Year : 2022

Development and Implementation of Computational Algorithms in Forensic Fingermark Examination

Swofford Henry

Swofford Henry, 2022, Development and Implementation of Computational Algorithms in Forensic Fingermark Examination

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://serval.unil.ch
Document URN : urn:nbn:ch:serval-BIB_6FA33720F0F80

UNIL | Université de Lausanne
Ecole des Sciences criminelles

Development and Implementation of Computational Algorithms in Forensic Fingermark Examination

Thèse de doctorat

Présentée à la
Faculté de droit, des sciences criminelles et d'administration publique
de l'Université de Lausanne par

Henry Swofford

Jury:
Prof. Christophe Champod, directeur de thèse (Université de Lausanne)
Prof. Simon Cole, expert externe (University of California, Irvine)
Prof. Gillian Tully, experte externe (King's College, London)
Prof. Alex Biedermann, expert interne (Université de Lausanne)
Prof. Pierre Margot, expert interne (Université de Lausanne)

Sous la présidence du Prof. Franco Taroni (Université de Lausanne)

Lausanne 2022

UNIL | Université de Lausanne
Ecole des Sciences criminelles

Development and Implementation of Computational Algorithms in Forensic Fingermark Examination

Thèse de doctorat

Présentée à la
Faculté de droit, des sciences criminelles et d'administration publique
de l'Université de Lausanne par

Henry Swofford

Jury:
Prof. Christophe Champod, directeur de thèse (Université de Lausanne)
Prof. Simon Cole, expert externe (University of California, Irvine)
Prof. Gillian Tully, experte externe (King's College, London)
Prof. Alex Biedermann, expert interne (Université de Lausanne)
Prof. Pierre Margot, expert interne (Université de Lausanne)

Sous la présidence du Prof. Franco Taroni (Université de Lausanne)

Lausanne 2022

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. Henry Swofford, candidat au doctorat en science forensique, intitulée

« Development and Implementation of Computational Algorithms in Forensic Fingermark Examination »

Le Président du Jury

Prof. Franco Taroni

Lausanne, le 5 juillet 2022

Acknowledgements

This thesis is the result of countless people who have contributed, both directly and indirectly, to individual research efforts and who have supported me through this journey.

First and foremost, I am deeply indebted to my wife, Daneille, for her love, support, encouragement, and sacrifice over the years; to Olivia, our precious daughter, who inspired me to keep plugging away at the research; to Trevor, our loyal Pomeranian Poodle, who patiently sat by my side through this entire journey; and to my parents, Ron and Joyce, for instilling a sense of intellectual curiosity at such a young age.

I extend a special thanks to Professor Christophe Champod for making this possible, and for his insights, support, and countless hours of engaging discussions related to this work. I also thank the other members of my committee—Professor Franco Taroni, Professor Pierre Margo, Professor Alex Biedermann, Professor Gill Tully, and Dr. Simon Cole—for their review and thoughtful guidance and suggestions.

I acknowledge the several other co-authors who participated in this research. Without their involvement, this work would not have been possible: Anthony (Tony) Koertner, Heidi Eldridge, Michael (Jeff) Salyards, Fabian Zemp, Alice Liu, Madeline Ausdemore, Valerie King.

I recognize the several other colleagues who contributed to or otherwise supported these efforts. Without their assistance, this work would not have been as successful as it was: Jessica LeCroy, Jeremy John, Kalisha Gill, Shauna Steffan, Lisa Carson, Monika Garcia, Amaliya Kovalchick, Monica Kupsco, Molly Hall, Tom Wortman, Juliet Wood, Marvin Stancil, Tonya Johnson, Rodney Schenck, Christine Swanson, Amanda Atkins, Gary Ford, Pat Wertheim, Henry Maynard, Garold Warner, Karen Kafadar, Hal Stern, Alicia Carriquery, Hariharian Iyer, Steve Lund, Simone Gittelson, Matthew Bohn, Kate Schilling, McKay Allred, Tim Kalafut, John Buckelton, Marco De Donno, Lauren Reed, Debra Glidewell, Rick Tontarski, Mike Hill, Anece Baxter-White, Randall Bentley, Mandi Hornickel, Matt Barno, Brandon Garrett, Jessica Cino, Manon Jendly, Glenn Langenburg, Eric Ray, Michele Triplett, Sandra Siegel, Tommy Jones, Allison Miller, Sarah Chu.

Finally, I also thank the numerous individuals that anonymously contributed to the research through participation in the surveys and interviews for their time, contribution, and candid responses, as well as the individuals that helped raise awareness and distribute the research throughout the community.

Abstract

This PhD thesis describes the development and implementation of computational algorithms in forensic friction ridge examination. The thesis is separated into two parts, reflecting the major objectives of the research. Part I describes the design, development, and validation of two different publicly accessible algorithmic tools—DFIQI and FRStat—that enable examiners to practically apply statistical measures to friction ridge impression evidence and explore more objective interpretation schemes. Part II explores practitioner and stakeholder perspectives on issues related to the adoption and implementation of algorithmic tools into practice; discusses salient challenges, considerations, and a path forward related to the implementation of algorithms in domains largely dominated by human judgment; and describes details surrounding the actual implementation of an algorithmic tool described in Part I (FRStat) into operational practice at a federal forensic laboratory in the United States and subsequent litigation involving its use. The work presented in this thesis has broad impact and implications—both theoretical and practical, ranging from statistics and evidence quantification to social psychology and human behavior—affecting policy, procedure, training, quality assurance, research, reporting and testimony, and litigation as it relates to the operational implementation and use of algorithms in friction ridge examination and forensic science more broadly. The availability of the tools presented in Part I and results of the discussions and proposed framework presented in Part II support eight key recommendations to strengthen the foundations of friction ridge examination and improve our understanding of the reliability of evidence that our nations' legal system depends on.

Résumé

Cette thèse de doctorat présente le développement et la mise en œuvre d'algorithmes venant en soutien à l'examen forensique des crêtes papillaires. Elle est divisée en deux parties, reflétant les principaux objectifs de la recherche. La partie I décrit la conception, le développement et la validation de deux outils algorithmiques accessibles au public -DFIQI et FRStat- qui permettent aux examinateurs d'appliquer de manière pratique des mesures statistiques reflétant la qualité des traces papillaires et des comparaisons et permettant d'exploiter des schémas d'interprétation plus objectifs. La deuxième partie explore les perspectives des praticien-ne-s et des parties prenantes sur les questions liées à l'adoption et à la mise en œuvre d'outils algorithmiques dans la pratique ; elle analyse les principaux défis et considérations liés à la mise en œuvre d'algorithmes dans un domaine largement dominé par le jugement humain. Dans cette partie également sont décrits les détails entourant la mise en œuvre opérationnelle dans un laboratoire forensique fédéral aux États-Unis d'un des outils algorithmiques développés dans la première partie (FRStat) et les débats liés à son utilisation. Les travaux présentés dans cette thèse ont un impact et des implications à la fois théoriques et pratiques, allant de l'usage de la statistique et de la quantification des preuves à la psychologie sociale et au comportement humain. Ils ont une incidence sur les politiques, les procédures, la formation, l'assurance qualité, la recherche, les rapports et les témoignages, ainsi que sur les litiges liés à la mise en œuvre opérationnelle et à l'utilisation d'algorithmes dans l'examen des crêtes papillaires et, plus généralement, dans la science forensique. La disponibilité des outils présentés dans la partie I et les résultats des discussions et du cadre d'usage présenté dans la partie II amènent à proposer huit recommandations clés visant à renforcer les fondements de l'examen dactyloscopique et à améliorer notre compréhension de la fiabilité des preuves dont dépend le système juridique de nos nations.

