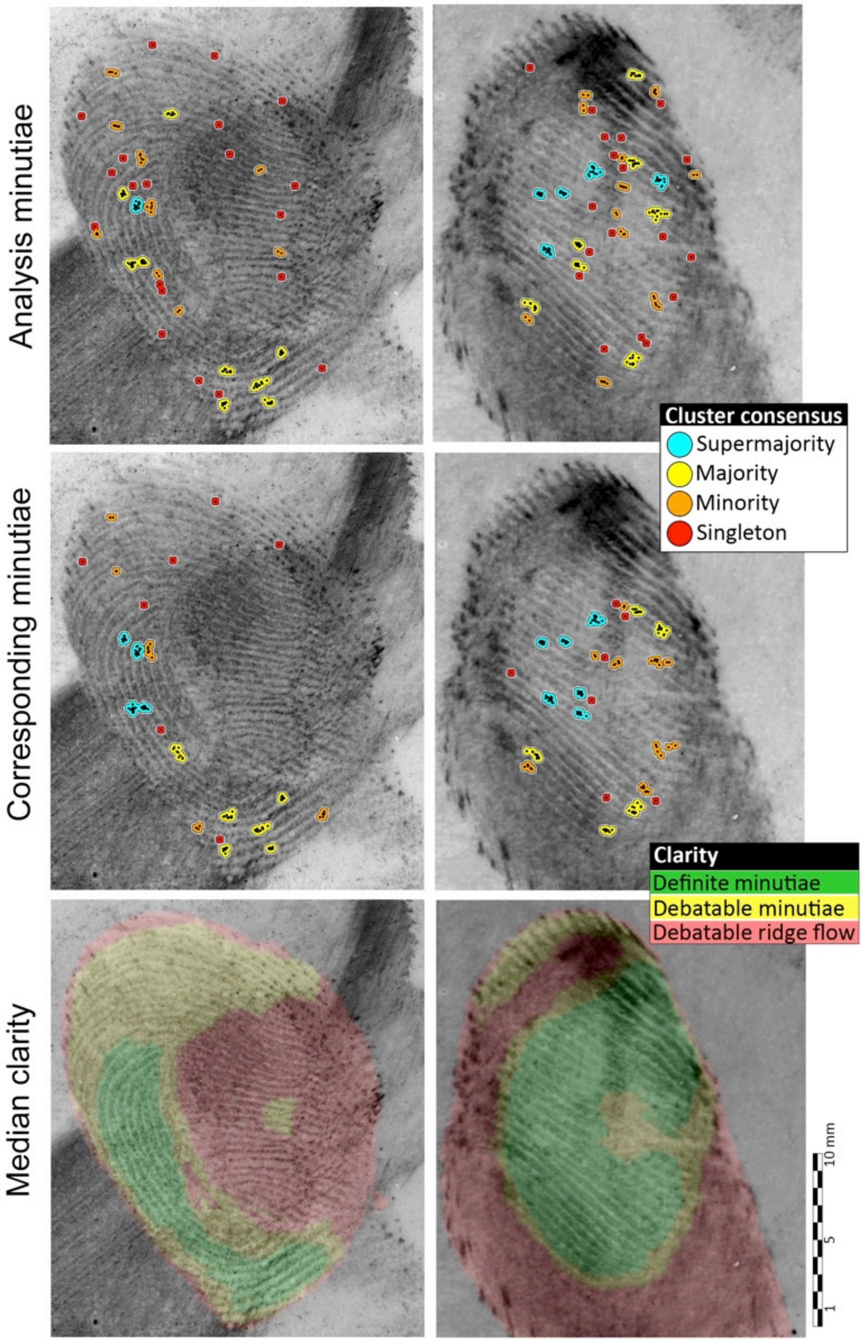


Figure 37: Two examples of interexaminer variation in minutia markup. Marked minutiae are shown as small black dots inside color-coded clusters. Row 1: Analysis phase; cluster colors indicate the proportion of examiners who marked within that cluster. Row 2: Comparison phase; cluster colors indicate the proportion of comparing examiners who corresponded the minutia; only those minutiae marked as corresponding are shown. Row 3: Analysis phase; median clarity map, which combines clarity responses from all examiners.

6.3.4b Clustering

Examiners' markups differed in whether or not individual minutiae were marked, and in the precise location where the minutiae were marked. In order to focus on whether examiners agree on the presence or absence of minutiae, we need to see past minor variations in minutia location. Neumann et al.[182] used ellipses to determine whether two minutiae should be considered the same, based on an expectation of more variation in location along the direction of the ridge than perpendicular to ridge flow; here we did not collect minutia direction, making this approach impractical. In [A-C], our technique of classifying features as retained, moved, added or deleted was based on a fixed radius of 0.5 mm (0.02 inch, or approximately the average inter-ridge distance) — although that approach was satisfactory for two markups where one was derived from the other, it is not well suited to comparing more than two markups.

In IEVMM, we used a commonly-used data clustering algorithm, Density-Based Spatial Clustering of Applications with Noise (DBSCAN),[183] to classify minutiae marked by multiple examiners as representing the same minutia on the latent. The DBSCAN algorithm was parameterized with a reachability distance of 0.38mm (0.015 inch);[184] any marked minutiae within this distance of one another coalesce into a cluster (a cluster starts with an arbitrary marked minutia, grows to include any other marked minutiae (from all examiners) within that distance, and then iteratively grows to include any other marked minutiae within that distance of the cluster). Any "overgrown" clusters were split using agglomerative hierarchical clustering to produce the final set of clusters for analysis.

6.3.4c Measuring interexaminer variation

Although "minutia" in theory refers to an actual feature on the skin, we have no special knowledge of the actual features beyond what we can learn from what was marked by examiners. To avoid ambiguity in what we are measuring, we define two terms:

- The annotation by an individual examiner at some location on the latent (***marked minutia***);
- A set of marked minutiae from multiple examiners that were grouped into the same cluster (***cluster***).

Our Analysis-phase results are based on 44,941 marked minutiae, which resulted in 10,324 clusters. We say that two examiners have marked the same minutia if both examiners marked within the same cluster. We define two closely related measures of interexaminer variation:

- For each marked minutia, we use the term ***reproducibility*** to refer to the proportion of other examiners who marked that minutia (i.e., marked within the same cluster).
- For each cluster, we use the term ***consensus*** to refer to the proportion of examiners who marked a minutia in that cluster.

6.3.4d Reproducibility of Analysis-phase minutiae

Overall, the probability of randomly selected minutiae being reproduced (mean reproducibility) was 63%. However, as shown in Figure 38, clarity is a major determinant of whether examiners mark the same minutiae: reproducibility is lower in areas the examiner marked as unclear (47% mean reproducibility), and higher in areas marked as clear (70% mean reproducibility). Unclear minutiae were much less likely to be unanimously reproduced than clear (9% of unclear minutiae, 26% of clear), and much more likely to be singletons (17% of unclear, 7% of clear minutiae).

[182] *[Neumann13b]*

[183] *[Ester96]*

[184] *The distance between ridges varies within an impression and between subjects, but average peak-to-peak distances are reported as varying between 0.43mm and 0.56mm ([Ashbaugh99], discussed in [AssessingLC]).*

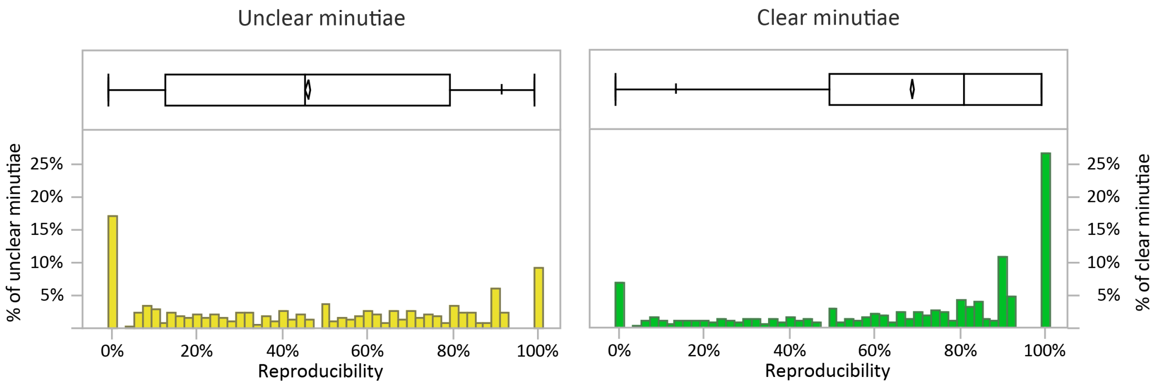


Figure 38: Reproducibility of Analysis-phase marked minutiae, by examiner clarity. The mean reproducibility was 63% (47% for unclear minutiae, 70% clear); median reproducibility was 75% (46% for unclear minutiae, 82% clear); 66% of minutiae were reproduced by the majority of other examiners, i.e., greater than 50% reproducibility (46% unclear, 73% clear). (n=44,941 minutiae: 12,782 unclear, 32,159 clear)

Figure 39 contrasts the two ways of measuring interexaminer variability: the ***reproducibility*** of marked minutiae (i.e., the 44,941 marked minutiae), and the extent of ***consensus*** among examiners that a minutia is present at a given location (i.e., the 10,324 minutia clusters). By counting each marked minutia equally, reproducibility gives more weight to minutiae marked by many examiners; consensus gives equal weight to each cluster regardless of how many examiners marked that minutia. A singleton is counted once in either case. As a result, the mean reproducibility (63%) is higher than the mean consensus (36%). Most of the marked minutiae (68%) were reproduced by a majority of other examiners, but most of the clusters (coincidentally 68%) were marked by a minority of examiners.

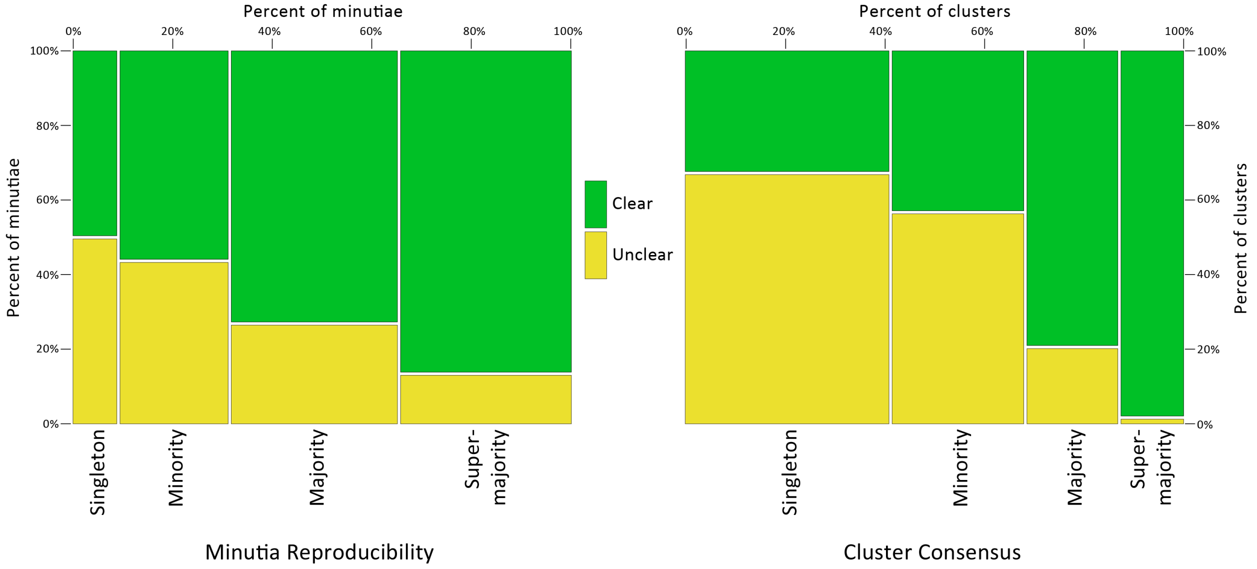


Figure 39: Mosaic plots showing the associations between clarity and interexaminer variability in minutia markup. (Left) minutia reproducibility by examiner clarity (n=44,941 minutiae); (Right) cluster consensus by median clarity (n=10,324 clusters). For example, there were 4269 singletons, accounting for 9% of marked minutiae and 41% of clusters.

The fact that an examiner marked a minutia, regardless of how that examiner marked clarity, indicates a high probability that a majority of examiners described the area as clear: even when examiners marked minutiae as unclear, on average about half of other examiners marked that area as clear. While marking a minutia as unclear effectively signaled low reproducibility, a voted description of clarity (median clarity map) provided an even better explanation of reproducibility. For example, 67% of the singletons were in median unclear areas, yet only 50% were marked as unclear by the examiner who marked the singleton; 98% of supermajorities were in median clear areas, but only 86% of those minutiae were marked as clear. Previously, we reported a similar result: median clarity predicted changes in minutia markup between Analysis and Comparison better than examiner clarity.[185] In general, we found that median clarity markups conform well to our expectations of proper and careful characterizations of latent clarity, by reducing the impact of outliers and imprecision found in the individual examiners' clarity markups. The (unexpected) result that median clarity was a better predictor of changes and reproducibility than examiner clarity suggests that greater consistency among examiners in describing clarity would make clarity markup more effective in flagging unreliable minutiae, and has the potential to make substantive disagreements among examiners more readily apparent.

There were many areas in the latents where there was no strong consensus among examiners on whether an area was clear or unclear; we refer to these areas as having "debatable clarity." Individual examiners were presumably uncertain how to mark clarity in some of these areas, but the test forced a choice between clear and unclear. Figure 40 indicates how these areas of debatable clarity contribute to our results. As the proportion of examiners describing an area as clear increased, both the number of minutiae marked and minutia reproducibility increased. Supermajorities sometimes occurred in areas where examiners did not agree on clarity (e.g., 20-80% voted clear). Even in areas that examiners agreed (90-100%) are clear, reproducibility was not unanimous: on review, the lack of unanimity usually could be attributed to adequate-but-difficult clarity, complex ridge flow, unclustered minutiae due to differences in location, or marking of features that were only debatably minutiae. Although reproducibility was lowest in areas that a large majority of examiners described as unclear, relatively few minutiae were marked in those areas: much of the lack of reproducibility therefore arose in areas of debatable clarity (e.g., 20-80% voted clear). This voted measure of clarity provided a more complete explanation of the relationship between clarity and reproducibility than the median clarity maps, which in turn provided a more complete explanation than the individual examiner clarity maps.

[185] *[A-C]*

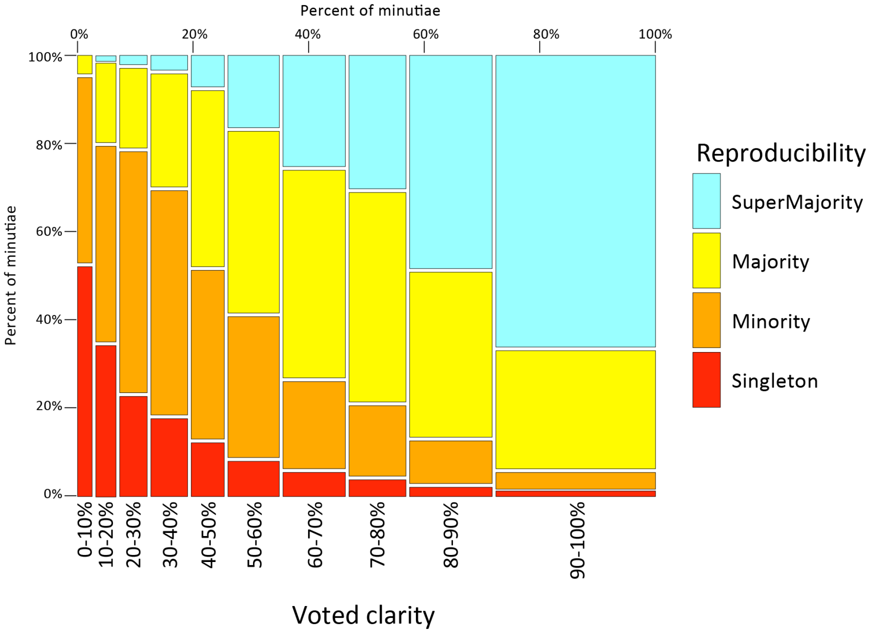


Figure 40: Voted clarity by reproducibility (n=44,941 minutiae). Voted clarity describes the percentage of examiners who described the location of that minutia as clear. 74% of minutiae were marked in areas described as clear by at least half of examiners.