Table of Contents

1 Introduction 1

1.1 Background to Friction Ridge Examination 2

1.1.1 Analysis 3

1.1.2 Comparison 4

1.1.3 Evaluation 4

1.1.4 Verification 4

1.2 Discussion of the Problem 5

1.3 Objectives of the research 11

1.4 Structure of the thesis 11

2 Quality Assessment Software (DFIQI) 15

2.1 Method Development and Validation 15

2.1.1 Abstract 15

2.1.2 Introduction 16

2.1.3 Materials & Methods 19

2.1.4 Results & Discussion 30

2.1.5 Conclusion 42

2.2 Comparison with Other Methods 43

2.2.1 Background 44

2.2.2 Materials & Methods 44

2.2.3 Results & Discussion 46

3 Statistical Interpretation Software (FRStat) 58

3.1 Method Development and Validation 58

3.1.1 Abstract 58

3.1.2 Introduction 59

3.1.3 Materials & Methods 60

3.1.4 Results & Discussion 72

3.1.5 Conclusion 83

3.2 Comparison with Other Methods 85

3.2.1 Background 85

3.2.2 Materials & Methods 86

3.2.3 Results & Discussion 89

4 Toward Objectivity: Integrating Algorithmic Outputs 108
4.1 Background 108
4.2 Materials & Methods 110
4.3 Results & Discussion 111
5 Evaluation of Practitioners' Perspectives 113
5.1 Abstract 113
5.2 Introduction 114
5.3 Background 115
5.4 Methods 117
5.5 Results 122
5.6 Discussion 135
5.7 Conclusion 140
6 Evaluation of Stakeholders' Perspectives 144
6.1 Abstract 144
6.2 Introduction 144
6.3 Materials & Methods 147
6.4 Results 152
6.4.1 Laboratory Managers 152
6.4.2 Prosecutors 158
6.4.3 Defense Attorneys 163
6.4.4 Judges 169
6.4.5 Other (Academic Scholars) 176
6.5 Discussion 183
6.5.1 Interpretation & Reporting Practices 184
6.5.2 Use of Algorithms 192
6.6 Conclusion 197
7 Implementation of Algorithms: Challenges, Considerations, and a Path Forward 200
7.1 Abstract 200
7.2 Introduction 200
7.3 Part I: The Introduction of Algorithms in Clinical Decision Making 203
7.4 Part II: Human-Algorithm Interaction in Laboratory Studies 205
7.5 Part III: Human-Algorithm Interaction in Real World Domains 209
7.6 Part IV: Algorithms and the American Legal System 213

7.7 Part V: A Path Forward for Forensic Science 220
7.8 Conclusion 231
8 Operationalization of Algorithms: Personal Reflections and Observations 233
8.1 Background 233
8.2 Implementation 235
8.3 Policies and Procedures 241
8.3.1 Background 241
8.3.2 Key Questions and Considerations 242
8.3.3 Example Workflow with FRStat 254
8.3.4 Example Report Phrasing with FRStat Results 256
8.4 Litigation (case study) 256
8.4.1 Case Background 257
8.4.2 Fingermark Evidence 258
8.4.3 FRStat 260
9 Looking Forward: Impact and Recommendations 271
9.1 Impact 271
9.2 Recommendations 281
9.2.1 Better algorithms should be developed and made accessible to all stakeholders 282
9.2.2 Algorithms should be regulated by an independent authority 282
9.2.3 Standards specifying minimum requirements for implementing algorithms operationally should be established. 283
9.2.4 Standards specifying minimum educational requirements for forensic practitioners should be expanded to include a more rigorous emphasis on scientific interpretation. 283
9.2.5 Reported results should be scientifically defensible and expressed with clear characterizations of their limitations 284
9.2.6 Analysts' opinions should be distinguished from reported conclusions 285
9.2.7 Algorithms should be implemented operationally 285
9.2.8 Examination, interpretation, and reporting practices should be governed by centralized policy and oversight 286
10 Conclusion 287
11 References 292
12 Appendix A: Glossary of Acronyms and Terms 306
13 Appendix B: Supplemental Material for Chapter 2 311
13.1 Appendix B-1 311

14 Appendix C: Supplemental Material for Chapter 3 315
14.1 Appendix C-1 315
14.2 Appendix C-2 321
14.3 Appendix C-3 324
14.4 Appendix C-4 328
14.5 Appendix C-5 337
15 Appendix D: Supplemental Material for Chapter 5 339
15.1 Appendix D-1 339
15.2 Appendix D-2 339
15.3 Appendix D-3 340
16 Appendix E: Supplemental Material for Chapter 6 343
16.1 Appendix E-1 343
16.2 Appendix E-2 344
16.3 Appendix E-3 346
16.4 Appendix E-4 348
16.5 Appendix E-5 349
16.6 Appendix E-6 377

1 Introduction

Friction ridge examination is ubiquitously practiced by forensic laboratories throughout the world and is often presented as incontrovertible evidence that an individual touched an item or was present at the scene of a crime. Despite being first introduced in the late 1800s and operationalized in the early 1900s, the practice of performing the examinations have remained nearly the same for well over a century and rely on a visual comparative methodology often referred to in more contemporary times as "ACE-V," an acronym for "Analysis," "Comparison," "Evaluation," and "Verification."

In general, when conducting an examination, the analyst first "analyzes" the impression for discriminating features and determines the value and quality of the mark. If the mark is deemed "of value" in the analysis phase, the examiner will "compare" the mark to a print from a known source and "evaluate" the significance[1] of observed similarities and differences between the mark and known source. Based on the results of the evaluation, the analyst will traditionally either conclude "identification" (the two impressions were made by the same source), "exclusion" (the two impressions were made by different sources), or "inconclusive" (insufficient similarities or differences to determine if the impressions were made by the same source or different sources). Once the analyst has rendered her conclusion, another examiner then verifies this determination by conducting an independent examination.

Throughout the examination process, the analyst is responsible for making a number of assessments. These assessments and the ultimate conclusions rendered are not based on empirical statistical measurements or clearly defined standards; rather, they are subjective determinations made by the analyst on a case-by-case basis and depend on her experience and personal confidence in the conclusion. Without statistical measurements or an empirical basis for which the significance of the evidence[2] is evaluated, it is unclear *what* contributed to the overall assessment of the evidential strength and *how* it was evaluated. Consequently, assessments made during friction ridge examinations are susceptible to variation from one analyst to another as well as by the same analyst from one examination to another. When considering borderline impressions with marginal quality or quantity of features, these variations often result in differences in the overall conclusion. In the broad spectrum, however, while the lack of empirical standards and measurements do not necessarily suggest the practice as a whole is unreliable or fraught with error, it does raise questions as to how reliable the evidence is for the case at hand; thus, there is a critical need for the friction ridge community to move towards integrating an empirical foundation using quantitative and statistical methods (through algorithmic tools) into their examination methodology.

Over the last several years, the friction ridge community has faced increasing criticism by scientific and legal commentators, challenging the validity and reliability of the ACE-V method,

[1] The term "significance" used throughout refers to the importance, weight of the observations, or their ability to discriminate impressions originating from common sources versus impressions originating from different sources. It should not be confused with the use of the term in classic frequentist hypothesis testing.

[2] The term "evidence" used throughout refers to the findings and information from the examination of physical items submitted in a particular case. This is consistent with colloquial usage of the term "evidence" in the American legal system; however, it is recognized that technically speaking, whether such findings or information are deemed as "evidence" that has probative value in specific litigation is ultimately within the purview of the court.

which relies heavily on the subjective interpretation of forensic practitioners [1-8]. Of particular concern, noted in 2009 by the National Research Council (NRC) of the National Academies of Science (NAS) [3] as well as the President's Council of Advisors on Science and Technology (PCAST) in 2016 [7] and the American Academy for the Advancement of Science (AAAS) in 2017 [8], is the lack of an empirically demonstrable basis to substantiate conclusions from friction ridge examinations, thus limiting the ability for the judiciary to reasonably understand the reliability of the expert's testimony for a given case. Along with several academic commentators, the NRC, PCAST, and AAAS strongly encourage the forensic science community to develop tools to evaluate and report the strength of forensic evidence using validated statistical methods [3, 7-9].