One explanation for some lack of reproducibility is that examiners do not always agree on the region of interest. Additionally, examiners sometimes differ in whether they choose to mark minutiae in low-clarity areas. As a consequence, examiners often marked minutiae physically far away from those marked by other examiners. To quantify this effect, we measured the distance from each marked minutia to the nearest majority cluster. We can (somewhat arbitrarily) consider that a marked minutia is "relatively far" from a minutiae cluster if they are at least 2.5mm (0.1") apart; this would be about 5 ridge intervals on average. Similarly, marked minutiae are "very far" apart if they are at least 5.1mm (0.2"), about 10 ridges, apart. By that measure, 11.2% of marked minutiae are relatively far from the center of the nearest majority cluster (3.2% of median clear minutiae and 35.9% of median unclear minutiae); 3.5% of marked minutiae are very far from the nearest majority cluster (0.5% of median clear minutiae and 12.9% of median unclear minutiae). Disagreements among examiners regarding the regions in which to mark minutiae account for a substantial proportion of interexaminer variability, especially in unclear areas.

Another possible explanation for lack of reproducibility that we discussed in the Introduction is the potential ambiguity of whether a feature should be considered a minutia or as a nonminutia feature, such as a dot or an event on an incipient ridge. Examiners were instructed to mark "other" (nonminutia) features when they were used as the basis for a Comparison determination; marking during Analysis was optional. For this reason, markup of nonminutia features was incomplete in both phases, limiting our ability to measure disagreements on feature type. On review of the markups, singletons were often marked on incipient ridges, dots, or on nonminutia features in cores or deltas. In the Comparison phase, features other than minutiae were present in the area of only 4.5% of minutia clusters on the latents; not all of these represent potential disagreements regarding the type of the feature.

In addition to assessing interexaminer variability by marked minutiae (reproducibility) and by clusters (consensus), we can assess variability by entire markups. Based on the idea that examiners should agree on minutiae in clear areas and differences regarding unclear minutiae should be acceptable, we could define markups as being in "perfect" agreement with the majority when they satisfy two conditions: all minutiae marked by that examiner in clear areas are in majority clusters, and that examiner marked a minutia in each of the majority clusters (in any clarity). By that measure, 15% of the 3730 Analysis-phase markups of latents were in perfect agreement (including 9% with no clear minutiae or no majority clusters). If we loosen the requirements to "75% agreement" (the examiner marked at least 75% of the majority clusters, and at least 75% of the minutia that the examiner marked in clear areas coincided with majority clusters), 52% of markups were in agreement.

6.3.4e Reproducibility of Analysis-Comparison changes

In the *Analysis to Comparison* study,[186] we reported that changes in markup were most prevalent on individualizations (minutiae were added or deleted on 90.3% of individualizations); for inconclusive and exclusion determinations, changes were more prevalent when the image pair was mated; a greater percentage of minutiae were deleted or added in unclear areas than in clear areas. Here, we see that the net effect of these changes was a small increase in minutia reproducibility on latents that were compared to mated exemplars; no net change in reproducibility was detected among the nonmate comparisons.

Deleted and added minutiae were each associated with low reproducibility. Examiners were more likely to delete minutiae that were marked by a minority of other examiners. Interestingly, the minutiae that they added (even those in clear areas) also tended to be marked by a minority of other examiners: this might be due in part to a motivation to thoroughly document individualization conclusions. These effects were particularly pronounced for singletons (e.g., among latents that were compared, 23% of singletons were deleted). The association of deleted and added minutiae with low reproducibility does not simply reflect higher volatility in unclear areas: a strong inverse association between changes and reproducibility remains after controlling for clarity. In other words, proportionally more minutiae were deleted and added in unclear areas than in clear areas and, after accounting for clarity, those minutiae with low reproducibility were more likely to be deleted or added than those with high reproducibility.

6.3.4f Reproducibility of corresponding minutiae from the Comparison phase[187]

Comparisons between a latent and an exemplar introduce another dimension of interexaminer variation in minutia markup: the examiners may differ not only on whether they mark a given minutia in the latent, but also on whether those minutiae that they agree are present in the latent correspond to the exemplar. Interpreting interexaminer variability in marking minutia correspondences is complicated by the fact that marking of correspondences is strongly associated with determinations: comparison markup is only available from those examiners who agreed that the latent is suitable for comparison, and examiners who individualize tend to mark more corresponding minutiae than those who exclude or are inconclusive.[188] For these reasons, we describe interexaminer variability for Comparison-phase results slightly differently than for Analysis-phase results, as shown in Table 10.

Figure 41 shows examples of interexaminer differences in annotations of corresponding minutiae, suggesting how some of the differences among examiners arise: examiners B, C, and E marked the features in a generally similar manner but differed on specific points (especially within the delta) and the extent of the areas they used in Comparison; examiner C changed value determination from

[186] *[A-C]*

[187] *Derived from [IEVMM] and [SuffID]*

[188] *[SuffID]*

VEO to VID during Comparison; examiner D individualized with only four corresponding minutiae but did not mark the delta or any of the features within the delta (improper annotation); examiner F misinterpreted the orientation, resulting in an erroneous exclusion.

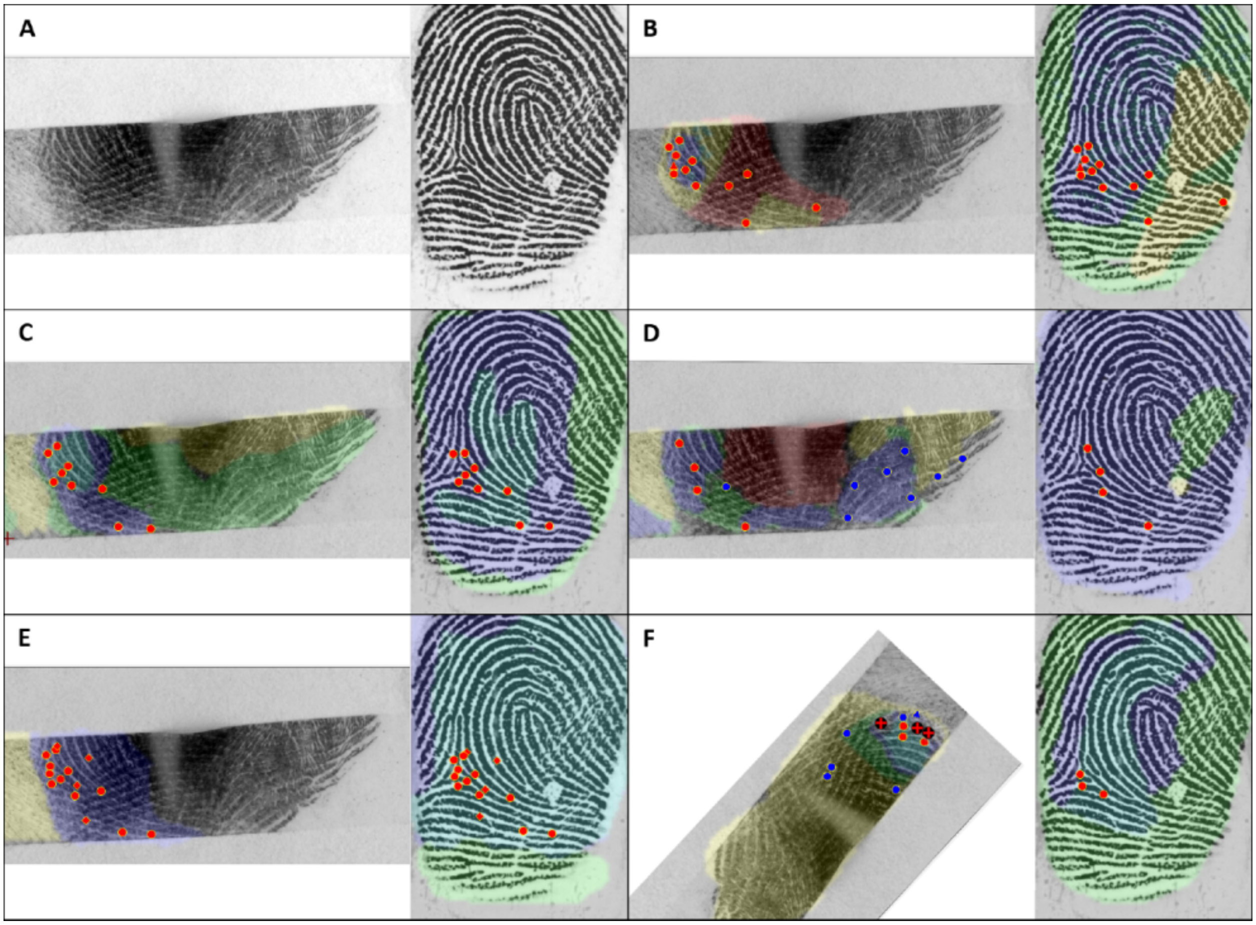


Figure 41: Example of a mated image pair (A), showing variations in annotation among five examiners (B-F). Corresponding points are shown here in red, unassociated in blue; minutiae as circles, deltas as triangles, other points as rhombuses; non-corresponding points as red Xs. Examiners B-E individualized; F excluded. Determinations by the 11 examiners assigned this image pair: 2 NV, 3 VEO (2 of which were changed to VID during Comparison), 6 VID; 1 inconclusive, 1 exclusion, 7 individualization.

Table 10 describes the reproducibility of marked minutiae in the Comparison phase, categorized by whether the examiners corresponded the minutiae. For each examiner ("Examiner A") the probability that a second examiner ("Examiner B") marked and corresponded a minutia was measured by considering all other examiners, regardless of whether the other examiners compared the latent. On average, if an examiner marked a minutia on the latent and corresponded that minutia to the exemplar, the probability that a second examiner also marked and corresponded that minutia was 69% for clear minutiae and 47% for unclear. When two examiners ***both*** individualized, that probability increased to 76% for clear and 57% for unclear. Examiners marked few correspondences on nonmated pairs: the probability that a second examiner reproduced a correspondence on a nonmated pair was 8% regardless of clarity.

Clarity accounts for much of the difference in whether the second examiner marked the minutia, but little of the difference in whether the second examiner corresponded a marked minutia, as shown in the right column of Table 10. In cases where two examiners agreed that a minutia was present on the latent and one examiner corresponded the minutia, the probability that the second examiner would also correspond the minutia was approximately the same for clear and unclear minutiae (88% vs. 84%).

The probability of examiners corresponding marked minutiae was correlated with the reproducibility of those minutiae. On individualizations, examiners corresponded 60% of their singletons and 92% of minutiae that were unanimously marked by comparing examiners; when examiners did not individualize, they corresponded 10% of their singletons and 25% of minutiae unanimously marked by comparing examiners. Note that because the latent and exemplar do not always completely overlap, not all minutiae in the latent can be corresponded with a given exemplar.

				Examiner B				Marked and compared minutiae that were corresponded
				Did not mark	Marked			
					Did not compare (NV)	Compared		
All Minutiae			Minutiae			Unassoc.	Corresp.	
Examiner A	Clear minutiae	Unassociated	14,744	36%	5%	44%	15%	26%
		Corresponding	20,470	20%	2%	10%	69%	88%
	Unclear minutiae	Unassociated	8221	59%	6%	25%	11%	30%
		Corresponding	7459	42%	2%	9%	47%	84%

Examiner A individualized

Examiner A	Clear minutiae	Unassociated	5507	41%	1%	39%	20%	34%
		Corresponding	18,823	20%	1%	9%	70%	89%
	Unclear minutiae	Unassociated	2600	66%	1%	20%	14%	41%
		Corresponding	6576	42%	1%	8%	49%	86%

Table 10: When examiner A marked a minutia, what examiner B did at that location, for all minutiae marked during Analysis (including deletions) or added during Comparison (n=50,894 minutiae, 3618 responses), and conditioned on examiner A having individualized (n=33,506 minutiae, 1654 responses). "Unassociated" includes all marked minutiae that were not corresponded. Percentages calculated as weighted sums over all other examiners who marked each latent, such that each minutia marked by examiner A is weighted equally. "Marked and compared minutiae that were corresponded" is the probability that examiner B corresponded a minutia given that examiner B marked that minutia and compared the latent to the exemplar.

Review of the markup provided another explanation for variation in minutia markup. The specific locations at which minutiae were marked often vary substantially among examiners. Marked minutiae in separate clusters on the latent were often corresponded to a single cluster on the exemplar: multiple examiners agreed that the minutia was present and agreed on the location in the exemplar, but differed substantially in where they marked the minutia on the latent. In order to better understand the extent of this issue, we clustered the minutiae marked on the exemplars, so that we could see how these exemplar clusters corresponded to latent clusters. Considering only those clusters in which corresponding minutiae were marked, there were 6% fewer clusters on the exemplars than on the latents. However, this effect was observed in both directions: 15% of exemplar clusters were corresponded to more than one latent cluster; 9% of latent clusters were corresponded to more than one exemplar cluster. Although some of these clustering issues might have been resolved with a different clustering algorithm, often the distance was large enough that we would not expect any clustering algorithm to group them.

6.3.4g Discussion

We identified several factors that affect minutia reproducibility: clarity, region of interest, feature type, and location. The fact that an examiner marked a minutia, regardless of how that examiner marked clarity, indicates a high probability that a majority of examiners described the area as clear. Marking a minutia as unclear was a good predictor that reproducibility would be low: in effect, by marking minutiae as unclear, examiners seem to anticipate low reproducibility. Differences in markup were most prevalent in areas where examiners did not agree on clarity, in part because relatively few minutiae were marked in areas that examiners agreed were unclear. Much of the variability, especially in unclear areas, can be attributed to differences in which areas of the prints examiners chose to mark: 36% of minutiae marked in median unclear areas during Analysis were relatively far away from the nearest majority cluster (at least 0.1 inch or approximately five or more ridge intervals). Some variability can be attributed to disagreements regarding minutia type: singletons were often marked on incipient ridges, dots, or on nonminutia features in cores or deltas. Additionally, some of the reported variability can be attributed to uncertainty in the precise location at which to mark a minutia on the latent: marked minutiae that were singletons or in separate clusters in the latent were often corresponded to a single location in the exemplar.

Some of the reported variability can be attributed to our measurement techniques, including the clustering algorithm, fingerprint selection, and markup procedures. Clustering was sensitive to our choice of radius, and did not account for factors such as local ridge width and direction. The fingerprints were selected to test the boundaries of sufficiency for individualization determinations, deliberately limiting the proportion of image pairs on which we expected unanimous determinations. Because requirements and procedures for markup are not standardized in practice, the tools and procedures we used were novel to the participants, contributing to the variability.

In the Black Box study, we found that much of the lack of (interexaminer) reproducibility of value and comparison determinations was associated with images and image pairs on which we also observed low (intraexaminer) repeatability. We assume there is a similar association between reproducibility and repeatability of minutia markup, based on our results in the *Sufficiency for ID* study[189] in which we saw a notable lack of repeatability in minutia markup (on a small sample of markups).

In our previous work,[190] we found that the association between examiners' minutia counts and their determinations was ***not*** notably affected by minutia clarity. Here, however, we see that clarity has a notable effect on the reproducibility of marked minutiae. Thus, while the total minutia count (clear and unclear minutiae) is indicative of examiners' determinations, most of the variance accounting for examiner differences in marked minutiae arises in unclear areas: when examiners individualized (or assessed a latent to be VID) those examiners generally marked more minutiae in unclear areas than examiners whose comparison determinations were inconclusive (or who assessed the latent to be NV).

We should not assume that reducing variability in markup would necessarily improve reproducibility of determinations. There are some indications that the relationship between markup and determinations may not be a simple forward causality: we have previously reported that examiner determinations appear to influence markup, as evidenced by the tendency of examiners to modify their latent markup more extensively when individualizing than when inconclusive,[191] and by a tendency not to mark just fewer than the minimum number of minutiae

[189] *[SuffID]*

[190] *[SuffValue, SuffID]*

[191] *[A-C]*

typically associated with individualization determinations.[192] It is possible that some of the variability in markup relates to processes motivated by the determination, such as reviewing unclear and peripheral areas to double-check one's work and document that nothing calls the conclusion into doubt.

There is not currently any method of defining a "correct" minutia markup for any given latent. An examiner's decision of whether a minutia is present in an unclear location is analogous to an examiner's decision as to whether the similarity of two prints is sufficient to make an individualization determination: in either case, the best information we have to evaluate the appropriateness of examiners' decisions is the collective judgment of other experts. Our method of clustering minutiae could be used to develop training sets in which an "ideal" markup would be based on a group consensus.

Differences in minutia markup are not always due to differences in interpretation, but often may be due merely to differences in how examiners document their interpretations. Examiners' clarity markup is a useful indicator of the reproducibility of the minutiae they marked, which suggests that greater consistency among examiners in describing clarity has the potential to appreciably limit the apparent disagreements among examiners in the interpretation of latent prints. We expect that standardizing markup of features and clarity (through formal specification, inclusion in training, and broad usage in operational casework) would facilitate greater transparency by making markup a more reliable description of examiners' interpretations.

6.4 Comparing evaluations of latent print examiners[193]

The publication by Ralph and Lyn Haber in *Science and Justice*, "Experimental results of fingerprint comparison validity and reliability: A review and critical analysis"[194] purports to offer a critique of 13 empirical studies of the performance of latent fingerprint examiners.[195] However, it is incumbent on the authors of such metaanalyses to make sure that they understand and accurately describe the purposes, design, and procedures of the constituent studies. Unfortunately, the Habers did not represent the studies accurately, and did not appear to have a complete understanding of the latent print examination discipline, or the statistics that they attempt to use in their criticisms. Haber and Haber's criticisms are often misdirected, faulting individual research studies for perceived shortcomings of current policy and practice, and limitations of the current body of research literature in its entirety. The Haber and Haber paper contains numerous factual errors. We documented dozens of errors in their summary of our results, which are detailed in a 16-page table in [HaberAnnex]. In addition to these errors, their criticisms are founded on numerous apparent misrepresentations and misunderstandings. We published a detailed response because we felt it was critical to have a published response correcting the Habers' faulty assertions, particularly because those faulty assertions might be used to support court testimony.

It is reasonable to perform metaanalyses, describing the commonalities and differences of a variety of studies, and discussing the strengths and weaknesses of each. However, the results of studies conducted for different purposes and with different procedures should be expected to differ. Many of the Habers' criticisms can be summarized as their dissatisfaction that the other researchers did not conduct their studies as the Habers would have wished. Indeed, the Habers argue (no, "demand") that they have preconditions for any such studies: "Three of the problems noted

[192] *[SuffID]*

[193] *New material and text derived from [HaberResponse]*

[194] *[Haber14a]*

[195] *[BB, BBRR, Langenburg12b , Evett96, Langenburg09a, FBI99, Dror06a, Dror06b, Hall08, Langenburg09b, Wertheim06, Gutowski06, Tangen11]*

throughout our critiques demand valid solutions before useful research can be performed to document the accuracy of fingerprint comparisons."[196] I concur with Langenburg et al:

> *It is unfortunate that their view is so narrow and do not emphasize the greater understanding that these experiments have brought the community. The studies have highlighted important areas for improvement in the process, such as recognizing the disparity between exclusion decisions and identification decisions, recognizing the importance of documentation and care during the analysis phase of complex latent prints, and demonstration of expertise. We are left with the view that HH's criticisms are a result of some partisan agenda as they offer no truly helpful insight, have over-weighted criticisms against these study designs, and do not offer any data to show that their suggested processes are superior to the processes employed by the researchers that they criticize.*[197]

I also agree with Thompson and Tangen: "A tit for tat exposition of 'design flaws' – legitimate methodological and statistical flaws and limitations notwithstanding – and outright rejection of studies is unlikely to be a fruitful approach toward our presumably shared goal of continuous improvement of forensic science systems."[198] Although the Habers did respond to the three letters,[199] their response generally repeated the same arguments from their initial paper, with no indication that they actually understood why their original paper was so problematic.

[196] *[Haber14a], Section 11*

[197] *[Langenburg14], p 395*

[198] *[Thompson14]*

[199] *[Haber14b]*

Chapter 7 Interacting with AFIS

From the outset, EFS was designed to be a standard language for examiners to communicate with AFIS systems. EFS is a vendor-neutral interchange format for automated fingerprint or palmprint systems. Previous proprietary formats instructed examiners to respond to the specifics of AFIS algorithms, which led examiners to decide whether and how to mark features based on their understanding of their system's algorithm, rather than just marking the features that they see. EFS changes this relationship, so that the engineers would react to examiners' needs, not the other way around.

Traditionally, feature markup was one-directional, from the examiner to the AFIS. EFS also provides a standard method for communication of features in the other direction, so that the AFIS can show the examiner the features in correspondence between a latent and each print in the candidate list. The FBI'S NGI system has implemented this functionality. Showing such features should lessen the possibility of a missed ID, particularly on palm latents or latents without cores, deltas, or obvious anchoring points. A potential drawback could be that the examiner could be biased by the features marked by the AFIS. However, the AFIS returns such "corresponding" features for all candidates, even though at most candidate can be a mate: since the majority of AFIS-generated "corresponding" features are by definition incorrect, examiners should be able to treat the correspondences in the same way that they treat candidates, as a possibility, most of which are incorrect.

7.1 Latent AFIS interoperability

In November 2003, Senator Edward Kennedy wrote a letter to the *Washington Post* decrying inadequate latent AFIS interoperability:

> *[...] The fingerprints of Lee Boyd Malvo, now on trial for one of the sniper shootings, were in the FBI's database long before he left a fingerprint at the scene of the brutal robbery and murder in Montgomery, Ala., 11 days prior to the first Washington-area killing. But Alabama did not send that fingerprint to the FBI until a month later [...] Alabama is one of 15 states that have not installed a high-tech connection to the FBI's fingerprint system, which contains 45 million fingerprints. [...] Rarely has there been such a vivid practical example of turning speculation over whether a serious crime could have been prevented into proof that it would have been prevented.*[200]

Recommendation 12 of the NRC report identified the need "to launch a new broad-based effort to achieve nationwide fingerprint data interoperability." The NRC report clearly states the need:

> Great improvement is necessary in AFIS interoperability. Crimes may go unsolved today simply because it is not possible for investigating agencies to search across all the databases that might hold a suspect's fingerprints or that may contain a match for an unidentified latent print from a crime scene. It is also possible that some individuals have been wrongly convicted because of the limitations of fingerprint searches. At present, serious practical problems pose obstacles to the achievement of nationwide AFIS interoperability. These problems include convincing AFIS equipment vendors to cooperate and collaborate with the law enforcement community and researchers to create and use baseline standards for sharing fingerprint data and create a common interface. Second, law enforcement agencies lack the resources needed to transition to interoperable AFIS implementations. Third, coordinated jurisdictional agreements and public

[200] *[Kennedy03]*

> policies are needed to allow law enforcement agencies to share fingerprint data more broadly. Given the disparity in resources and information technology expertise available to local, state, and federal law enforcement agencies, the relatively slow pace of interoperability efforts to date, and the potential gains from increased AFIS interoperability, the committee believes that a broad-based emphasis on achieving nationwide fingerprint data interoperability is needed.[201]

Interoperability problems are a side effect of the US system of federalism, in which the Federal government is limited in its authority over the states, and some states are limited in their authority over localities. In the US, there are hundreds of latent fingerprint identification systems: 35 state or multi-state AFISs, hundreds of local AFISs, the FBI Next Generation Identification (NGI) system, DoD's Automated Biometric Information System (ABIS), and DHS's Automated Biometric Identification System (IDENT).

Interoperability is necessary because there are differences in which subjects are in which databases. The subjects included in each database vary among systems, so that state systems may or may not be supersets of local systems, and NGI is not a superset of all state systems. Since most crimes are local, most of the time agencies will perform searches of state or local systems first. A surprising number of agencies only occasionally search the national FBI system — anecdotally, most state or local examiners I have asked say they search NGI (or IAFIS, NGI's predecessor) only for a few percent of their latents. The lack of interoperability is of greatest concern in cases in which suspects are likely to have records in different jurisdictions, yet high-crime cities on or near state borders often have state agencies that have no means of searching each other's AFISs.[202]

Even when different AFIS databases have the same subjects, searching multiple systems is still desirable. The different systems may have different prints on file, and a mediocre latent may miss on one and hit on another. Similarly, given the imperfect accuracy of AFIS algorithms, different algorithms may hit or miss even on the same exemplars.

Interoperability for ten-print exemplar systems is not a serious issue because all such searches simply use the images themselves and do not require examiners to mark features. Latent AFIS searches are more problematic because different vendors traditionally required proprietary markup of features, differing in format, types of features, and guidance to examiners regarding where and whether features to mark. For example, although all vendors concur that a bifurcation should be located at the point that the ridge forks, vendors have several different expectations of where a ridge ending should be marked; vendors differ on whether debatable minutiae should be marked or omitted; vendors differ on whether minutiae in cores or deltas should be marked; vendors differ on whether they rely on quality maps or counts of ridges between minutiae. EFS was designed to eliminate such proprietary differences.

A variety of initiatives have been undertaken over the years to address Latent AFIS interoperability, all of which I have some involvement with:

- The FBI's Universal Latent Workstation (ULW) was developed starting in 1998 specifically as a tool for AFIS interoperability. Because the different vendors could not agree on a common feature set, I wrote code in ULW to translate between the different feature sets, moving the location of minutiae, asking for different features, and formatting the results in proprietary fields.
- CDEFFS, which we discussed in Section 3.2, developed EFS from 2005-2011.
- NIST and NIJ sponsored the Latent Print AFIS Interoperability Working Group starting in 2008.[203] Among its outputs were

[201] *[NRC09], p S22-S23*

[202] *For example, Kansas City, St. Louis, Chicago, and New York City.*

[203] *[NIST-InteropWG]*

  - *Writing Guidelines for Requests for Proposals for Automated Fingerprint Identification Systems*[204] — a guide to assist agencies in developing Requests for Proposals (RFPs) in order to solicit bids to acquire and implement new interoperable AFISs.
  - *Writing Guidelines to Develop a Memorandum of Understanding for Interoperable Automated Fingerprint Identification Systems*[205] — a guide to developing a latent AFIS interoperability memorandum of understanding (MOU) between two or more agencies.
- The NIST/Noblis Latent Interoperability Transmission Specification (LITS) project, which developed the EFS Profiles specification (Section 7.1.1), the LITS specification (Section 7.1.2), the EFS Markup Instructions (Section 4.1), and "Latent Print Interoperability – State and Local Perspectives", which summarizes the findings from interviews with select state and local law enforcement officials regarding latent fingerprint interoperability.[206]
- The White House National Science and Technology Council (NSTC) Subcommittee on Forensic Science (SoFS) created an AFIS Interoperability Task Force starting in 2011, which resulted in the April 2015 report, *Achieving Interoperability For Latent Fingerprint Identification In The United States*.[207] That report outlines an approach to interoperability including "technical compatibility, network connectivity, proper governance, and performance testing and training within and between systems." Its recommendations regarding technical compatibility are based on the implementation of EFS and LITS.
- The National Commission on Forensic Science (NCFS) in July 2015 issued a *Directive Recommendation: Automated Fingerprint Information Systems (AFIS) Interoperability*,[208] which recommended that "the US Attorney General should support, recommend and fund interoperability of Automated Fingerprint Identification Systems (AFIS) as a national effort to improve public safety." Specifically, NCFS advised the Attorney General to require that any AFIS system that is acquired using federal funding meet interoperability standards using EFS or LITS.

EFS is the common thread through all of these initiatives. EFS was developed based on experience gained in ULW interoperability. The rest of these initiatives all are based on EFS, and on two interrelated specifications that build upon EFS: the *EFS Profile Specification* and *LITS*. To understand the role of these specifications, it is necessary to understand the relation between the base ANSI/NIST-ITL standard and the application profiles derived from ANSI/NIST-ITL:

- The ANSI/NIST-ITL standard defines an overall file format and defines the records and fields that can be included in that file. EFS defines a number of fields that can be used within one record type. ANSI/NIST-ITL (and thereby EFS) do not define which fields are appropriate for a given purpose.
- The ANSI/NIST ITL standards are the basis for biometric and forensic ***application profile*** specifications used around the world, including the FBI's Electronic Biometric Transmission Specification (EBTS), DOD EBTS, DHS IXM, Interpol's INT-I, and a wide variety of national, state, and local application profiles. Application profiles define transactions, which are the specific combinations of ANSI/NIST-ITL records and fields used for a given purpose. For example, the latent AFIS transactions defined by FBI's EBTS include image- and feature-based latent print searches, latent search results, unsolved latent matches, and image requests — each of these has a specific set of required and optional fields and records.

[204] *[Ballou13a]*

[205] *[Ballou13b]*

[206] *[Noblis12]*

[207] *[NSTC15]*

[208] *[NCFS15]*

7.1.1 EFS Profiles

The *EFS Profile Specification*[209] is a supporting document to ANSI/NIST-ITL that defines the sets of EFS features to be used in different types of latent AFIS searches and responses. EFS Profiles provides a bridge between the base ANSI/NIST-ITL standard and the various application profiles. The EFS Profiles are designed so that they may be incorporated by reference into LITS, FBI or DoD EBTS, Interpol INT-I, or other application profiles.[210] This decoupling of feature sets from transactions enables different transactions (or transactions from different organizations) to share a common feature set, aiding in interoperability. EFS Profiles are incorporated by reference by ANSI/NIST-ITL, and into LITS.

Multiple EFS Profiles are defined to allow for tradeoffs between examiner time and search accuracy; these profiles provide a range from image-only searches (requiring no examiner markup) through standard minutiae searches to profiles including skeletons or ridge counts (to maximize accuracy at the cost of additional examiner markup time). EFS profiles for AFIS searches include Image-only, Minimal markup (region of interest, orientation, pattern class, cores, deltas), Quick minutia search (Minimal markup plus minutiae), and Detailed markup (Quick minutia search plus ridge quality/confidence map, ridge flow map, center point of reference, distinctive features, dots, incipients, and core-delta ridge counts). For casework markup, the Full annotation profile incorporates all EFS features.