In an effort to strengthen the foundations of friction ridge examination, the objectives of this thesis are twofold: (1) to develop, validate, and make publicly accessible friction ridge examination statistical software tools capable of (a) assessing the clarity of friction ridge skin features and overall quality of impressions and (b) evaluating the statistical strength of correspondence between two impressions, and (2) to develop strategies for practical application and implementation of these (and similar) tools in an operational forensic science laboratory. Taken together, this research will not only provide the friction ridge community access to novel, validated algorithmic tools to evaluate the significance of friction ridge examinations, but it will also provide a better understanding of how to practically apply and implement algorithmic tools in an operational forensic science laboratory. As a result, this work has the aim and potential to promote meaningful reform and transform the way forensic fingermark impressions are evaluated, interpreted, and reported throughout the United States and the international forensic science community.

1.1 Background to Friction Ridge Examination

Friction ridge examinations are conducted by human analysts using a visual comparative methodology often referred to as "ACE-V." Examinations are based on the ridge flow, ridge events or features, and ridge structures of friction ridge skin impressions. Of particular importance are the location, orientation, type, and spatial relationships of ridge features when making a determination of suitability for comparison or whether two impressions could have originated from a common source.

For over a century, the friction ridge community has relied upon two broad premises, or tenets, as the foundational principles supporting the use of friction ridge impressions as a means of personal identification:

(1) Persistence – the morphological structure of friction ridge skin is formed before birth and, barring scarring or disease, does not change in a significant manner until after death.

(2) Uniqueness – the morphological structure of friction ridge skin bears a complex and unique pattern of ridges which are highly discriminating between different individuals.

Generally speaking, the friction ridge community subscribes to three different levels of friction ridge skin detail that are used for examinations. Level 1 detail refers to the overall ridge flow and pattern type. Level 2 detail refers to the individual friction ridge features, such as bifurcations, ending ridges, and dots, and their relative arrangement among one another. Level 3 detail refers to ridge structures, such as edge shapes and pores, and their relative arrangement among one another. Non-ridge events that impact the appearance of friction ridges, such as creases, scars, warts, incipient ridges, and other features may have aspects that are reflected in all three levels of detail and are also used for examination purposes.

Friction ridge skin impressions are imperfect representations of the morphological structures of the ridges on the skin. A variety of factors, such as deposition, substrate, matrix, development, and environmental conditions often have a degradative effect on impressions thereby limiting the legible attributes available for examination. As a result, the friction ridge community invests a considerable amount of effort to train analysts on how to properly detect and interpret complex impressions. In the following sections, an overview of the individual steps of the ACE-V methodology are discussed along with a critical analysis of major gaps in the methodology which expose vulnerabilities in the foundation of friction ridge examination.

1.1.1 Analysis

During the "analysis" phase, the analyst visually inspects the questioned impression to detect the qualitative and quantitative attributes[3] of the impression and determine whether the impression is "suitable" for a specified purpose (such as identification or exclusion). To this end, the analyst is particularly concerned with identifying discriminating attributes of the friction ridge detail which may be used for comparison and evaluation against the prints of a known source impression. The ability for the analyst to reliably detect these attributes depends heavily on the clarity of the impression. Generally, as the clarity of an impression increases, the confidence analysts' have in their interpretation of the location, orientation, type, and spatial arrangement of features also increases. Additionally, as the number of interpretable features increases, the discriminating strength of the impression is considered to increase as well. Once the features have been detected, the analyst will assess the overall quality of the impression and make an experience-based determination of the "suitability" for further comparison and evaluation [10]. This determination is often not based on an empirical standard or a specific measure; rather, it is a personal determination made by the analyst on a case-by-case basis and depends on whether the analyst judges that the quality of the impression is sufficient to compare to a known source and render a particular conclusion regarding the potential source of the impression.

[3] Quantitative attributes of the impression refer to the number of features in the impression. Qualitative attributes refer to the clarity, type, location, orientation, and spatial arrangement of features in the impression, as well as other details that indicate the orientation and anatomical location of the source of friction ridge skin that made the impression (e.g., tip of a finger, hypothenar of a palm).

1.1.2 Comparison

Once the overall qualitative and quantitative attributes of the impression have been assessed and the analyst has determined the impression to be "suitable," then he or she will proceed to the "comparison" phase. In the comparison phase, the analyst will visually compare the features detected in the impression against those available in the standard to determine if the features are in agreement with one another – or in other words, whether the features are similar enough to be considered "sufficient" and therefore included as a possible source. Because there are factors which may impact the appearance of features, criteria for "sufficient" agreement may allow for differences in appearance depending on various deposition conditions of the impression [10]. Like in the analysis phase, this determination is not based on an empirical standard or specific measure; rather it is a personal determination made by the analyst on a case-by-case basis and depends on whether the analyst judges that the conditions of the impression are such that the observed differences between the impressions are within the typical range of variation that is possible from impressions originating from the same source.

1.1.3 Evaluation

Once the analyst has compared each feature detected in the questioned impression to the features present (or absent) in the known source impression, he or she will proceed to the "evaluation" phase. In the "evaluation" phase, the analyst considers the strength of the correspondence or discordance, or in other words, the likelihood the questioned impression and known source impression originated from a common source, and renders a conclusion of either "identification" – the questioned impression and known source impression both originated from the same source, "exclusion" – the questioned impression and known source impression originated from different sources, or "inconclusive" – there was insufficient agreement or disagreement to conclude whether the questioned impression and known source impression did or did not originate from the same source [10]. Again, like in the analysis and comparison phases, this determination of whether two impressions were made by the same source is not based on an empirical standard or specific measure; rather it is a personal determination made by the analysts on a case-by-case basis and depends on whether the analyst judges that the likelihood the two impressions were made by a different source is so remote that it is a "practical impossibility" [10] and he or she is willing to defend this claim to his or her peers or during litigation.

1.1.4 Verification

Once the analyst has rendered a conclusion, she will provide the questioned impression and known source impression to another "competent" analyst to repeat the ACE steps outlined above. If the both analysts reach the same conclusion then the findings are reported accordingly [10]. However, if the analysts reach different conclusions, the laboratory will make an administrative decision on how the conclusion is reported. Some laboratories may defer to a panel of analysts to independently conduct their own examinations and the final conclusion is based on some degree of consensus among the analysts or others may defer to a single manager or senior analyst to conduct their own examination and make the final conclusion. In situations where

consensus (however defined) is not achieved, some laboratories will simply report "inconclusive" whereas other laboratories may reassign the case to an analyst that may be "able" (or "willing") to render a more definitive conclusion.

1.2 Discussion of the Problem

From the general description of the ACE-V methodology, there are several key aspects of the methodology which expose the discipline to vulnerabilities. Broadly stated, there is wide latitude in how the method is applied, there are no standard instruments (other than human visual-cognitive system) to measure feature attributes, there is no traceability to empirical data to substantiate certain conclusions, and the human analyst plays the role of both the instrument of measurement (albeit actual measurements are not taken) and the instrument of interpretation by establishing (flexible, undefined) criteria by which (non)measured results are compared to determine their significance. This creates several concerns from a scientific standpoint and presents a grave threat to the scientific foundations of the discipline. Within this context, the overarching concern is the lack of an empirical means of calibrating the human system or an empirical foundation by which the significance of the evidence is evaluated such that a specified level of confidence can be provided for the final result [3, 7, 8]. This is likely to always be an issue where human beings, relying on their personal experience and judgment, play a role of the measuring instrument. The issue is further complicated when the human being also serves as the one who determines whether the measurements assessed are significant without empirical validation. Notwithstanding the clear potential for cognitive and contextual biases to impact the analyst's determinations and conclusion [11], there is a potent decision-theoretic influence across the entire examination spectrum that cannot be disentangled from the analyst's evaluative judgment of the evidence [12]. This adds another layer of complication in which personal, professional, and societal values and cultural expectations shape decision theoretic utilities which contribute to the analyst's evaluative judgment. Without defined measurements or empirical basis for which the significance of the evidence is evaluated, it is unclear *what* contributed to the overall assessment of the evidential strength and *how* it was evaluated. While the lack of empirical standards and measurements do not necessarily suggest the practice as a whole is unreliable or fraught with error, it does raise questions as to how reliable the assessment is for a specific case at hand.

For years, these issues have been echoed and communicated to the forensic science community by several different scientific advisory committees in the United States. The primary concerns related to these issues brought forth by these committees can be summarized as:

1. Forensic fingermark examinations rely on the subjective interpretation of forensic examiners, which are vulnerable to human error, inconsistency across examiners, and cognitive biases [7, 8].

2. The ACE-V methodology is not specific enough to qualify as a validated method; does not guard against bias; is too broad to ensure repeatability and transparency; and does not guarantee that two analysts following it will obtain the same results [3].

3. The reported results that make claims, directly or by implication, of zero error rates, 100% certainty, or a single source attribution to the exclusion of all other sources are not scientifically defendable [9]; conclusions of "identification" or "individualization" claims too much, is not adequately established by fundamental research, and is impossible to validate solely on the basis of experience [3, 8, 9]; and statements claiming or implying greater certainty than demonstrated by empirical evidence are scientifically invalid [7].

In his 2012 thesis, Langenburg [13] describes his attempts to gain a better understanding of what factors contribute to the overall assessment of the evidence by deconstructing the ACE-V framework into smaller, more compartmentalized tasks. In doing so, he provides a series of critical findings and recommendations for best practices and useful tools to help reduce expert variance and errors while increasing transparency and understanding of expert decision making. In the Analysis phase experiments, Langenburg finds significant variation in feature selection and determinations of whether marks were considered "of value" or "suitable for comparison." After introducing an annotation scheme, called "GYRO," to convey levels of uncertainty in the existence of minutiae annotated, Langenburg finds that, in general, experts' error rates for incorrect minutiae selection mirrored their assignment of uncertainty using the GYRO system (minutiae selection error rates increased as the analyst uncertainty assigned to the minutiae also increased). Further, when evaluating the number of minutiae reported versus the decision for "suitability" for a given mark, an operational threshold for "suitability" appeared between 7 to 8 minutiae. In the Comparison phase experiments, Langenburg finds that the clarity of marks had the most drastic effect on the ability for experts to accurately locate a match during searching tasks—lower clarity images resulted in fewer correct responses. In the Evaluation phase experiments, Langenburg finds that experts were mostly consistent in their precision and sensitivity of reported decisions with a false positive error rate of approximately 0.1%. When using the LQMetrics Quality Map software [14] to assess the quality of fingermarks, analysts reported highly reproducible "identification" and "exclusion" decisions when quality values were "high." When quality values were "low," analysts reported highly reproducible "no value" decisions. For quality values in between, significant variations were observed in analysts' determinations of "value" or "identification" and "exclusion" decisions. Following these series of experiments, Langenburg identifies several significant findings from the research. Key findings among those discussed by Langenburg [13] include:

1. There is significant variation in the features that analysts perceive, select, and utilize throughout the ACE-V process.

2. GYRO conveys analysts' uncertainty regarding the features that are selected.

3. Tools, strategies, and specific training can be implemented to reduce the variation of feature selection.

4. Operational decision thresholds were shown for decisions "of value" and "identification."

5. Methods for assessing quality were useful predictors of case complexity.

6. Reproducibility and repeatability for feature selection and reported decisions was significantly lower in complex cases where marks had marginal ridge detail.

7. Consensus feature sets are the most reliable and accurate features upon which to base a decision.

Taking into consideration the findings listed above, Langenburg [13] proposes the following recommendations to help improve the practice of fingerprint examination and provide a starting point for future research:

1. As early in the examination process as possible, cases should be identified as to their level of complexity.

2. Documentation must be done in each case to the extent that is appropriate for the complexity of the case and to the extent that it is sufficiently transparent how the analyst arrived at her conclusions.

3. Decision thresholds should be formalized, transparent, and documented.

4. In disputed and/or complex cases, a consensus feature set approach should be considered for the primary basis of the reported conclusions.

5. Likelihood ratios offer a demonstrative means of representing the weight of the contribution of the corresponding features between two fingerprint impressions.

The work by Langenburg provides one of the first and most comprehensive evaluations of the ACE-V process in terms of salient factors and general trends influencing the assessment of fingermark evidence.

In 2017, Hicklin, in his thesis [15], expands upon the work by Langenburg through a compilation of studies conducted over the course of several years in which he attempts to address ways to improve the rigor, standardization, transparency, and quantifiability of the fingermark examination process. While Hicklin notes there have been improvements to the processes over the years, he found that there is still a great deal of variation and ambiguity in the process and provides several findings and recommendations that augment those proposed by Langenburg. Key findings and recommendations among those discussed by Hicklin [15] are summarized below:

1. Address inconsistencies through standardized training, competency and proficiency tests, operating procedures, certification, and accreditation: There remains a great deal of variation in friction ridge examination procedures and terminology—among agencies, among training programs, and among examiners. These differences present major problems to the criminal justice system causing results to vary by organization and by examiner, inserting ambiguity into legal testimony, and impeding cross-agency evaluations of examiners' performance.

2. Conduct further Black Box testing based on examiner proficiency and comparison difficulty: Black Box tests to date are based on performance for examiners in general, on marks and exemplars of a range of qualities; however, there is a wide disparity in the difficulty of friction ridge 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.

3. Focus attention on effectiveness as well as error: Training and competency/proficiency tests should reflect an increased focus on effectiveness and efficiency, not just the avoidance on eliminating or minimizing error. Trying to optimize a single error would be a red flag; errors almost always involve tradeoffs. We can (facetiously) eliminate all erroneous identifications by doing no work whatsoever, which is obviously not an acceptable solution. 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.

4. Require detailed documentation of the features used by examiners in making their determinations: Rigorously defined and consistently applied methods of performing and documenting ACE-V would improve the transparency of the friction ridge examination process and reduce the risk of error.

5. Provide a 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. The value of marks is a continuum that is not well described by binary (value vs. no value, or individualization vs. inconclusive) determinations. In the medium or long term, probabilistic determinations will provide such continuous measures.

6. Use quality metrics: The lack of standard methods of assessing quality means that all friction ridge evidence must be treated as if it is all the same. The ability to assess the quality of a mark, or the comparative quality of a mark-exemplar comparison, suggests a variety of possible uses, such as for assessing the examiner performance metrics, quality-directed workflow, enhanced quality assurance practices, and describing datasets for research purposes.

7. Augment examiners' determinations with probabilistic models: A great deal of on-going research is being conducted on statistical models designed to quantify the probability that a mark came from a specified source. Having this capability will be particularly useful for more difficult comparisons that cannot be identified using fully automated means in a "lights-out" environment.

Then, in her thesis in 2020, Eldridge [16] expands on the work by Langenburg [13] and Hicklin [15] while focusing specifically on the concept of "suitability" during the Analysis phase of the ACE-V methodology. Eldridge not only explores the information that is most considered by examiners when making decisions through white-box testing, but also proposes expanded scales

for assessing the utility of a mark and presents the development and validation of a predictive suitability model that relies on both key observations from a human expert and automated measures from existing quality tools [16]. Ultimately, Eldridge demonstrates, as a proof-of-concept, benefits that can be achieved by a hybrid examiner—algorithm model that leverages the strengths of both to provide consensus-based guidance and encourages the friction ridge community to move toward adopting such approaches. Key findings and recommendations from Eldridge [16] include:

1. Do not annotate with minutiae-type specific markers: Results showed that examiners are not consistent or cohesive in their use of marker types. The results support that there is really no justification for designating a particular minutia as either a ridge ending or a bifurcation. In most cases, the examiner cannot distinguish one from the other with confidence and therefore the designation is arbitrary.