7.1.2 Latent Interoperability Transmission Specification (LITS)

LITS[211] is an application profile that defines AFIS transactions for exchange among state and local law enforcement agencies. *LITS* is a system-level specification that focuses on the definition of vendor-neutral latent transactions to be exchanged among disparate cross-jurisdictional AFIS. LITS is parallel to and compatible with the FBI EBTS, which is solely limited to the scope of transactions to and from the FBI; LITS extends EBTS for states and localities to interchange information with each other; by definition, a *LITS*-conformant system is compatible with FBI CJIS *EBTS*. LITS-conformant AFIS systems provide the examiner with a seamless search capability for all interoperable AFIS systems.

LITS also provides a standard for exchange of non-AFIS documentation and casework information, as we discussed in Section 4.4.

7.2 Evaluation of Latent Fingerprint Technologies: Extended Feature Sets (ELFT-EFS)

The NIST ELFT-EFS Evaluations were conducted in 2009 and 2010 to evaluate the state of the art in latent matching, by comparing the accuracy of searches using images alone[212] with searches using different sets of EFS features marked by experienced latent print examiners.

One of the purposes of ELFT-EFS was to determine the extent to which human feature markup was effective. Because human markup is expensive in terms of time, effort, and expertise, there is a need to know when image-only searching is adequate, and when the additional effort of marking

[209] [EFSProfiles]

[210] [NIST-AppProfiles]

[211] [LITS]

[212] Image-only matching is described in the ELFT reports as "automatic feature extraction and matching (AFEM)." I believe that "Image-only matching" is clearer: often examiners doing a feature-based search will use automated feature extraction as a basis for their markup; automatic matching is a given for AFIS.

minutiae and other features was appropriate. ELFT-EFS was not a test of automatic EFS extraction (i.e. conformance to the standard), but rather a test of data interoperability and how potentially useful such human marked features are when processed by the matcher.

The ELFT-EFS evaluations were open to both the commercial and academic community. Participants included five commercial AFIS vendors: Sagem, NEC, Cogent, Sonda, and Warwick; the performance of Sagem, NEC, and Cogent was substantially better than the others, so the others are omitted from the data presented here. The participants each submitted three Software Development Kits (SDKs): a latent fingerprint feature extraction algorithm, algorithms for ten-print feature extraction and gallery creation, and a 1-to-many match algorithm that returned a candidate list. Evaluations were run at NIST on commodity NIST hardware.

ELFT-EFS evaluations #1 and #2 used the same test methodology ELFT-EFS #2 provided the vendors the opportunity to correct possible errors and to make adjustments to their algorithms in light of the findings published in ELFT-EFS #1. The results discussed here are generally from ELFT-EFS #2.

7.2.1a Fingerprint and Feature Data

The ELFT-EFS #1 test dataset contained 1,114 latent fingerprint images from 837 subjects. The ELFT-EFS #2 test dataset contained 1,066 latent fingerprint images from 826 subjects (38 latents (from 4 subjects) were removed after ELFT-EFS #1 since they were provided as example images to the participants for miss-analysis purposes; 10 of the Evaluation #1 latents that did not have mates in the gallery were also removed). The gallery was comprised of (mated) exemplar sets from all latent subjects, as well as (non-mated) exemplar sets from 99,163 other subjects chosen at random from an FBI provided and de-identified dataset. Each subject in the gallery had exemplar sets containing rolled and plain impressions from all ten fingers.

In addition to fingerprint images, each latent had an associated set of hand-marked features. The features were marked by twenty-one International Association for Identification Certified Latent Print Examiners (IAI CLPE) using guidelines developed specifically for this process; these guidelines served as the basis for *Markup Instructions for Extended Friction Ridge Features* (Section 4.1). No vendor-specific rules for feature encoding were used; all encoding was made in compliance with the EFS specification. The various subsets of latent features were the precursors to the EFS Profiles discussed in Section 7.1. Features were marked in latent images without reference to exemplars, with the exception of a subset of 458 latent images that included an additional Ground Truth (GT) markup based on the latent and all available exemplars; GT markup provides a measure of ideal (but operationally infeasible) performance when compared to the original examiner markup.

The extended features included minutiae, ridge counts, cores & deltas, pattern class, ridge quality maps, creases, dots, incipient ridges, ridge edge protrusions, and pores. A subset of the latents had skeletons marked (including associated ridge flow maps). Latent examiners made determinations of Value, Limited Value (latents of value for exclusion only), or No Value at the time of markup, in addition to informal quality assessments of “Excellent”, “Good”, “Bad”, “Ugly”, and “No Value”.

7.2.1b Methods of analysis

Analyses of the accuracy of 1:N identification searches returning candidate lists can be with respect to rank or score.

- In rank-based analyses, identification rate at rank *k* is the proportion of the latent images correctly identified at rank *k* or lower. A latent image has rank *k* if its mate is the kth largest comparison score on the candidate list. Recognition rank ranges from 1 to 100, as 100 was the (maximum) candidate list size specified in the API. Overall accuracy results for rank-based metrics are presented via Cumulative Match Characteristic (CMC) curves. A CMC curve shows how many latent images are correctly identified at rank 1, rank 2, etc. A CMC is a plot of identification rate (also known as ”hit rate”) vs. recognition rank. Rank-based analyses are specific to the gallery size used in the test, and cannot be assumed to scale to substantially larger gallery sizes.

- Score-based analyses are defined with respect to True Positive Identification Rate (TPIR) and False Positive Identification Rate (FPIR). TPIR indicates the fraction of searches where an enrolled mate exists in the gallery in which enrolled mates appear in the top candidate list position (i.e. rank 1) with a score greater than the threshold. (Note that the False Negative Identification Rate (FNIR = 1-TPIR) indicates the fraction of searches in which enrolled mates do not appear in the top position with a score greater than the threshold.) FPIR indicates the fraction of candidate lists (without enrolled mates) that contain a non-mate entry in the top candidate list position with a score greater than the threshold. Score-based results are more scalable than rank-based results, providing a better indication of how accuracy would be affected by an increase in database size. As a rule of thumb, the TPIR at FPIR = 0.01 provides a rough projection of accuracy for an increase in database size of 100x, so the score-based results discussed here are rough estimates of what rank-1 results would be for a gallery with 10 million subjects. In theory, analysis could use a combination of score and rank, in which scores are filtered based on rank. In practice, score-based results at rank 1 and at rank 100 were not notably different, so results presented are for scores at rank 1.

7.2.1c Results

Table 11 summarizes the results from ELFT-EFS#2.

		Image only (LA)	**Image + ROI (LB)**	**Image, ROI, quality, pattern class (LC)**	**Image + minutiae (LD)**	**Image + full EFS (LE)**	**Image + full EFS + Skeleton (LF)**	**Minutiae + ridge count (no image) (LG)**
Rank-1 ID Rate	Sagem	63.4	64.1	64.1	65.6	65.6	64.8	40.4
	NEC	57.7	60.1	60.1	67.0	67.0	68.2	47.4
	Cogent	59.6	60.1	58.6	66.3	67.2	*n/a*	45.9
ID Rate where FPIR=0.01	Sagem	52.9	54.8	54.8	57.2	57.2	50.8	30.9
	NEC	47.8	46.9	48.8	56.5	56.5	53.8	35.4
	Cogent	50.5	50.7	50.5	57.9	57.6	*n/a*	38.3

Table 11: Rank-based and score-based identification rates. Score-based results (418 latents, 100,000 exemplar subjects)[213]

- The highest accuracy for all participants was observed for searches that included examiner-marked features in addition to the latent images. However, image-only matching was nearly as accurate.
- Image-only searches were more accurate than feature-only searches for most matchers.
- Since score-based results are more scalable than rank-based results, they provide a better indication of how accuracy would be affected by an increase in database size. This capability could provide operational benefits such as reduced or variable size candidate lists, or for reverse latent searches (searches of databases containing unsolved latents) where using a score threshold was used to limit candidate list size.
- The effect of the use of EFS features other than minutiae was mixed. Feature-only (LG) performance was far less accurate than Image-only (LA). Image-only (LA) performance was improved by adding region of interest (ROI) (LB), and improved further by adding minutiae (LD). Cogent improved accuracy further through the use of all EFS features; NEC improved through the use of skeletons.

[213] *From [ELFT-EFS2], Table 9b*

- The ground truth (GT) markup method, in which all exemplar mate images were consulted when marking latent features, yielded an increase in performance over the original examiner markup of about 4 to 6 percentage points for image + full EFS searches, and about 12 to 15 percentage points for minutiae-only searches. The GT markup shows the ideal (and operationally infeasible) limit of the accuracy of feature markup. This shows that matcher accuracy is highly affected by the precision of latent examiner markup, especially in the absence of image data.
- Latent orientation (angle) has an impact on matcher accuracy. When the orientation of latents was unknown, the rank-1 identification rates were approximately percentage points lower than the overall average.
- Although latents assessed to be of value are much more likely to hit than latents assessed as not of value, the latter nevertheless do sometimes succeed. Matcher accuracy was very clearly related to the examiners' latent print value determinations, with much greater accuracy for latents determined a priori to be of value. The matching algorithms demonstrated an unexpected ability to identify low feature content latents: Sagem's rank-1 accuracy for No Value latents was 20% on image-only searches, and 26.2% on Limited Value latents. Some agencies consider some low-quality images "No Value for AFIS" or "Not Suitable for AFIS" — however, there is no justification for considering a latent to be of value for non-AFIS casework, but too poor quality for AFIS. If a latent is good enough for an examiner to use in comparison, it can still be searched — and AFIS can even hit on some images that an examiner cannot use.
- The performance of all matchers decreased consistently as lower quality latents were searched, with respect to the informal scale of "Excellent", "Good", "Bad", or "Ugly".
- Analysis showed that the greatest percentage of the misses were for latents with low minutiae count, and those assessed by examiners as poor quality ("Ugly") or "No Value". Algorithm accuracy for all participants was highly correlated to the number of minutiae. For latents with more than 10 minutiae, minutiae count was the most important factor for successful identification with examiner-assessed quality being secondary. For latents with fewer than 10 minutiae, examiner-assessed quality was a better predictor of match accuracy than minutiae count.
- Approximately 22% of the latents in the test were missed by all matchers at rank 1, more than half of which could be individualized by a certified latent examiner. The initial or reviewing examiners determined that 14% of the latents in the test were of No Value, of value for exclusion only, or resulted in an inconclusive determination; about one third of these could be matched by one or more matchers at rank 1.
- The highest measured accuracy achieved by any matcher at rank-1 on any latent feature subset was 66.7%, even though approximately 78% of the latents in the test were matched by one or more matchers at rank-1. This indicates a potential for additional accuracy improvement through improved algorithms, or through the use of data based fusion (e.g. search using image-only and again search using image+features). The differences in which latents were identified by the various matchers also points to a potential accuracy improvement by using algorithm fusion.
- The use of both rolled and plain impressions in the gallery resulted in higher accuracy than the use of either rolled or plain impressions separately for most matchers. Use of plain impressions in the gallery as compared to rolled impressions resulted in a drop in accuracy for most matchers.

The ELFT-EFS results showed substantially lower accuracy than the earlier ELFT Phase II evaluation,[214] even though three vendors participated in both tests. This was an expected result

[214] *[Indovina09]. Note I was not involved in EFLT Phase II.*

because the ELFT-EFS test data included a greater proportion of poor-quality latents, had higher throughput requirements, and larger gallery size. In addition, the data used in ELFT Phase II was selected by using an AFIS to determine the mates, which resulted in a dataset that omitted all of the latents that could not be successfully searched. A dataset resulting from such "AFIS bias" (a type of survivorship bias) can be expected to have near-perfect AFIS performance, but cannot be considered as representative for any evaluations.

There has not previously been a public evaluation of latent fingerprint matchers of this scale — particularly in which systems from different vendors used a common, standardized feature set. The results show that searches using images plus manually marked EFS features demonstrated effectiveness as an interoperable feature set. The four most accurate matchers demonstrated benefit from manually marked features when provided along with the latent image. The latent image itself was shown to be the single most effective search component for improving accuracy, and was superior to features alone in most cases. For most matchers, the addition of new EFS features provided an improvement in accuracy. The accuracy when searching with EFS features was promising considering the results are derived from early-development, first-generation implementation of the new standard.

In evaluating results, note that the ELFT-EFS tests were evaluations of data compliant with an emerging draft specification: the file format and syntax and semantics of the features were not familiar to the participants, and therefore software for parsing and using the features had to be developed with limited opportunity for testing; the schedule was extremely demanding for the participants, and did not permit time for extensive research and development and software debugging.

Ideally, such evaluations should follow the model of the NIST Proprietary Fingerprint Template (PFT) Evaluations,[215] which are long-term ongoing evaluations of exemplar fingerprint algorithms. Such ongoing evaluations provide a feedback looping allowing vendors to continually improve and evaluate their systems. Such evaluations provide needed transparency, so that vendors and their customers all can share an up-to-date understanding of comparative performance.

[215] *[NIST-PFT, NIST-PFT2]*

Chapter 8 Findings and Recommendations, Implications and Future Possibilities

How do I envision latent print examination progressing into the future? In the near term, the existing latent print examination processes can be made more rigorous without waiting for the completion and validation of statistical models and automated solutions (Section 8.1). In the medium to long term, the examiners' conclusions can be replaced or augmented with automated processes (Section 8.2).

In making these recommendations, I wish to reiterate that these are my personal opinions, and should not be seen as the position of any organization or agency with which I am affiliated.

8.1 Enhancing existing processes

Until probabilistic models or automated decisions are developed that are capable of replacing human examiners in all casework, the latent print community must work to enhance the existing manual system to make it more robust, transparent, and quantifiable. I think that the following steps would make substantial progress toward that end.

8.1.1 Addressing inconsistencies through standardized training, competency and proficiency tests, operating procedures, certification, and accreditation

There is currently a great deal of variation in latent print examination procedures and terminology, among agencies, among training programs, and among examiners. These differences present major problems: varying results by organization and by examiner, inserting ambiguity into legal testimony, and impeding cross-agency evaluations of examiners' performance.

I believe that in order to address these inconsistencies, the latent print community must work to standardize multiple facets of the process:

- ***Training*** — Training and continuing education of latent print examiners should be made more consistent to ensure that practicing examiners in any organization have been thoroughly trained to scientifically-based standards and best practices that reflect the current state of the practice.
- ***Competency tests*** — I am very concerned regarding the wide disparity of who is considered to be competent to testify as a latent print examiner. Agencies currently follow their own policies to determine when an examiner who has completed a training program is ready for casework. I believe that passing a standard competency test should be required for anyone to be permitted to testify to expert opinions as a latent print examiner. Variants of the Black Box test could be used as the basis for such tests, to assess whether an examiner is sufficiently accurate and reliable to perform casework. In addition to testing skills, these tests should test that examiners can effectively implement the current state-of-the-practice standards and best practices.
- ***Proficiency tests*** — In addition to a pass/fail test of competency, I believe that there is a need for testing to assess the varied level of skills among competent examiners. There is a great variation in the skills of examiners, and I believe that many examiners do not themselves have a good way of assessing how good they are. The Black Box results showed that the skill of latent print examiners is multidimensional, suggesting approaches that could be used in constructing proficiency tests. Both accuracy (ability to avoid errors) and effectiveness (ability to avoid being inconclusives) should be considered. Proficiency tests that report different levels of skill would be invaluable in differentiating among examiners so that agencies could focus training, and direct complex comparisons or verifications to specific individuals. Standardized proficiency

tests would permit black box testing conditioned on examiner skill, as discussed in Section 8.1.2, and provide greater transparency into the abilities of the individual examiners.