2. Develop consensus-based standards for suitability decisions: Results showed a high degree of variability in suitability decisions thus making it clear that the decision is currently far too subjective and that standards are needed to guide examiners in these decisions.

3. Document analysis, including confidence level: Documentation provides the transparency necessary to support the suitability decision and for others to review the factors that went into that decision, such as which features were considered, the weight provided by those features, and how the decision compares to standardized criteria and thresholds.

4. Use the model as a second—or first—opinion: Results demonstrate that a suitability model performs well for predicting the consensus opinion of expert examiners along all four scales of suitability proposed. Adopting such a model could not only improve consistency, but also increase efficiency of suitability determinations during friction ridge examinations.

5. Use the new categories proposed by this research: The proposed new and expanded scales for suitability encourage examiners to approach suitability in terms of a continuum and to support tailored quality controls as part of a broader quality assurance program (e.g., increased quality controls for more complex impressions and lower quality controls for non-complex impressions).

Collectively, Langenburg [13], Hicklin [15], and Eldridge [16] have provided a comprehensive evaluation of the practice of friction ridge examination and a deeper understanding of generalized performance characteristics and sources of variability in the processes; however, the generalized nature of their findings has limited applicability to a specific case. The findings and recommendations from Langenburg, Hicklin, and Eldridge, as well as those put forth by the PCAST and AAAS, boil down to a single common issue—there remains a critical need for the friction ridge community to move towards integrating quantitative and statistical tools into the friction ridge examination methodology in order to provide an additional empirical foundation to the assessment of the evidence. Doing so will not only allow several of the recommendations put forth by Langenburg, Hicklin, and Eldridge to be acted upon in terms of standardizing the procedures and improving the overall practice, but it will also allow the friction ridge analyst the

ability to clearly demonstrate the significance of an examination and communicate the reliability of the assessment for the specific case at hand thereby resolving the underlying issue espoused by the NRC, PCAST and AAAS which have called into question the continued admissibility of fingermark evidence.

Within the context of the ACE-V methodology, the development and integration of quantitative and statistical tools are most critical for:

(1) Analysis: Assessment of quality/clarity of friction ridge features and the value of the mark for subsequent comparison.

(2) Comparison and Evaluation: Assessment of the statistical strength between two impressions.

Over the years, there have been a number of notable efforts by researchers in which quantitative and statistical tools were introduced for these purposes [16-44]; however, none have successfully made it into the hands of practitioners and implemented into routine casework operations. There are a number of different reasons for this, which include both technological and cultural dimensions. For example, Eldridge [16] notes that a significant limitation to her thesis was the technological challenges of developing a single stand-alone version of the proposed model in a single user interface that is conducive to operational use. This is unfortunately a common issue that renders such tools inaccessible for practical applications. Aside from technological issues, the single greatest challenge that is often underestimated with the practical application of such tools is the longstanding cultural hesitation and the paradigm shift that would be required to facilitate such a transition. Attention must be directed toward how to most effectively navigate the implementation of these tools in a field that has largely been dominated for so long by human interpretation and experience-based judgment. In forensic science, little effort has been given to such a critical issue. As a result, many prior recommendations have yet to manifest in practice thereby stifling their impact.

Overcoming the challenges associated with practical implementation of quantitative and statistical tools is not a straightforward task. The integration of these tools to the friction ridge discipline is often viewed as more than the mere introduction of an "additional tool in the toolbox" to assist analysts in their interpretation. Rather, it has been viewed as a challenge to a century old paradigm and an indictment on traditional practices and examiners' prior judgments, experiences, and expertise. Consequently, successful implementation requires consideration of the cultural challenges that come with facilitating practitioner acceptance of the new technology and methods. The hesitation by the friction ridge community is best illustrated following a commentary in 2001 by Champod and Evett in which they proposed probabilistic reasoning based on empirical measurements as a more appropriate, scientifically compatible, and defensible approach to fingerprint examination and which could afford the legal system the ability to consider potentially valuable evidence that would otherwise be denied under traditional practices if it did not meet the analyst's experience-based threshold of "sufficiency" [45]. Shortly after this commentary emerged, it was quickly met with resistance and incited hostility by many throughout the friction ridge community [46-48]. In direct response, one author responded with the opening sentence, "Once again, identification science is under attack, this time from a shotgun blast by statisticians"

[46]. Following this response, others followed suit. Another author responded by offering "reassurance to members of the fingerprint community who may feel there are moves afoot to weaken the hold that fingerprint has as a science" and further reported, "Recently the courts have accepted the status earned by these practitioners and pronounced that there is nothing lacking in the process of fingerprint identification" [47]. And yet another author reported being "disturbed and astounded that someone in our profession would propose such a detrimental and dangerous proposition" and "urge the IAI to reject this probabilistic approach to fingerprint evidence" [48]. Since these initial reactions nearly two decades ago, the friction ridge community has come a long way in beginning to embrace this new paradigm as a theoretical ideal; however, the shift is still within its infancy and polarizing viewpoints continue to exist.

Recognizing that the ultimate issue is the need for the friction ridge community to move towards integrating quantitative and statistical tools into the friction ridge examination methodology in order to provide an empirical foundation to the assessment of the observations, the tools must be developed in a manner that maximizes practitioner receptivity and acceptance. This requires consideration of the needs and expectations of the adversarial legal environment, laboratory operational workflows and throughput requirements, practitioner knowledge and skills, and appropriately balancing human intuition and judgment with quantitative and empirical standards as it relates to the procedures governing the use of the tools and reporting and testimony of the results.

1.3 Objectives of the research

The major objectives of this research are twofold: (1) to develop, validate, and make publicly accessible algorithms and software applications for friction ridge examination capable of (a) assessing the clarity of friction ridge skin features and overall quality of impressions and (b) evaluating the statistical strength of correspondence between two impressions, and (2) to develop strategies for practical application and implementation of these (and similar) tools in an operational forensic science laboratory. These objectives are achieved through a series of studies and discussions related to the development, validation, and operationalization of these tools in practice.

Collectively, this work expands upon the generalized foundations established by Langenburg, Hicklin, and Eldridge, among others, and attempts to provide an implementable solution for the friction ridge community to act on their recommendations and resolve the underlying issue espoused by prior commentators, including the PCAST and AAAS, which have called into question the continued admissibility of friction ridge impressions in court.

1.4 Structure of the thesis

This thesis combines two parts. Part I (Chapters 2 through 4) focuses on the development and validation of algorithmic tools. Part II (Chapters 5 through 8) focuses on the implementation of algorithmic tools into practice. Chapter 9 summarizes the contributions and major findings in each preceding chapter and provides overarching recommendations for future practice and

research. Chapter 10 provides a high-level summary and conclusion of the work. The scope of each chapter is briefly described below:[4]

Chapter 2 – Quality Assessment Software (DFIQI)

This chapter presents a manuscript entitled "A Method for Measuring the Quality of Friction Skin Impression Evidence: Method Development and Validation" (Swofford et al., 2021) [49] published in *Forensic Science International* that describes the development and validation of a publicly accessible algorithm and software application (referred to as the Defense Fingerprint Image Quality Index, or DFIQI). The DFIQI algorithm first assesses the clarity of each friction ridge feature identified by an analyst and provides a color-coded output (green, yellow, red) to the user as an indication of its reliability. The software then accounts for the quantity and clarity of features to provide a measure of the overall quality of the impression for suitability for further examination by an expert. In addition to the published manuscript, this chapter also discusses the performance of DFIQI compared to other available methods.