- ***Operating procedures*** — Agencies' standard operating procedures must reflect the legal requirements imposed for each jurisdiction, as well as the specific mission for each agency. Beyond those requirements, however, operating procedures should not vary as widely as they currently do. Agencies should work to ensure that latent print examination and reporting do not differ in meanings and implications across agencies. Operating procedures should be regularly revised to reflect the current state-of-the-practice standards and best practices, and this should be considered in the accreditation of laboratories.
- ***Certification and accreditation*** — In my view, the admissibility of evidence should be predicated on the accreditation of laboratories, as well as certification of individual examiners. I believe that there should be a minimum certification, based on competency testing, to serve a threshold for doing operational casework; advanced certification of greater levels of expertise could be based on proficiency testing.

8.1.2 Further Black Box testing based on examiner proficiency and comparison difficulty

The consumers of an examiner's decisions (laboratory managers, police, prosecution, defense, judge, and jury) need to have at least a basic understanding of the accuracy and reliability of those decisions. The Black Box studies were conducted to provide such rates from a general, overall perspective — which PCAST has reported adequately satisfy Daubert criteria for admissibility. These studies are valuable in providing a rough order of magnitude understanding of examiner capabilities — data which is absent for a number of other forensic areas.

However, Black Box tests to date are based on performance for examiners in general, on latents and exemplars of a range of qualities. We know that there is a wide disparity in the difficulty of latent print comparisons as well as in the skills of examiners, and therefore overall averages should be seen as only the first step. In practice, the consumers of examiners' decisions are not just interested in overall averages, but are particularly interested in a specific examiner's abilities to render a decision for a specific comparison. Future Black Box tests should be conditioned on examiner proficiency and comparison difficulty. This implies both standardized tests of examiners' proficiency (Section 8.1.1), and metrics of the difficulty of latent print comparisons (Section 8.1.10). The ultimate goal would be that when an examiner makes a decision on a specific image pair, the decision would be reporting with corroborating data from Black Box testing showing the accuracy, reproducibility, and repeatability of decisions for other examiners with an equivalent proficiency, on comparisons of an equivalent difficulty.

8.1.3 Focusing on effectiveness as well as error

As I said in Chapter 2, the range of criticisms of latent print examination have resulted in improvements: the criticisms made it possible to review and change long-standing assumptions and practices, resulting in improved practices and a climate open to change. However, the criticisms have focused overwhelmingly on erroneous individualizations. In any analysis conducted for operations research or process reengineering, trying to optimize a single error would be a red flag: errors almost involve tradeoffs. We can (facetiously) eliminate all erroneous IDs by doing no work whatsoever, which is obviously not an acceptable solution.

In 2010, a friend and Noblis colleague, Calvin Yeung, was murdered in Maryland in an apparent road rage or carjacking incident. The perpetrator is still unknown and (it is reasonable to assume, given the nature of the crime) is likely to have gone on and committed other such crimes. In 2014, 52.6% of violent crimes in the United States went unsolved as did 35.5% of homicides.[216]

Erroneous IDs are not the only failure of forensic science: failing to make effective use of the resources available is also a failure of the discipline.

Training and competency/proficiency tests should reflect an increased focus on effectiveness and efficiency, not just the avoidance of error. As we move forward, I would urge decision makers to remember that as we focus on eliminating or minimizing error we must at the same time focus on effectiveness. Proposed quality assurance measures and changes to standard operating procedures should be assessed not only in regard to the effect on error rates, but also in regard to the impact on the amount of casework that can be performed.

8.1.4 Standardized metrics for laboratory workflow

When an agency makes major changes in their procedures, how can they tell what the impact is? If, for example, an agency implements a sequential unmasking process to address potential bias, there should be a method to assess whether the processes have an adverse effect on the amount of casework conducted. However, agencies vary tremendously in how – or whether – they define or quantify laboratory workflow. Latent print units differ enough in how they track and count that it is very difficult to get reasonable comparisons of e.g. the proportion of cases that result in any individualizations, the proportion of latents that are assessed of no value, or the proportion of AFIS searches that result in individualizations.[217]

Standard metrics for casework would mean that agencies would define, count, and measure their processes the same way. This would enable business process reengineering methods to assess the efficacy and efficiency of procedures — which would in turn make it possible for decision makers to understand the full impact of changes in procedures.

8.1.5 Required documentation

I concur with the Mayfield OIG report, NRC report, and multiple colleagues[218] that detailed documentation of the features used by examiners in making their determinations should be required. Analysis features should be marked prior to the comparison phase, so that features changed during Comparison can be clearly differentiated.[219] 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). Conflict resolution procedures vary among agencies, and I believe that increased transparency in reporting conflict is appropriate. One concern if conflicts are not clearly documented is that the resolution may be seen (rightly or wrongly) as a dispute in which the examiner with the stronger personality prevails.
- Standardized documentation for reporting and testimony. I was surprised when I joined SWGFAST that there was no standard for courtroom presentations of latent print evidence. I believe that agency reports and testimony of latent print conclusions would be well-served by standard formats and content.

[216] *[FBI14]*

[217] *G.I. Kiebuzinski, personal communication*

[218] *e.g. [Mayfield06, Langenburg12b, Neumann13b, Evett96, Swofford13, Haber09]*

[219] *I think that a strict "linear ACE" process in which only those features marked during Analysis can be used during Comparison (see e.g. [Haber09]) is unrealistic: after a number of comparisons it becomes clear that some features are accidentally missed during Analysis, and others are reinterpreted during Comparison. Such changes cannot be ignored, but must be documented, and should be used with caution.*

- More detailed information available for technical review and (non-blind) verification of casework. Although blind verification necessarily omits all documentation, technical review and non-blind verification also serve important roles, determining whether an examiner has adequately justified conclusions.
- Increased automation of quality assurance procedures. Automated quality assurance could automatically flag examinations with conclusions 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.
- Long-term archiving of the bases for conclusions in cases that may take years to go to court.
- For use in probability models, as discussed in Section 8.2.2.

Note that although most of these benefits of detailed documentation relate to quality assurance, I do believe that standardized detailed documentation may assist the individual examiner in being more rigorous when making borderline decisions, by considering explicitly the extent of information available. However (as discussed in Section 6.3.2f), we cannot always assume a causal relation between documentation and conclusions. Because there is no demarcation between the Comparison and Evaluation phases, documentation serves in part to indicate the basis for a conclusion, but also serves to justify that decision.[220]

Whether to always require detailed documentation is a policy decision that adds overhead to the examination process. When a latent has e.g. 80 minutiae, requiring documentation of all features during Analysis is excessive, as it would force a disproportionate amount of time of the easiest latents. I would suggest instead that an LQMetric threshold be used, so that all latents with e.g. LQMetric < 75 would be required to have detailed Analysis markup of clarity, minutiae, cores, and deltas.[221] I would suggest that corresponding (and discrepant) minutiae, cores, and deltas should always be documented during Comparison, but after a large number of corresponding features (e.g. more than 20) it would be reasonable to indicate that additional correspondences were detected but not marked. Regarding non-minutia features, I would suggest requiring cores and deltas to be marked in Analysis and Comparison, but other features (e.g. incipients, dots, creases, pores) would only be marked if used as the basis for a conclusion.

If detailed documentation is not required in all cases, there are some instances that I believe should always be documented in detail: in conflict resolution, the initial examiner and verifier should detail the bases for their determinations prior to discussing their differences; in homicide or equivalent cases; in any case in which the sole or predominant evidence is a single latent.

Rigorously defined and consistently applied methods of performing and documenting ACE-V would improve the transparency of the latent print examination process, and reduce the risk of error.

[220] *I do not think that documentation as justification necessarily implies bias. I believe that in a comparison that results in individualization, the examiner starts the Comparison process neutrally, but as more corresponding features are detected and the comparison begins to look like a possible individualization, the examiner looks increasingly for any potential discrepancies and in that process uncovers more corresponding details. Hence we saw in Section 6.3.3 that individualizations often were associated with minutiae added during Comparison.*

[221] *This does leave the possibility that a very high-quality latent would not be annotated during Analysis, but would in Comparison be found to have only a minimal overlap with the exemplar, and therefore the conclusion would be based on a few Comparison features, none of which were marked during Analysis. A possible way of mitigating this would be that latent fingerprints that are not centered (and/or have no core or delta present) would always need to be annotated in detail.*

8.1.6 Standard transactions for archiving and exchange of casework

As discussed in Section 4.4, the LITS casework transactions (COMP, ASYS, and CWE) provide a basis for long-term standardization of how latent print examinations are documented, exchanged, verified, reported in legal contexts, and made available for quantitative quality assurance. I believe that there should be a nationwide requirement for analysis to be documented and archived in ASYS files, and comparison/evaluation in COMP files, and that such files be the standard for agency reporting of decisions in a legal context. I believe that CWE transactions should be used for long-term archiving of the ancillary images and information associated with a case. I do not believe that ACEware or ULW are the only means of handling such files — indeed, I believe that such would be undesirable — but that software such as PiAnoS and Mideo Latentworks be compliant with EFS and the LITS casework transactions.

8.1.7 Blind verification of major decisions and mitigating cognitive bias

For that portion of casework in which an individual latent print decision is the only evidence (or the predominant evidence) in a homicide, the implications of an erroneous or debatable decision would be much more severe than in typical casework. In such instances, it is particularly important to have additional corroboration. Similarly, the implications of potential bias are greatest when there is a single decision in an important case. I believe that in such cases, one or more fully independent blind verifications of the decision should be required, in which the verifying examiner(s) receive just the latent print(s) and exemplar(s), with no information regarding the type of case or the suspect. When the initial examiner's decision is reported, it would be required to be accompanied by the independent blind verification decision(s). For small agencies (in which it may be impossible to blind one examiner from the details of a case), the blind verification should be performed by a different agency — for transparency, even large agencies might consider having blind verifications performed by outside agencies.

There has been a great deal of discussion of the risk of cognitive and contextual bias in forensic science, and ways of mitigating it.[222] The case manager[223] and sequential unmasking[224] approaches are a priori methods to shield the initial examiner from potential bias. My concern with these approaches is that they need to be considered in regard to both potential benefits and costs, given that they could be resource intensive and require high levels of expertise:[225] their relative value is much greater if they can be implemented in a way that would have minimal impact on workflow. Blind verification could be a less resource-intensive way of addressing possible bias, if it is restricted to those conclusions that are at greatest risk of bias: single decisions on major cases. The a priori approaches would entail a cost across all cases, whereas blind verification could be focused on a critical subset of casework.

8.1.8 Verification of no value, inconclusive, and exclusion determinations

In many agencies, verifications are only performed for individualization determinations. Given the variability of examiners' determinations, it is unreasonable for agencies to assume that a single examiner's determination will be reproducible. Not verifying all determinations means that inappropriate determinations will not be detected, potentially resulting in missed conclusions.

[222] *E.g. [Risinger02, Saks03, Dror06a, Dror06b, Krane08, NRC09, Thompson11, Champod14, Risinger14, NCFS15b]*

[223] *[Thompson11]*

[224] *e.g. [Krane08]*

[225] *Because these approaches themselves require expertise in determining what information to show/hide from the examiner, in the most critical cases there may still be a reason to want to blind verify conclusions.*

8.1.9 Greater continuum for determinations

Much of the reason for the imperfect repeatability and reproducibility of examiners' determinations appears to be due to discretization error: making categorical decisions in borderline cases. Our results indicate that the value of latent prints is a continuum that is not well described by binary (value vs. no value, or individualization vs. inconclusive) determinations. Lack of repeatability or reproducibility is much more understandable if we consider, for example, that an examiner may (consciously or unconsciously) be 51% convinced that individualization is more appropriate than inconclusive in the comparison at hand. I feel that the s-curves shown in Figure 15 are as effective a means as I have encountered of showing why a more continuous representation of the decision process is needed to replace or augment examiners' determinations. The s-curves show that the decision space for value and individualization decisions is really a continuum of how certain the examiners are that a determination is warranted; the instances on the slope of the s-curve show that the categorical responses that examiners are required to make are not well-suited to the data. Given this, I do not see the 85-90% repeatability and reproducibility rates in this test as a criticism of the examiners, but a criticism of the system: about 10-15% of the test data had no obvious answer, and therefore it is unreasonable to expect the answers to be consistent in those instances; for some comparisons the answer is clear, and therefore repeatability and reproducibility are high.

In the medium or long term, I assume that probabilistic determinations will provide such continuous measures. For the near term, we need to consider how to improve the existing holistic determinations. I believe that one issue relates to how the few categories of determinations do not accurately reflect the subtleties of examiners' decisions. A very complex identification that takes hours to decide is not inconclusive does not warrant the same confidence as an identification of large and pristine impressions. I see several approaches that can be taken:

- Examiner assessments of "complex" or "difficult" comparisons — complex comparison determinations can be flagged using the definitions of complexity provided by SWGFAST[226] and EFS. Assessing complexity could be included in training and guidelines so that examiners can use the assessment as consistently as possible. Determinations based on complex comparisons could then have a greater level of quality assurance review.
- Examiners could be permitted to make determinations between individualization and inconclusive (and between exclusion and inconclusive): a determination of e.g. "limited support for a same source conclusion" could be permitted for examiners, which then agencies could report differently than individualization, or have a greater level of quality assurance review.
- Automated quality metrics (Section 8.1.10) could be used instead of examiner assessments of difficulty — with the caveat that the current quality metrics are assessments of a single image, and therefore are not assessing the difficulty or complexity of the comparison.

Such methods of indicating borderline determinations may be useful in flagging prints whose value determinations are likely to be debatable. These approaches could be used in establishing business processes to manage risk and optimize workload: for example, a quality assurance process could require review of determinations for low-quality or complex prints, direct such prints to highly qualified examiners, or require rigorous verification when such prints are used in comparison.

8.1.10 Uses of latent quality metrics

The ability to assess the quality of a latent, or the comparative quality of a latent-exemplar comparison, suggests a variety of possible uses. Some of these may be realizable now using LQMetric; others may require further research or development of quality assessment tools.

[226] *[SWGFAST-Conclusions13]*

- In court, when there is a challenge to latent print testimony, the instinctive reaction from many in the fingerprint community is to ask about the quality of the fingerprints, assuming that there may be a basis for debating conclusions on poor-quality latents, but none if the latents are high quality. The lack of standard methods of assessing quality means that all latent print evidence must be treated as if it is all the same. Latent quality metrics would allow the quality of the evidence to be considered in court, for example in determining whether or not to challenge evidence, or in determining whether additional latent print examiners should be brought in by the defense.
- In evaluations of examiners (such as our Black Box test), the measurements of accuracy and reproducibility could be categorized by quality, with (potentially) different error rates for each quality bin. Ultimately (in theory), this could be accompanied by assessments of examiner proficiency so that evaluation results could be cited for different combinations of examiner proficiency and latent quality.
- Quality-directed workflow, in which incoming work or backlog is prioritized (or triaged) based on quality. For example, large numbers of images from a crime scene photographer could be sorted or grouped based on quality; the highest-quality latents could be searched against an AFIS first in order to increase the probability that hits are made rapidly.
- Providing an objective measure of difficulty for use in quality assurance (e.g. flagging complex prints for special handling/additional verification, or directing poor-quality latents to more expert examiners).
- Providing an automated means of verification for value determinations, especially for NV determinations (which in some agencies are never verified).
- Helping to separate fingerprints from non-fingerprints in large heterogeneous image databases.

In addition to the quality of individual latents, latent quality metrics can also be used as a way of describing overall datasets. When using latents for AFIS evaluations, the measured accuracy is significantly affected by the quality of the latents, which makes it difficult to compare evaluations that did not use the same latents: a higher measured accuracy on a test might just indicate that it used easier data, rather than a substantive difference among matchers. By characterizing the quality distribution of a dataset of latent prints, quality metrics provide a means to assess whether different datasets are comparable.

Similarly, latent quality metrics may be of use in describing data used for proficiency tests or other evaluations of (human) latent print examiners. Evaluation results accompanied by the LQMetric distribution makes it easier to compare results among tests.

8.1.11 Differing AFIS search strategies[227]

Some agencies have historically treated all latent searches the same: always marking minutiae and counting ridges, and always comparing all twenty candidates returned. We are trying to shift this paradigm and get users to think in terms of ***search strategies***: deciding how to conduct latent searches of an AFIS based on the requirements and implications of that specific case and that specific latent. Search strategies seek to optimize tradeoffs between effectiveness (maximizing the likelihood that a search will result in a hit) and efficiency (minimizing the effort required for searching and comparing candidates).

For most agencies, different cases may have widely different requirements, based on the case priority, how much examiner time is available, and the workload:

- minor cases that would otherwise never be searched may justify only a minimal effort ("low-hanging fruit");

[227] *[Hicklin15]*

- a homicide may necessitate an exhaustive search and many times the amount of effort of a routine case ("no stone left unturned");
- backlog, cold cases or an overwhelming workload may benefit from prioritization ("biggest bang for the buck");
- routine cases need to balance among these approaches.

There are several possible search strategies that could be used to accomplish these different objectives when searching latents against an AFIS, by trading off between examiner time (efficiency) and the probability of making hits (effectiveness):

- Minimizing effort — Minimize the examiner time to conduct searches and review responses in order to process as many latents as possible, by lowering the probability of identification for each individual search (such as for low-priority cases, property crimes, backlog, or cold cases where it is not practical to do an exhaustive search of every latent).
- Maximizing probability of an individualization — Maximize the probability that a specific latent will be identified, by increasing examiner time (such as for high-priority cases).
- Prioritizing workload — Prioritize workload by sorting searches and responses so that the most likely identifications occur first (such as for time-critical cases, or for large cases where early identifications may mean that not all latents would need to be searched).
- Balanced — For routine cases, an appropriate search strategy is likely to be a compromise among the other strategies: the probability of making hits is balanced against examiner time.

Selecting a search strategy can be based on a variety of factors, such as case priority, number of latents in the case, forensic relevance (probative value) of the latent, quality of the latent, overall workload, and staffing availability.

The Netherlands Police (Netherlands Politie) decides how to conduct searches based on the type of crime and the investigator's judgment as to the depth of the investigation required. In the Netherlands four levels of crime scene investigation are practiced; high volume crime (burglary/theft etc.), serious crime (robbery/rape), serious crime plus (Murder/Terrorism), and Disasters. The depth of the investigation is determined by the severity of the case and the probative forensic value of the latent. For the lowest-priority crimes only image-only searches are used, with no manual feature markup. For the highest-priority crimes (a very small percentage of all crimes), an exhaustive search will have an image-only search and up to three different examiners each submitting three different feature searches in a sequential manner (i.e. 1 image-only and 9 feature searches). The Netherlands Police strategy is based in part on analyses of their own system that found 62% of their IDs came from image-only searches, 24% from the first manually-marked feature search, 8.8% from the second, on down to 0.3% from the ninth.[228] Every additional search increases examiner time and has a descending likelihood of an individualization: their approach uses the case priority and probative value to determine how much additional examiner time should be expended, given a decreasing (but non-zero) chance of making an individualization.

It should be noted that such search strategies explicitly have different policies depending on the type of case, which is contrary to the sequential unmasking concept, and therefore blind verification or other means of mitigating bias should be considered.

8.2 Replacing or augmenting examiners' determinations

The long-range ideal would be for latent print examinations to be automated (as with ten-print exemplars), or at least based on detailed statistical models (as with nuclear DNA). I do not see either of these completely replacing examiner conclusions in the near term, but I expect both to be a reality for a portion of casework in the medium term.

[228] *[Riemen12]*

8.2.1 Replacing examiners with fully automated feature detection and decisions[229]

Latent AFISs traditionally have not made automated identification decisions, but instead return lists of high-scoring potential candidates for human examiners to compare and make final conclusions. There are a variety of potential approaches to partial "lights-out" systems,[230] but the ideal solution — if possible — would be fully automated image-only conclusions without a need for examiner feature markup.

For state-of-the-art latent AFIS systems with a criminal justice use case, there is good reason to believe that a small percentage of searches (possibly 5-15%) return matcher scores so high that there can be a statistical basis for very large-scale latent AFIS systems (e.g. 10-100 million subjects) to return automated identification decisions ("AutoID") for those high-scoring searches; the remainder would return candidate lists as is current practice.

Currently, latent AFIS acts as a tool for finding candidates in the database — human examiners retain full responsibility for making identifications. With an AutoID capability, we envision a process in which those latent searches that result in the highest scoring matches would not return a candidate list but would instead result in a new response transaction that would return a single candidate that was automatically identified by the AFIS. The proportion of searches that would be affected would vary by use case: we estimate 5-15% of searches for crime scene latents against a very large-scale AFIS could be automatically identified.

AutoID would improve the efficiency of the latent print examination process, which would be further enhanced due to the increased effectiveness of image-only latent searches, which in turn would substantially reduce the time required to prepare a search. The combination of AutoID and the increased effectiveness of image-only searches means that it is feasible to have some fully "lights-out" latent casework, with no examiner markup needed for the search, and no examiner comparisons needed for some decisions. Such automated decisions could have a very significant effect on the entire discipline, by reducing large unprocessed backlogs or cold case files in some agencies, and overcoming resource limitations (expense and throughput of human latent print examination).

For an AutoID capability to be operationally practical, the resulting AutoID decisions would have to be both accurate (entailing essential no practical risk of an erroneous individualization), and effective (involving a large enough proportion of casework to be worthwhile). AutoID would be implemented using a score threshold — a matcher score above which all candidates would be considered IDs. Assessing the feasibility of AutoID will require evaluating the algorithm and system in question, including both controlled tests (in which latents and exemplars with definitive "ground-truth" associations are tested) and operational evaluation (in which operational casework is monitored, e.g. with special review of high-scoring candidates that examiners conclude are not IDs).

The primary risk associated with AutoID is a remote possibility that erroneous identifications could occur. This risk can be managed in various ways:

- Monitoring of operational data prior to implementation can provide a detailed understanding of what the impact of AutoID would be, without any risk of operational errors.
- Gradual adjustment of decision thresholds can be used so that early AutoID implementation would be cautious and impacts on casework and resources can be thoroughly understood prior to full implementation.

[229] *Derived from [AutoID]*

[230] *[Dvornychenko14]*

- Agencies may choose to implement AutoID differently for different categories of users: agencies may choose to initially implement AutoID only for their own examiners, or set more conservative decision thresholds for examiners external to their agency.
- The implications of an AutoID response could be adapted to minimize risk. For example, early implementations of AutoID could be considered "probable cause," or could require special verification procedures.
- AutoID would still be under agencies' quality control procedures, and therefore the current safeguards against errors would remain in place.

The capacity of agencies to perform casework is currently limited by latent print examiner time, and the amount of casework that the examiners can do in that time. AutoID could increase agency's throughput and capacity, or reduce cost. For example, agencies would be able to handle low-priority and backlog cases that are not practical to process today. If such cases were submitted as image searches by trainees or other non-examiners, then AutoID would mean that this portion of the workflow could be increased significantly without an impact on latent print examiner resources. AutoID could enable rapid responses in order to respond to active incidents immediately, for latent searches sent directly from a crime scene.

As the accuracy of latent AFIS increases, the proportion of searches appropriate for automated identification decisions can be expected to increase. Whether or not automated latent identification is practical in the immediate future will require extensive evaluation, as well as policy decisions — regardless, we should prepare for it to be a reality within the next few years.

8.2.2 Augmenting examiners with probabilistic determinations

A great deal of ongoing research is being conducted on statistical models designed to quantify the probability that a latent print came from a specified source. This work builds on decades of work on the measurement of fingerprint individuality,[231] statistics supporting fingerprint examination,[232] and AFIS algorithms.[233]

I expect that AutoID will be possible for the highest-quality latents in the medium term, but that difficult comparisons will require human examiners far into the future. Therefore, the focus for probabilistic models will be the more difficult comparisons that do not generate extremely high matcher scores on automatically extracted features.

Using AutoID to replace examiners with fully automated feature detection and decisions (Section 8.2) has the notable advantage that AutoID is deterministic: it does not have to rely on examiner markup of features. Probabilistic models that do not automatically extract features are highly sensitive to the interexaminer variability of feature markup, as discussed in the IEVMM study (Section 6.3.4). There are two aspects to this sensitivity:

- Part of the sensitivity is that the models use feature markup in their training: if the features used in training do not account for uncertainty (of presence, of location, of type, and of direction), the training set may be built on biased or otherwise unrepresentative data; the differences between the examiners' markup and the feature markup used in training may affect the accuracy of the model.
- Part of the sensitivity is because the model in any specific instance is based on an individual examiner's markup, but is estimating probabilities based on training across multiple examiners.

[231] *Survey in [Stoney01]*

[232] *Survey in [Neumann13a]*

[233] *[Waman05, Maltoni09]*

For these reasons, probabilistic models are dependent on the standardization of feature markup. I see that enabling probabilistic models is one of the key reasons to promote EFS, and standardized training in feature markup.

8.3 Conclusion

In this thesis I have presented a portfolio of interrelated work conducted over the last decade, focused on the problem of increasing the rigor of latent print examination. The work collectively has sought to improve the transparency, standardization, and quantifiability of the current processes. As we move into the future, those of us who are seeking to enhance the discipline will work to increase automation, improve quality assurance processes, and improve the effectiveness and efficiency of latent print examiners. Our end goal is to make the latent print discipline as accurate and effective a tool as possible for the criminal justice system.

In retrospect, the work has been more successful and had far more of an impact than I initially may have expected, and I would not have done things differently. When Steve Meagher suggested in 1995 that I volunteer to lead the CDEFFS subcommittee to look into an improved fingerprint feature format, I rapidly recognized that this was an opportunity to go far beyond the specific request to make an AFIS transaction format that was more representative of the features by examiners — this was an opportunity to standardize how the content of friction ridge images is defined, with implications throughout the latent print examination process. As EFS was being completed and incorporated as a formal ANSI standard, I worked to bring those implications into fruition: I proposed to NIST the ELFT-EFS tests that evaluated its effectiveness; worked with the FBI to have EFS incorporated into the FBI's Universal Latent Workstation; worked with the FBI to have EFS used as the basis for feature-based searches of the Next Generation Identification system (in EBTS); proposed and developed EFS Markup Instructions, LITS, and EFS Profiles; and proposed and oversaw the EFS Training Tool and ACEware development.

Similarly, the initial impetus for the Quality and Black Box studies was the "Team 8" report [Budowle06], but how those studies were designed and conducted diverged from the loose outlines suggested in that report. For each of the subsequent studies (Black Box Repeatability, Sufficiency for Value, Sufficiency for Individualization, Analysis to Comparison, and Interexaminer Variation in Minutia Markup) we conceptualized each successive study based on the lessons learned to date, conducting the studies that we felt were the most appropriate building blocks in the series to bring greater transparency and rigor to the latent print examination process.

This thesis describes only part of a body of work in progress. This work has certainly been highly sucessful, as recognized (for example) by the Department of Justice's Inspector General [Mayfield11], and PCAST [PCAST16, PCAST17]. However, I look forward to how this work continues into the future, as I work with colleagues and the forensic community to effect the recommendations I have discussed.

References

[A-C] Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (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)

[Abraham13] Abraham J et al (2013) Modern Statistical Models for Forensic Fingerprint Examinations: A Critical Review." Forensic Science International 232, no. 1-3: 131-50.

[ANSI/NIST] 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.

[Anthonioz08] Anthonioz A, Egli N, Champod C, Neumann C, Puch-Solis R, Bromage-Griffiths A (2008) Level 3 Details and Their Role in Fingerprint Identification : A Survey among Practitioners. Journal of Forensic Identification 58(5), pp. 562-589, 2008.

[Ashbaugh99] Ashbaugh D (1999) Quantitative-qualitative friction ridge analysis: an introduction to basic and advanced ridgeology. (CRC Press, New York).

[AssessingLC] Hicklin RA, Buscaglia J, Roberts MA (2013) Assessing the clarity of friction ridge impressions. Forensic Sci Int 226(1):106-117.

[AutoID] Hicklin RA, Ulery BT (2015) Automated decisions for latent fingerprint identification systems. Noblis white paper. (Unpublished)

[Ballou13a] Ballou SM et al (2013) Writing Guidelines for Requests for Proposals for Automated Fingerprint Identification Systems. NIST Special Publication 1155. (http://www.nist.gov/manuscript-publication-search.cfm?pub_id=913324)

[Ballou13b] Ballou SM (2013) Writing Guidelines to Develop an Memorandum of Understanding for Interoperable Automated Fingerprint Identification Systems. NIST Special Publication 1156. (http://www.nist.gov/manuscript-publication-search.cfm?pub_id=913325)

[BB] 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.

[BBRR] Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2012) Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners. PLoS ONE 7:3.

[Biedermann08] Biedermann A (2008) Decision theoretic properties of forensic identification: Underlying logic and argumentative implications. Forensic Science International 177:120-132

[Box87] Box GEP, Draper NR (1987), Empirical Model Building and Response Surfaces, John Wiley & Sons, New York, NY

[Budowle06] Budowle B, Buscaglia J, Schwartz Perlman R (2006) Review of the Scientific Basis for Friction Ridge Comparisons as a Means of Identification: Committee Findings and Recommendations, Forensic Science Communications Vol. 8. (http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/jan2006/research/2006_01_research02.htm)

[Budowle09] Budowle B, et al (2009) A perspective on errors, bias, and interpretation in the forensic sciences and direction for continuing advancement. J. Forensic Sci. 54(4), 798-809.

[Canen02] State of Indiana v. Lana Canen (2002).

[Champod01] Champod C, Evett IW (2001) A Probabilistic Approach to Fingerprint Evidence, J. Forensic Identification 51(2):101-122

[Champod04] Champod C, Lennard CJ, Margot P, Stoilovic M (2004) Fingerprints and Other Ridge Skin Impressions. CRC Press.

[Champod14] Champod C (2014) Research focused mainly on bias will paralyse forensic science. Science & Justice 54(2):107-109.

[Champod16] Champod C, Lennard CJ, Margot P, Stoilovic M (2016) Fingerprints and Other Ridge Skin Impressions, 2nd Edition. CRC Press.

[Champod95] Champod C (1995) Edmond Locard – numerical standards & "probable" identifications. J. Forensic Identification 45(2), 136-155.

[Chatterjee62] Chatterjee SK (1962), Edgeoscopy, Fingerprint and Identification Magazine, 44, 3-13.

[Cole02] Cole SA (2002) Suspect Identities: A History of Fingerprinting and Criminal Identification. Harvard University Press.

[Cole05] Cole SA (2005) More than zero: accounting for error in latent fingerprint identification. J Crim Law Criminol 95(3), 985-1078.

[Cole06] Cole S (2006) Is Fingerprint identification valid? Rhetorics of reliability in fingerprint proponents' discourse, Law & Policy 28(1) 109-135.

[Cole14] Cole SA (2014), Individualization is dead, long live individualization! Reforms of reporting practices for fingerprint analysis in the United States, Law Probab. Risk, 13, 117-150.

[Distortion] Kalka, ND, Hicklin RA (2014) On relative distortion in fingerprint comparison. Forensic Science International 244(2014), 78-84.

[Dror06a] Dror IE, Charlton D, Peron A (2006) Contextual information renders experts vulnerable to making erroneous identifications. Forensic Sci Int. 156(1):74–78.