Chapter 3 – Statistical Interpretation Software (FRStat)

This chapter presents a manuscript entitled "A Method for the Statistical Interpretation of Friction Ridge Skin Impression Evidence: Method Development and Validation" (Swofford et al., 2018) [50] published in *Forensic Science International* that describes the development and validation of a publicly accessible algorithm and software application (referred to as the Friction Ridge Statistical Interpretation Software, or FRStat). The FRStat algorithm first calculates the similarity (referred to as the *Global Similarity Statistic*, or GSS) between two sets of features identified by an analyst on two separate impressions which the analyst believes to correspond. The software then provides two estimates, one indicating how often impressions originating from common sources would result in a GSS that is equal to or less than the calculated GSS and another indicating how often impressions from different sources would result in a GSS that is equal to or greater than the calculated GSS. The two values are then combined as a ratio providing a single summary statistic indicating to what extent the GSS is consistent with impressions originating from a common source compared to different sources. In addition to the published manuscript, this chapter also discusses the performance of FRStat compared to another available methods.

Chapter 4 – Toward Objectivity: Integrating Algorithmic Outputs

This chapter explores the utility (i.e., usefulness), from a quality management standpoint, of integrating the DFIQI and FRStat algorithms into a single system for which the input to the FRStat is dependent upon the output from the DFIQI. An integrated system such as this could provide a more objective and semi-automated approach for ensuring analysts' interpretations are empirically

[4] Throughout this thesis, the term "latent print," "fingermark," "mark," and "fingerprint" are used interchangeably to refer to chance reproductions of friction ridge skin. Further, use of these terms also includes "patent print" and "plastic print" for purposes of the thesis. The technically appropriate terms are "mark" or "fingermark" to refer to chance reproductions of the friction ridge skin. The inconsistency in using one specific term is due to their use in published manuscripts and other published sources that are reflected in this thesis—the terms "latent print" and "fingerprint" are commonly used throughout the United States whereas the terms "mark" and "fingermark" are commonly used throughout European countries. Additionally, references to tables, figures, and equations have been modified from their original publication to uniquely identify them and reflect the chapters in which they occur within this thesis.

supported for all major decisions throughout the examination methodology as well as a means for monitoring and ensuring the quality of results meet minimum standards for quality assurance. This chapter describes how the two systems can be integrated and evaluates the impacts of such an application in practice.

Chapter 5 – Evaluation of Practitioners' Perspectives

This chapter presents a manuscript entitled "'Mt. Everest—We are Going to Lose Many': A Survey of Fingerprint Examiners' Attitudes Toward Probabilistic Reporting" (Swofford et al., 2021) [51] published in *Law, Probability & Risk* that explores practitioners' perspectives related to probabilistic reporting practices (with or without algorithmic tools) in terms of their reactions, attitudes, and sources of resistance toward probabilistic methods. Practitioners' perspectives are evaluated quantitatively and qualitatively using a structured survey instrument with Likert-scale response and free-text responses choices.

Chapter 6 – Evaluation of Stakeholders' Perspectives

This chapter presents a manuscript entitled "Probabilistic Reporting and Algorithms in Forensic Science: Stakeholder Perspectives within the American Criminal Justice System" (Swofford & Champod, 2022) [52] published in *Forensic Science International: Synergy* that explores perspectives from key criminal justice stakeholders (forensic laboratory managers, prosecuting attorneys, defense attorneys, judges, and other academic scientists and scholars) related to interpretation and reporting practices (with or without algorithmic tools) and the use of computational algorithms in legal settings. Stakeholders' perspectives are evaluated qualitatively from semi-structured interviews.

Chapter 7 – Implementation of Algorithms: A Responsible and Practical Roadmap

This chapter presents a manuscript entitled "Implementation of Algorithms in Pattern & Impression Evidence: A Responsible and Practical Roadmap" (Swofford & Champod, 2021) [53] published in *Forensic Science International: Synergy* that discusses challenges, considerations, and a path forward for the implementation of algorithms in pattern and impression evidence domains. The paper explores human-algorithm interactions and seeks to understand *why* practitioners (in general) tend to oppose algorithmic interventions and *how* their concerns might be overcome. Further, it addresses issues concerning to human-algorithm interactions in both real-world domains and laboratory studies as well as issues concerning the litigation of algorithms in the American legal system. With these considerations in mind, the article proposes a strategy for approaching the implementation of algorithms, including a taxonomy describing the various ways algorithms can be implemented, in a responsible and practical manner.

Chapter 8 – Operationalization of Algorithms: Anecdotal Reflections and Observations

This chapter discusses the implementation of algorithms into operational practice through anecdotal reflections and observations from my own experiences—both as a laboratory manager and as a private analyst. From those distinct experiences, I reflect on my perspective and discuss

strategies for implementation, considerations for policies and procedures, and use of the algorithm in litigation.

Chapter 9 – Looking Forward: Implications, Recommendations, and Future Research

This chapter summarizes key developments and findings from this thesis and discusses the impact and implications of the work, provides recommendations for friction ridge examination practices, and proposes areas for future research.

Chapter 10 – Conclusion

This chapter provides a high-level summary of the thesis, including a brief description of the algorithms developed, various challenges and considerations related to the implementation of algorithms into practice, and a path forward toward stronger foundations for friction ridge examination and other pattern and impression evidence disciplines.

2 Quality Assessment Software (DFIQI)

This chapter presents a manuscript entitled "A Method for Measuring the Quality of Friction Skin Impression Evidence: Method Development and Validation" (Swofford et al., 2021) [49] published in *Forensic Science International* that describes the development and validation of a publicly accessible algorithm and software application (referred to as the Defense Fingerprint Image Quality Index, or DFIQI). The DFIQI algorithm first assesses the clarity of each friction ridge feature identified by an analyst and provides a color-coded output (green, yellow, red) to the user as an indication of its reliability. The software then accounts for the quantity and clarity of features to provide a measure of the overall quality of the impression for suitability for further examination by an expert. In addition to the published manuscript, this chapter also discusses the performance of DFIQI compared to other available methods.

2.1 Method Development and Validation

A Method for Measuring the Quality of Friction Skin Impression Evidence: Method Development and Validation

[1]Swofford, H.; [1]Champod, C.; [2]Koertner, A.; [1,3]Eldridge H.; [4]Salyards M.

[1]School of Criminal Justice, Forensic Science Institute, University of Lausanne, Switzerland
[2]U.S. Army Criminal Investigation Laboratory, Defense Forensic Science Center, USA
[3]RTI International, Inc., USA
[4]Compass Scientific, LLC, USA

2.1.1 Abstract

The forensic fingerprint community has faced increasing criticism by scientific and legal commentators, challenging the validity and reliability of fingerprint evidence due to the lack of an empirical basis to assess the quality of the friction ridge impressions. This paper presents a method, developed as a stand-alone software application, DFIQI ("Defense Fingerprint Image Quality Index"), which measures the clarity of friction ridge features (locally) and evaluates the quality of impressions (globally) across three different scales: value, complexity, and difficulty. Performance was evaluated using a variety of datasets, including datasets designed to simulate casework and a dataset derived directly from casework under operational conditions. The results show performance characteristics that are consistent with experts' subjective determinations. This method provides fingerprint experts: (1) a more rigorous approach by providing an empirical foundation to support their subjective determinations from the Analysis phase of the examination methodology, (2) a framework for organizations to establish transparent, measurable, and demonstrable criteria for Value determinations, (3) and a means of flagging impressions that are vulnerable to erroneous outcomes or inconsistency between experts (e.g., higher complexity and difficulty), and (4) a method for quantitatively summarizing the overall quality of impressions for ensuring representative distributions for samples used in research designs, proficiency testing and error rate testing, and other applications by forensic science stakeholders.

Keywords: Forensic Science; Fingerprints; Quality Metric; Probability

2.1.2 Introduction

Friction ridge examination is practiced by nearly every forensic laboratory throughout the world and is often relied upon as evidence that an individual touched an item or was present at the scene of a crime. The process for conducting friction ridge examination is described by the acronym ACE-V, which stands for "Analysis," "Comparison," "Evaluation," and "Verification." ACE-V has been described in the forensic literature as a means of comparative analysis of evidence since 1959 [3]. The process begins with the *analysis* of the latent print in which human analysts will visually observe and interpret friction ridge detail in a latent impression and determine if it is "suitable" or "of value" for comparison purposes. This determination is an experience-based judgment based on the quality and quantity of friction ridge detail discernible in the impression. If a latent print does not have "sufficient" detail to form a conclusion regarding the source of the impression, the impression is determined to be "not suitable" or "no value" and no comparison is made. If an impression is determined to be "of value," the analyst will perform a side-by-side *comparison* of the friction ridge detail between the latent print and the known prints from an individual. During comparison, and ultimately thereafter, the analyst will *evaluate* the similarities and differences of the friction ridge detail between the two impressions and form a conclusion regarding the source of the impression. *Verification* occurs when another qualified analyst repeats the observations and forms the same conclusion.

Within the ACE-V process, the "analysis" of the friction ridge skin detail is one of the most critical tasks of the examination as it establishes whether, and to what extent, the impression bears sufficiently discernible features that can be used for examination. More specifically, during the "analysis," the analyst is particularly concerned with identifying reproducible and discriminating attributes of the friction ridge detail which may be used for comparison and evaluation against a known source impression. The ability for the analyst to reliably detect these attributes depends heavily on the clarity of the impression. Generally, as the clarity of an impression increases, analysts' have more confidence in their interpretation of the location, orientation, type, and spatial arrangement of features. Additionally, as the number of interpretable features increases, the discriminating strength of the impression as a whole is considered to increase as well. Once the features have been detected, the analyst will assess the overall quality of the impression and make a determination of the "suitability" or "value" for further comparison and evaluation [10]. This determination is not based on an empirical standard; rather, it is a subjective determination made by the analyst on a case-by-case basis and depends on whether the analyst believes the quality of the impression is sufficiently reproducible and selective to be compared to a known source and render a particular conclusion regarding the potential source of the impression. Consequently, assessments made during friction ridge examinations are susceptible to variation from one analyst to another (inter-analyst) as well as by the same analyst from one examination to another (intra-analyst). When considering borderline impressions which contain marginal quality or quantity of features, these variations often result in differences in the *analysis* conclusion. In the broad spectrum, however, while the lack of empirical standards and measurements do not necessarily imply the practice as a whole is unreliable or fraught with error, it does raise questions as to how reliable the evidence is for the case at hand. Thus, there is a critical need for the friction ridge community to move towards integrating tools to quantitatively assess the clarity and quality of

friction ridge impression details to standardize and provide an empirical warrant for analysts' claims [3, 7-9].

Over the years, there have been several notable efforts by researchers in which quantitative tools were introduced for assessing the quality of friction ridge impressions [14, 17-28, 44]. The majority of these efforts can be classified as suitability prevision models, which provide a predictive estimate of whether the impression is suitable for some intended purpose or utility, such as suitability for identification or exclusion purposes during manual comparisons or, more often, for assessments of search performance using automated fingerprint identification systems (AFIS). Early models are described by Alonso-Fernandez et al. (2005) and all focus on calculating quality as a means of predicting AFIS feature extraction or matcher performance. Most of the early methods entailed a variety of different image processing techniques, such as measuring ridge frequency, ridge thickness, and ridge to valley thickness ratio, using Gabor filters to increase contrast, measuring pixel intensity differences, two-dimensional Discrete Fourier Transform (DFT), and neural network classifiers to classify local regions as "good" or "bad" quality [17]. Alonso-Fernandez et al. note that all of the various methods tend to behave similarly to one another except for the method based on neural network classifiers, likely due to the low number of quality labels used for training, and propose the concept of integrating the various algorithms into a quality-based multimodal authentication system for future works.

In 2007, Nill developed Image Quality of Fingerprint (IQF) as a freeware software application designed to predict AFIS matching performance, alert operators to poor quality enrollment of known source standards or aid in performance assessments of capture devices [18]. The approach developed by Nill relies on the two-dimensional, spatial frequency power spectrum of the digital fingerprint image to produce a global assessment of quality [18]. In 2008, Fronthaler et al. studied the orientation tensor of fingerprint images to quantify signal impairments like noise, lack of structure, and blur with the help of symmetry descriptors when combining multiple AFIS matchers for improved matching performance [19].

In 2011, Hicklin et al. [20] attempted to understand how human latent fingerprint analysts assess fingerprint quality by surveying eighty-six latent print examiners from federal, state, local, international, and private sector laboratories using overlapping subsets of 1,090 latent and exemplar fingerprint images to identify key features that will guide the development of automated quality metric algorithms in future works [20]. Up to this point, nearly every other method was focused entirely on optimizing AFIS matching performance or developing quality metrics to predict match performance rather than attempting to understand what was considered by human analysts during manual examinations. From the survey, Hicklin et al. note there is general concurrence of human assessments of local and overall image quality, but enough variation between examiners to result in differing conclusions and demonstrate the need to provide uniform definitions of quality and automated assessment tools to standardize the practice [20].

In 2012, two additional methods were proposed: both focused on optimizing or predicting AFIS match performance. While earlier methods tended to focus on biometric enrollments and known source impressions, these were geared more towards latent fingerprint impressions. Murch et al. (2012) proposed a method for automated feature extraction to improve the performance of AFIS searches of latent fingerprint impressions using image segmentation to differentiate the

foreground impression from background noise [21]. Yoon et al. (work first presented in 2012, but published in 2015) proposed a method for assessing latent fingerprint image quality using the product of the average ridge clarity bounded within the convex hull enclosing all annotated minutiae and total number of minutiae [22]. The calculation of average ridge clarity involved the application of two-dimensional Fourier analysis to a pre-processed contrast enhanced image. Although Yoon et al. was focused specifically on latent impressions, the quality algorithm was still geared towards predicting AFIS matcher performance and thus not necessarily tailored to attributes considered during human examinations [22, 44].

In 2013, three additional approaches were introduced, which begin to steer focus towards latent fingerprint image clarity relevant during human examinations compared to prior methods. Hicklin et al. (2013) developed Latent Quality Assessment Software (LQAS), which applies a variety of image processing algorithms to assess the clarity of friction ridges in localized regions [14, 23] (LQAS [23] was later enhanced and combined with Universal Latent Workstation (ULW). Within ULW, it is referred to as LQMetric. Details related to LQMetric development are provided by Kalka et al. 2020 [14]). Based on the clarity assessment, the software then applies a color-coded clarity map which corresponds to the color codes within the American National Standards Institute/National Institute of Standards and Technology (ANSI/NIST) 2011 standard "Data Format for the Interchange of Fingerprint, Facial & Other Biometric Information" [24] for simple interpretation and a standardized framework for documentation [13-14]. Sankaran et al. (2013) propose a method which assesses ridge clarity and quality [25]. The former (Hicklin et al.) refers to the visual discernibility of the features irrespective of the presence or absence of features and the latter (Sankaran et al.) refers to the quantity and number of features present in a given local region (i.e. a predictor of AFIS matching performance). The local ridge clarity assessment is based on average eigenvalues from decomposed structure tensors following image smoothing using a Gaussian filter [25]. A local clarity map is generated as a result of the clarity assessment similar to that of Hicklin et al. (2013) [20]. The ridge quality assessment is calculated as the kurtosis of the weighted average histogram based on the local clarity map described previously along with the number of features present within a local region [25]. Pulsifer et al. (2013) propose a method for calculating overall quality based on a semi-automated assessment of the local clarity maps generated from LQAS developed by Hicklin et al. (2013) [14, 23] to produce an alternative way of calculating the overall quality of the impression [26].