[Dror06b] Dror IE, Charlton D (2006) Why experts make mistakes. J. Forensic Identification 56(4):600–616.

[Dror10] Dror I, Mnookin J (2010) The use of technology in human expert domains: challenges and risks arising from the use of automated fingerprint identification systems in forensic science. Law, Probability and Risk 9(1), 47-67.

[Dror11a] 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.

[Dror11b] Dror, IE, Champod, C, Langenburg, G, Charlton, D, Hunt, H, & Rosenthal, R. (2011). Cognitive issues in fingerprint analysis: Inter-and intra-expert consistency and the effect of a 'target' comparison. Forensic science international, 208(1): 10-17. http://dx.doi.org/10.1016/j.forsciint.2010.10.013

[Dvornychenko14] Dvornychenko VN, Meagher SB, Garris MD (2014) Characterization of Latent Print Lights-Out Modes for Automated Fingerprint Identification Systems (AFIS). JFI 64(3):255-284

[EFSMI] Chapman W, et al (2013) Markup Instructions for Extended Friction Ridge Features. National Institute of Standards and Technology, Special Publication 1151.

[EFSProfiles] Chapman, et al (2013) Extended Feature Set Profile Specification. NIST Special Publication 1134. http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1134.pdf

[EFSTT] Extended Feature Set Training Tool. Website. http://www.nist.gov/forensics/EFSTrainingTool/

[ELFT-EFS1] Indovina M, Hicklin RA, Kiebuzinski GI (2011) ELFT-EFS Evaluation of Latent Fingerprint Technologies: Extended Feature Sets. National Institute of Standards and Technology Interagency Report #7775.

[ELFT-EFS2] Indovina M, Dvornychenko, V, Hicklin RA, Kiebuzinski GI (2012) ELFT-EFS Evaluation of Latent Fingerprint Technologies: Extended Feature Sets, Evaluation 2. National Institute of Standards and Technology Interagency Report #7859.

[ENFSI15] ENFSI (2015) Best Practice Manual for Fingerprint Examination. ENFSI-BPM-FIN-01, v1, Nov 2015.

[Ester96] Ester, M, Kriegel, H.P., Sander, J. , Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. ISBN 1-57735-004-9. CiteSeerX: 10.1.1.71.1980.

[Evett95] Evett IW, Williams RL (1995) A review of the 16 point fingerprint standard in England and Wales. Fingerprint Whorld 21(82).

[Evett96] 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]

[Farelo12] Farelo A (2012) Fingerprints Survey 2011. 7th International Symposium on Fingerprints, Lyon, France, April 2011 (Conference presentation).

[FBI14] FBI (2014) Offenses Cleared. Uniform Crime Reports: Crime in the United States, 2014

[FBI85] FBI (1985) The Science of Fingerprints: Classification and Uses. U.S. Government Printing Office.

[FBI99] Federal Bureau of Investigation Laboratory (1999) Survey of Law Enforcement Operations in Support of a Daubert Hearing. U.S. v. Mitchell, 365 F.3d 215 (3rd Cir.) (never published)

[Fierrez-Aguilar05] Fierrez-Aguilar J, Ortega-Garcia J, Gonzalez-Rodriguez J, Bigun J (2005) Discriminative multimodal biometric authentication based on quality measures., Pattern Recognition, 38(5):777–779.

[Fierrez-Aguilar06] Fierrez-Aguilar et al (2006) Incorporating Image Quality in Multi-Algorithm Fingerprint Identification. International Conference on Advances in Biometrics (ICB '06).

[Fine06] Fine, G.A. (2006). A review of the FBI's handling of the Brandon Mayfield case. Washington, DC: US Department of Justice Office of the Inspector General.

[FPSourceBook11] National Institute of Justice (2011) Fingerprint Sourcebook. (https://www.ncjrs.gov/pdffiles1/nij/225320.pdf)

[Friedman08] Friedman J, Hastie T, Simon N, Tibshirani R (2008) Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22.

[German05] German E (2005) Legal Challenges to Fingerprints [Website; Last updated 2005; accessed 24 Feb 2016] http://onin.com/fp/daubert_links.html

[German14] German, E. (2014) Problem Idents [Website; accessed 5 June 2015] http://onin.com/fp/problemidents.html

[Grieve96] Grieve DL (1996) Possession of truth. J. Forensic Identification 46 (5), 521-528.

[Gutowski06] Gutowski S (2006) Error rates in fingerprint examination: the view in 2006. The Forensic Bulletin, Autumn 2006, 18-19.

[Haber08] Haber L, Haber RN (2008) Scientific validation of fingerprint evidence under Daubert. Law, Probability and Risk 7(2), 87-109.

[Haber09] Haber RN, L Haber (2009) Challenges to fingerprints. Lawyers & Judges Publishing Company, Tucson AZ.

[Haber14a] Haber R, Haber L (2014) Experimental results of fingerprint comparison validity and reliability: A review and critical analysis. Sci. Justice. http://dx.doi.org/10.1016/j.scijus.2013.08.007

[Haber14b] Haber R, Haber L (2014) Can fingerprint casework accuracy be evaluated by experiments? (Letter to editor) Sci Justice.

[HaberAnnex] Hicklin, et al (2014) Detailed list of errors: Annex to Hicklin, et al., Response to Haber and Haber. http://dx.doi.org/10.1016/jscijus.2014.06.007

[HaberResponse] Hicklin RA, Ulery BT, Buscaglia J, Roberts MA (2014) In response to Haber and Haber, "Experimental results of fingerprint comparison validity and reliability: A review and critical analysis" Science and Justice 54, 390–397

[Hall08] Hall LJ, Player L (2008) Will the introduction of an emotional context affect fingerprint analysis and decision-making? Forensic Sci. Int. 181(1):36–39.

[Havvard01] United States v. Havvard, 117 F. Supp. 2d 848, 849 (S.D. Ind.2000); affd, 260 F.3d 597 (7th Cir. 2001).

[Hicklin02] Hicklin RA, Reedy CL (2002) Implications of the IDENT/IAFIS Image Quality Study for Visa Fingerprint Processing. Noblis Report to USDOJ. (http://www.noblis.org/MissionAreas/nsi/ThoughtLeadership/IdentityDiscovery_Management/Documents/NIST%20IQS%20Final.pdf)

[Hicklin05] Hicklin RA, Meagher S (2005) Extended Fingerprint Feature Set. ANSI/NIST ITL 1-2000 Standard Workshop, 6 December 2005. Briefing. (http://biometrics.nist.gov/cs_links/standard/archived/workshops/workshop2/Presentations-docs/Hicklin-Ext-FP-Features.pdf)

[Hicklin06] Hicklin RA, Khanna R (2006) The Role of Data Quality in Biometric Systems. Noblis technical report to US VISIT. Feb. 1, 2006.

[Hicklin15] Hicklin RA (2015) Best Practices for Universal Latent Workstation. IAI Educational Conference, 6 Aug 2015. (Conference presentation)

[Huber59] Huber RA (1959) Expert witness. Criminal Law Quarterly 3:276–295.

[HumanFactors12] 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)

[IAI11] Polski J, et al (2011) The Report of the International Association for Identification, Standardization II Committee. National Institute of Justice, 233980, March 2011. (http://www.ncjrs.gov/pdffiles1/nij/grants/233980.pdf)

[IAI73] (1973) Report of the Standardization Committee of the International Association for Identification; Identification News; Aug. 1973; p 13. (http://www.latent-prints.com/images/IAI%201973%20Resolution.pdf)

[IEEGFI04] 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

[IEVMM] Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2016) Interexaminer variation of minutia markup on latent fingerprints. Forensic Science International, 264:89–99. (http://dx.doi.org/10.1016/j.forsciint.2016.03.014) Supporting information published separately: Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2016) Data on interexaminer variation of minutia markup on latent fingerprints. Data in Brief, 8:158-190. (http://dx.doi.org/10.1016/j.dib.2016.04.068)

[Indovina09] Indovina M et al (2009) ELFT Phase II - An Evaluation of Automated Latent Fingerprint Identification Technologies. National Institute of Standards and Technology Interagency Report #7577. (http://www.nist.gov/customcf/get_pdf.cfm?pub_id=901870)

[ISO-29794-1] ISO/IEC 29794-1 (2009) Information technology -- Biometric sample quality -- Part 1: Framework. (http://webstore.ansi.org/RecordDetail.aspx?sku=INCITS%2fISO%2fIEC+29794-1-2010)

[ISO-29794-4] ISO/IEC TR 29794-4 (2010) Information technology – Biometric sample quality – Part 4: Finger image data. (http://webstore.ansi.org/RecordDetail.aspx?sku=INCITS%2fISO%2fIEC+29794-1-2010)

[IsraelNP95] Anonymous (1995) Symposium Report – Israel National Police: International Symposium on Fingerprint Detection and Identification. Journal of Forensic Identification, 45(5):578-584.

[Jackson14] Kim Jackson v. State of Florida (2014). Case No. SC13-2090.

[Kassin13] 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)

[Kaye03] Kaye DH (2003) Questioning a Courtroom Proof of the Uniqueness of Fingerprints. International Statistical Review, 71(3):521–533.

[Kennedy03] Kennedy E (2003) Preventing Future Sprees. Washington Post, 17 November 2003.

[Koehler08] Koehler JJ (2008) Fingerprint error rates and proficiency tests: what they are and why they matter. Hastings Law J. 59 (5), 1077-1110.

[Komarinski04] Komarinski P (2004) Automated Fingerprint Identification Systems (AFIS), Elsevier Academic Press.

[Krane08] Krane DE et al (2008) Sequential unmasking: a means of minimizing observer effects in forensic DNA interpretation. J Forens Sci. 53(4): 1006–1007.

[Laird11] Laird A, Lindgren K (2011) Analysis of fingerprints using a color-coding protocol; J. Forensic Identification, 61(2):147-154.

[Langenburg09a] Langenburg G (2009) A performance study of the ACE-V process. J. Forensic Identification 59 (2), 219-257.

[Langenburg09b] Langenburg G, Champod P, Wertheim P (2009) Testing for potential contextual bias effects during the verification stage of the ACE-V methodology when conducting fingerprint comparisons. J Forensic Sci. 54 (3), 571-582.

[Langenburg11] Langenburg G, Champod C (2011) GYRO System-A Recommended Approach to More Transparent Documentation. Journal of Forensic Identification, 61(4):373-384.

[Langenburg12a] Langenburg G (2012) A critical analysis and study of the ACE-V process. Doctoral Thesis, University of Lausanne, Switzerland.

[Langenburg12b] Langenburg G, Champod C, Genessay T. (2012) Informing the judgments of fingerprint analysts using quality metric and statistical assessment tools. Forensic Sci Int 2012 219(1-3):183-98. doi: 10.1016/j.forsciint.2011.12.017. Epub 2012 Jan 24.

[Langenburg14] Langenburg G, Neumann C, Champod C (2014) A comment on experimental results of fingerprint comparison validity and reliability: A review and critical analysis. Letter to the editor. Sci Justice.

[Lee05] Lee B, Moon J, Kim H (2005) Novel Measure of Fingerprint Image Quality using Fourier Spectrum. Proceedings of SPIE Biometric Technology for Human Identification Conference #2, Vol 5779.

[LITS] Chapman, et al (2013) Latent Interoperability Specification. NIST Special Publication 1152. http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1152.pdf

[LleraPlaza02] U.S. v. Llera Plaza, Cr. No. 98-362-10, 11, 12. (E.D. Pa. 2002).

[Locard20] Locard E (1920) L'enquête criminelle et les méthodes scientifiques, Ernest Flammarion, Paris, pp 129-130. (http://babel.hathitrust.org/cgi/pt?id=njp.32101068784956)

[LQSurvey] Hicklin RA, et al (2011) Latent fingerprint quality: a survey of examiners. J. Forensic Identification, 61(4): 385-418.

[Maceo09] Maceo A (2009) Qualitative Assessment of Skin Deformation: A Pilot Study. J. Forensic Ident. 59(4), pp. 390-440.

[Maltoni09] Maltoni D, Maio D, Jain A, Prabhakar S, ed. (2009) Handbook of Fingerprint Recognition. Springer.

[Mayfield06] Office of the Inspector General (2006) A Review of the FBI's Handling of the Brandon Mayfield Case. (US Department of Justice, Washington, DC).

[Mayfield11] 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

[McKie11] Campbell A (2011) The Fingerprint Inquiry. 14 December 2011.

[Mitchell99] U.S. v. Mitchell, No. 96-407 (E.D. Pa. 1999).

[Mnookin08a] Mnookin JL (2008) Of black boxes, instruments, and experts: testing the validity of forensic science. Episteme 5, 343-358.

[Mnookin08b] Mnookin JL (2008) The validity of latent fingerprint identification: confessions of a fingerprinting moderate. Law, Probability and Risk 7(2), 127-141.

[Mnookin10] Mnookin J (2010) The Courts, the NAS, and the Future of Forensic Science. Brooklyn Law Review 75(4)1209-1275.

[NCFS15a] National Commission on Forensic Science (2015) Directive Recommendation: Automated Fingerprint Information Systems (AFIS) Interoperability.Version 2015-0715. (https://www.justice.gov/ncfs/file/641616/download)

[NCFS15b] National Commission On Forensic Science (2015) Ensuring That Forensic Analysis Is Based Upon Task-Relevant Information. (https://www.justice.gov/ncfs/file/795286/download)

[Neumann07] Neumann C, et al (2007) Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. J. Forensic Sci. 52(1), 54-64.

[Neumann10] Neumann, C, Champod, C, Yoo, M, Genessay, T, & Langenburg, G. (2010). Improving the understanding and the reliability of the concept of "sufficiency" in friction ridge examination. National Institute of Justice, Washington DC. https://www.ncjrs.gov/pdffiles1/nij/grants/244231.pdf

[Neumann12] Neumann C, Evett IW, Skerrett J. (2012) Quantifying the Weight of Evidence from a Forensic Fingerprint Comparison: A New Paradigm. J R Stat Soc Ser A Stat Soc Vol. 175 (Part 2), pp 371-415.

[Neumann13a] Neumann C (2013) Statistics and probabilities as a means to support fingerprint examination, in Ramatowski R (ed), Advances in Fingerprint Technology, 3nd Ed.. Boca Raton: CRC Press, pp 419-465.

[Neumann13b] Neumann C, Champod C, Yoo M, Genessay T, Langenburg G (2013) Improving the Understanding and the Reliability of the Concept of "Sufficiency" in Friction Ridge Examination. National Institute of Justice, 12 July 2013. (https://www.ncjrs.gov/pdffiles1/nij/grants/244231.pdf)

[Nill07] Nill NB (2007) IQF (Image Quality of Fingerprint) Software Application. MITRE Technical Report 070053. (https://www.mitre.org/sites/default/files/pdf/07_0580.pdf)

[NIST-AppProfiles] NIST. ANSI/NIST-ITL Standard Profiles and Implementations.Website. (http://www.nist.gov/itl/iad/ig/ansi_standard-profiles.cfm)

[NIST-InteropWG] NIST. Latent Print AFIS Interoperability Working Group. Website. (http://www.nist.gov/oles/afis_interoperability.cfm)

[NIST-MINDTCT] NIST Biometric Image Software (http://www.nist.gov/itl/iad/ig/nbis.cfm)

[NIST-PFT] NIST Proprietary Fingerprint Template (PFT) Evaluation, 2003-2010. Website. (http://www.nist.gov/itl/iad/ig/pft_2003.cfm)

[NIST-PFT2] NIST Proprietary Fingerprint Template (PFT) Evaluation II. Website. (http://www.nist.gov/itl/iad/ig/pftii.cfm)

[NIST-SD27] NIST Special Database 27: Fingerprint Minutiae from Latent and Matching Tenprint Images. (http://www.nist.gov/srd/nistsd27.cfm)

[NIST-Workshops] NIST. ANSI/NIST-ITL Workshops. Website (http://www.nist.gov/itl/iad/ig/ansi_standard-history.cfm#workshops)

[Noblis12] Noblis (2012) Latent Print Interoperability: State and Local Perspectives. Report to NIST OLES. (http://www.noblis.org/media/2caed236-0763-476c-82d5-c9dbe06fca60/docs/Case_Studies_Final_Report_v1_1_2012-04-02_pdf)

[NRC09] National Research Council (2009) Strengthening forensic science in the United States: a path forward. (The National Academies Press, Washington, D.C.).

[NSTC15] National Science And Technology Council Subcommittee on Forensic Science (2015) Achieving Interoperability For Latent Fingerprint Identification in the United States. April 2015. (https://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/latent_fingerprint_report_may_2015.pdf)

[Pankanti02] Pankanti S, Prabhakar S, Jain AK (2002) On the Individuality of Fingerprints. IEEE Trans Pattern Anal Mach Intell, 24 (8):1010-1025.

[PCAST16] Executive Office of the President, President's Council of Advisors on Science and Technology (2016) Report to the President. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. September 2016. (https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf)

[PCAST17] President's Council of Advisors on Science and Technology (2017) An Addendum to the PCAST Report On Forensic Science in Criminal Courts. January 2017.

[Peterson10] Peterson J, Sommers I, Baskin D, Johnson D (2010) The Role and Impact of Forensic Evidence in the Criminal Justice Process. National Institute of Justice 2006-DN-BX-0094, September 2010.

[PiAnoS15] Université de Lausanne (2015) User Instructions for PiAnoS 4.

[Richmond04] Richmond (2004) Do fingerprint ridges and characteristics within ridges change with pressure?," Australian Federal Police Forensic Services. <http://www.latent-prints.com/images/changes%20with%20pressure.pdf>

[Riemen12] Riemen J (2012) Netherlands case study: auto encoding of latents. 7th International Interpol Symposium on Fingerprints. Conference presentation.

[Risinger02] Risinger DM, Saks MJ, Thompson WC, Rosenthal R (2002) The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. Cal L Rev. 90:1–56.

[Risinger14] Risinger DM et al (2014) Regarding Champod, editorial: "Research focused mainly on bias will paralyse forensic science." Science & Justice 54(6):508-509.

[Roman08] Roman JK, et al (2008) The DNA Field Experiment: Cost-Effectiveness Analysis of the Use of DNA in the Investigation of High-Volume Crimes. Urban Institute. April 2008.

[Rose07] Maryland v Rose, No. K06-0545 (MD Cir. Ct. 2007).

[Saks03] Saks MJ, Risinger DM, Rosenthal R, Thompson WC (2003) Context effects in forensic science. Sci & Just. 43(2): 77–90.

[Saks05] Saks M, Koehler J (2005) The coming paradigm shift in forensic identification science. Science 309, 892-895.

[Scarborough] Scarborough S, York R, Wertheim K. Daubert Card. Website, accessed 28 Feb 2016. (http://www.clpex.com/daubertcard.htm)

[Stoney01] Stoney DA (2001) Measurement of Fingerprint Individuality, in Lee HC, Gaensslen RE (eds), Advances in Fingerprint Technology, 2nd Ed. Boca Raton: CRC Press, pp 327-387.

[Stoney10] Stoney DA (2010) Fingerprint identification, in: D.L. Faigman, M.J. Saks, J. Sanders, E.K. Cheng (Eds.), Modern Scientific Evidence: The Law and Science of Expert Testimony, vol. 4, Thomson-West, St. Paul, MN, 2010, pp. 337–449.

[Stoney91] Stoney DA (1991) What made us ever think we could individualize using statistics? J. Forensic Science Society 31(2), 197-199.

[Su09] Su C, Srihari S (2009) Probability of Random Correspondence for Fingerprints, in Computational Forensics: Third International Workshop, IWCF 2009. (Geradts MH, Franke KY, Veenman CJ, eds) Berlin-Heidelberg:Springer Verlag, pp 55-66.

[Su10] Su C, Srihari S (2010) Evaluation of rarity of fingerprints in forensics. In Advances in Neural Information Processing Systems, edited by J. Lafferty J, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel and A. Culotta, 1207-15 2010. (http://www.cedar.buffalo.edu/~srihari/papers/nips2010.pdf)

[SuffID] 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)

[SuffValue] 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.

[SWGFAST-Conclusions13] 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)

[SWGFAST-Doc12] SWGFAST (2012) Standard for the Documentation of Analysis, Comparison, Evaluation, and Verification (ACE-V), Ver. 2.0. (http://swgfast.org/documents/documentation/121124_Standard-Documentation-ACE-V_2.0.pdf)

[SWGFAST-ErrorRates12] SWGFAST (2012) Standard for the Definition and Measurement of Rates of Errors and Inappropriate Decisions in Friction Ridge Examination (Latent/Tenprint) Version 2.0. (http://swgfast.org/documents/error/121124_Rates-of-Error_2.0.pdf)

[SWGFAST-ExamMethod02] SWGFAST (2002) Friction ridge examination methodology for latent print examiners. (v.1.01; swgfast.org).

[SWGFAST-Glossary09] SWGFAST (2009) Glossary. Ver.2 (swgfast.org).

[SWGFAST-ID12] SWGFAST (2012) Individualization / Identification Position Statement, Version 1.0. (http://swgfast.org/Comments-Positions/120306_Individualization-Identification.pdf)

[SWGFAST-Memo05] SWGFAST (2005). Memo to Mike McCabe (NIST) regarding ANSI/NIST ITL 1-2000. (http://fingerprint.nist.gov/standard/cdeffs/Docs/SWGFAST_Memo.pdf)

[SWGFAST-StdExam13] 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

[SWGFAST-Symbols13] SWGFAST (2013) Uniform Automated Fingerprint Identification System (AFIS) Feature Symbols Position Statement. (http://swgfast.org/Comments-Positions/130427_Uniform_AFIS_Feature.pdf)

[SWGFAST-Terminology11] SWGFAST (2011) Standard terminology of friction ridge examination, Version 3.0. (http://swgfast.org/documents/terminology/110323_Standard-Terminology_3.0.pdf)

[Swofford13] 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.

[Tabassi04] Tabassi E, Wilson CL, Watson CI (2004) Fingerprint Image Quality. NIST Interagency Report 7151; National Institute of Standards and Technology. (ftp://sequoyah.nist.gov/pub/nist_internal_reports/ir_7151/ir_7151.pdf)

[Tangen11] Tangen JM, Thompson MB, McCarthy DJ (2011) Identifying fingerprint expertise. Psychol. Sci. 22(8):995–997.

[Thompson11] Thompson WC (2011) What role should investigative facts play in the evaluation of scientific evidence? Australian Journal of Forensic Sciences 43(2-3):123–134. (https://ucconsortiumssl.files.wordpress.com/2015/08/thompson-ajfs-final.pdf)

[Thompson13] Thompson MB, Tangen JM, McCarthy DJ (2013) Expertise in Fingerprint Identification. J Forensic Sci 56(6):1519–1530.

[Thompson14] Thompson MB, Tangen JM (2014) Generalization in fingerprint matching experiments. Letter to editor. Sci Justice. 391-392

[UK-FSR11] Forensic Science Regulator (2011). Developing a quality standard for fingerprint examination. Report. Dec. 2011. https://www.gov.uk/government/publications/fingerprint-examination-developing-a-quality-standard

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

[Vanderkolk04] Vanderkolk J (2004) "ACE+V: A Model". Journal of Forensic Identification, 54(1):45-52

[Vanderkolk09] Vanderkolk J (2009) Forensic Comparative Science: Qualitative Quantitative Source Determination of Unique Impressions, Images, and Objects. Academic Press.

[Wayman00] Wayman JL (2000) When Bad Science Leads to Good Law: The Disturbing Irony of the Daubert Hearing in the Case of U.S. v. Byron C. Mitchell. Biometrics Publications, 2 Feb 2000. (http://www.nlada.org/forensics/for_lib/Documents/1048198555.79/publications_daubert.html)

[Wayman05] Wayman J, Jain A, Maltoi D, Maio D, ed. (2005) Biometric Systems: Technology, Design and Performance Evaluation. Springer.

[Wertheim06] Wertheim K, Langenburg G, Moenssens A (2006) A report of latent print examiner accuracy during comparison training exercises. J. Forensic Identification 56 (1), 55-93.

[Wilson04] Wilson C, et al (2004) Fingerprint Vendor Technology Evaluation 2003. National Institute of Standards and Technology Interagency Report #7775. (http://biometrics.nist.gov/cs_links/latent/elft-efs/NISTIR_7775.pdf) [Note I was technical lead on this study]

[Wright97] Texas vs Gregory Edward Wright (1997). Trial Cause No. F97-01215-PJ. (http://www.freegregwright.com/NewVol47.pdf)

[Yen06] Yen R (2006) Human Visual Perception Model for Measuring Fingerprint Image Quality. NIST Biometric Quality Workshop, 8 Mar 2006. Briefing. (http://biometrics.nist.gov/cs_links/quality/workshopI/proc/yen_fiqm_for_nist_workshop.pdf)

[Yoon12] Yoon S, Liu E, Jain AK (2012) On Latent Fingerprint Image Quality", International Workshop on Computational Forensics (IWCF), Tsukuba, Japan, November 11, 2012.

[Yoon13] Yoon S, Cao K, Liu E, Jain AK (2013) LFIQ: Latent Fingerprint Image Quality", International Conference on Biometrics: Theory, Applications and Systems (BTAS), Washington, D.C., September 29 - October 2, 2013.

[Zabell05] Zabell SL (2005) Fingerprint Evidence. Journal of Law and Policy 13(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.
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. Starting in 2011, this incorporated the Extended Feature Set (EFS) definition of friction ridge features.[234]
ASYS	The Analysis (ASYS) transaction provides a means to provide detailed markup and annotation for a single impression that is not associated with other prints. ASYS files are ANSI/NIST files, with the transaction defined in LITS.
AutoID	Automated identification determinations returned by an AFIS without markup by a human latent print examiner.
Candidate	Used to describe an impression (generally an exemplar) that might possibly be individualized against a given latent. Most frequently used to describe the list of exemplars returned by an AFIS in response to an AFIS search.
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 utility, quantity, or distinctiveness of features.)
COMP file	The Comparison (COMP) transaction provides a standard format for two or more friction ridge images, feature markup, and determinations. COMP files are ANSI/NIST files, with the transaction defined in LITS.
Comparison/Evaluation phase	The second and third phases of the ACE-V method. In this test, there was no procedural demarcation between the Comparison and Evaluation phases of the ACE-V method; hence, this refers to the single combined phase during which both images were presented side-by-side.
Comparison determination	The determination of individualization, exclusion, or inconclusive reached in the Comparison/Evaluation phase of ACE-V. SWGFAST refers to this determination as the Evaluation Conclusion;[235] here, I use "conclusion" to refer only to individualization or exclusion, as I consider an "inconclusive conclusion" to be an oxymoron.

[234] *[ANSI/NIST]*

[235] *[SWGFAST-Conclusions13]*

Conflict resolution	The process conducted when there is a difference of determinations or conclusions between examiners, generally when the initial examiner and verifier disagree.
Corresponding clarity map	The corresponding clarity map represents the minimum clarity at each location in the aligned latent and exemplar clarity maps, as described in [[236]]. These maps were constructed from the examiners' annotations by post-processing software whenever at least three corresponding features were marked by the examiner. A thin-plate spline algorithm was used to align the latent and exemplar prints. (See Local clarity map)
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.
CWE	The Casework Exchange (CWE) transaction provides a format for latent examiners to collect all information related to a case within a single transaction. CWE files are ANSI/NIST files, with the transaction defined in LITS.
Debatable correspondence	A relationship between a feature in a latent and a feature in the exemplar in which there is an apparent correspondence between a feature in the latent and a feature in the exemplar that does not rise to the threshold of definite correspondence. (Not to be confused with debatable ridge flow or debatable features, which were indicated by painting the image clarity.)
Determination	The result of an examiner's decision: the Analysis phase results in a Value determination, and the Comparison/Evaluation phase results in a Comparison determination.
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.
GBU	Informal "Good, Bad, Ugly" scale for assessing latent print quality.
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.

[236] *[AssessingLC]*

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."[237]
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.[238] Outside of North America, an impression from an unknown source (latent) is often described as a "mark" or "trace," and "print" is often used to refer only to known impressions (exemplars).
Level-3 detail	Friction ridge dimensional attributes such as width, edge shapes, and pores.
LITS	The Latent Interoperability Transmission Specification (LITS) is an application profile that builds upon ANSI/NIST and EBTS. LITS defines AFIS transactions for exchange among state and local law enforcement agencies, as well as transactions for non-AFIS casework exchange and archiving.
Local clarity map	A color-coded annotation of a friction ridge image indicating the clarity for every location in the print, as described in [AssessingLC] and defined in [ANSI/NIST].
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 (or image pair, for median corresponding clarity maps).
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.
Misclassification rate	The proportion of responses that would be incorrectly classified as individualization or not individualization for a given model.
Missed ID	Failure by an examiner to individualize a mated pair that was individualized by any other examiners (also known as a "missed individualization" or "missed identification").
NGI	The FBI's Next Generation Identification system, a multi-modal ABIS.[239]
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).
NRC	National Research Council of the National Academies.

[237] *[SWGFAST-ID12]*

[238] *[SWGFAST-Terminology11]*

[239] *https://www.fbi.gov/about-us/cjis/fingerprints_biometrics/ngi*

NV (No value)	The impression is not of value for individualization and contains no usable friction ridge information. See also VEO and VID.
OSAC	Organization of Scientific Area Committees[240]
Other point	In the White Box 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).
Overall Clarity	A metric based on the size and consistency of the areas of the various levels of clarity in a local clarity map (c.f.). Overall Clarity ranges from 0-100 and was developed to correspond to human examiner assessments of the value and difficulty of an image.
Qualified examiner	Determined by an agency to be appropriately qualified as a latent print examiner. Used instead of "certified" in some organizations to differentiate from the IAI certification, "Certified Latent Print Examiner."
Quality	General concept referring to the clarity and utility of a latent
Reliability	Consistency of results, here differentiated into repeatability (c.f.) and reproducibility (c.f.)
Repeatability	Intraexaminer agreement: when one examiner provides the same response (annotation or determination) to a stimulus (image or image pair) on multiple occasions.
Reproducibility	Interexaminer agreement: when multiple examiners provide the same response (annotation or determination) to a stimulus (image or image pair).
Source	An area of friction ridge skin from which an impression is left. Two impressions are said to be from the "same source" when they have in common a region of overlapping friction ridge skin.
Sufficient	An examiner's assessment that the quality and quantity of information in a print (or image pair) justifies a specific determination (especially used with respect to individualization).
SWGFAST	Scientific Working Group on Friction Ridge Analysis, Study and Technology [241]
Transparency (of a process)	The extent to which the actions and inner workings of a process are visible and accessible.
ULW	The FBI's Universal Latent Workstation software.[242]
Unassociated feature	In the White Box study, a feature marked in one print for which the examiner did not indicate any level of correspondence or non-correspondence with respect to the other print (often either obscured or outside the corresponding area).
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 [243]].
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.

[240] *http://www.nist.gov/forensics/osac/*

[241] *http://swgfast.org*

[242] *[ULW]*

[243] *[BB]*

Verification	The final phase of ACE-V: the independent application of the ACE process by a subsequent examiner to either support or refute the conclusions of the original examiner. Not addressed in this study.
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.