In 2014, Kellman et al. proposed a number of quantitative measures of image characteristics related to image quality metrics, such as intensity and contrast information, as well as measures of information quantity, such as total fingerprint area, to calculate image quality and predict analyst performance and perceived difficulty during comparisons by human analysts [27]. The work by Kellman et al. indicates a shift towards establishing quality metrics geared towards predicting human analyst performance rather than tailored specifically to predicting AFIS match performance. More recently in 2018, with a similar intent as Hicklin et al. [14, 23] and Kellman et al. [27], Chugh et al. proposed a crowdsourcing framework to understand the underlying bases of suitability determinations by fingerprint analysts and use it to develop an automated means of predicting suitability determinations [28].

While there have been a number of different models proposed over the years, the majority of them are geared entirely towards optimizing or predicting AFIS match performance rather than

focused on assessing local ridge clarity (discernibility of feature data) and predicting human performance using image quality attributes considered by human analysts during manual comparisons. Consequently, these types of predictive models are often based on the aggregate of qualitative and quantitative attributes of the entire impression to provide a single estimate of utility or quality. These approaches often lack transparency and often do not necessarily correspond to the same features considered by human analysts during traditional examinations. The motivation behind this focus is largely driven by industry desires to optimize the performance of AFIS in a "lights-out" environment. Indeed, this focus is important for the broader biometric industry; however, the narrow focus on AFIS platforms leaves a gap as it relates to manual examination and interpretation processes by human analysts in the traditional forensic setting. Thus, the need remains for the development and implementation of tools capable of quantitatively assessing the clarity of friction ridge detail in a transparent and objective manner within in a simple, accessible, and user-friendly software application that can be easily integrated into friction ridge examination practices. Such a tool would offer significant improvements to traditional practices and permit laboratories to establish standardized suitability criterion and provide empirical substantiation to analysts' opinions.

This paper presents a method, developed as a stand-alone software application, DFIQI ("Defense Fingerprint Image Quality Index"), designed to measure the clarity of friction ridge impression minutiae and provides a quantitative assessment of the quality of an impression for comparison and evaluation purposes. Although this method builds upon general approaches described earlier and considers well established means of assessing image clarity, it provides a simple and novel approach for quantifying the quality of friction ridge impressions. Further, having been developed as an automated stand-alone software application, this method is accessible to the forensic community[5] thereby providing the capability for laboratories to ensure the quality of friction ridge details are sufficient to permit reliable interpretations and move toward standardizing and improving traditional practices. In the sections that follow, this paper provides a brief overview of the calculations performed by the method followed by more detailed discussions regarding its development, performance and validation. Limitations of the method and considerations for policy and procedure when applied to forensic casework are discussed as well as implications for future integrations with other tools to strengthen the foundations of friction ridge examination in general.

2.1.3 Materials & Methods

Background

In general terms, the method assesses the clarity of friction ridges in localized "regions of interest" (ROIs) immediately surrounding the x,y location of features identified in the impression. Features can be identified by manual annotation or using automated feature extraction applications (followed by human-expert verification). Each region of interest is assessed using five variables (described below) consisting of various measures of friction ridge image clarity and quality. The five variables were selected by the authors based on domain expertise, reduced mathematical complexity, and algorithmic transparency. The output of each variable measured is normalized by

[5] The software application can be accessed at: https://doi.org/10.5281/zenodo.4426344.

a scoring function and combined to create a single quantitative value representing the clarity and quality of the friction ridges within the localized ROI. Each local ROI score is then combined to a single quantitative value representing the quality of the ROIs combined across the entire impression, which accounts for both the quality and quantity of detail in the impression.

Once the x,y coordinates are identified for the features in the impression (e.g. by an analyst marking the location), the application creates an inverted 8-bit digital grey-scale copy of the image on which all subsequent digital processing is performed. For each feature, a 2.54mm x 2.54mm (i.e. 0.1-inch x 0.1-inch) square ROI, centered on the location of the feature, is applied to the image and cropped (as a copy). The size of the ROI was selected to ensure it is small enough to represent a local region of the impression immediately surrounding a feature, but large enough to cover multiple ridges and enable a meaningful discrete Fourier transform related to the spatial frequency variable (described below); however, it was not subject to formal parameter optimization methods Each ROI is large enough to generally contain between four and seven ridges, depending on the width and orientation of the ridges. The five variable measures are taken from the cropped ROI to calculate the clarity and quality of the ridge detail immediately surrounding each individual feature in the impression.

Before the variable values are calculated, each ROI is split into two separate images to separate the "ridges" from the "furrows" (or more appropriately referred to as "signal" from "background") by applying adaptive mean thresholding to the pixel intensity values with a local neighborhood radius of 0.38mm. The 0.38mm radius was selected based on ad hoc testing and not subject to formal parameter optimization methods. Unlike simple thresholding methods, adaptive thresholding determines the threshold for a pixel based on a small region around it resulting in different thresholds for different regions of the same image. This generally provides greater segmentation accuracy as illumination conditions may vary throughout an image. Figure 2-1 illustrates the results of applying adaptive thresholding to a cropped ROI.

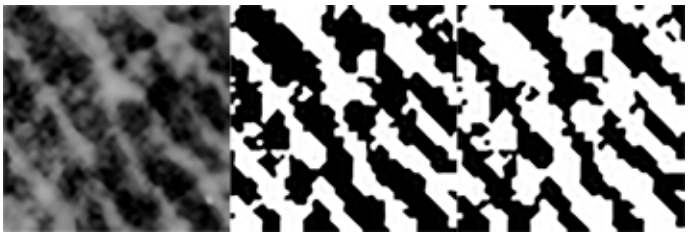


Figure 2-1: The image on the left represents the original ROI (the darker color pixels correspond to friction ridges). The image in the center represents the binary mask of the segmented ROI for which the black areas correspond to pixels thresholded as "signal." The image on the right represents the binary mask of the segmented ROI for which the black areas correspond to pixels thresholded as "background" (i.e. the image on the right is the inverse of the image in the center). NOTE: Actual size of images are 2.54mm x 2.54mm. Images are enlarged and pixels interpolated for illustration.

Variables

Using the cropped and segmented ROIs, the following measures of clarity and quality are calculated:

- Signal Percent Pixels Per Grid (S3PG): This variable calculates the percentage of pixels that have been segmented as "signal" compared to the total number of pixels available in the ROI. For a high-quality impression of friction ridges, an approximate value of 50, accounting for approximately 50% of total pixels segmented as "signal," is expected. As S3PG values deviate from the expected output of 50 in one direction or another, it suggests there are distorting artifacts in the ROI that may interfere with accurate detection of friction ridge detail.

- Bimodal Separation (BS): This variable calculates an index value summarizing the extent to which two histograms of pixel intensity values are separated from one another. Using the pixel intensity values of those segmented as "signal" and those segmented as "background," the index is calculated using the formula below. As the difference between the mean values increase and the standard deviations decrease between the segmented images, the value of the bimodal separation index increases, which indicates greater contrast between pixels classified as "signal" versus "background." On the other hand, as the difference between the mean values decrease and the standard deviations increase between the segmented images, the value of the bimodal separation index decreases, which indicates lower contrast and may interfere with accurate detection of friction ridge detail. The bimodal separation variable is calculated using the formula in equation 2-1.

$$x = \frac{\bar{S} - \bar{B}}{2(\sigma_S + \sigma_B)}$$

Equation 2-1: The formula for which the bimodal separation variable is calculated for each ROI.

- Acutance (ACUT): This variable calculates an index value summarizing the natural log of the mean acutance across the entire ROI and is applied to the non-segmented copy of the image. Acutance is described as the physical characteristics that underlay the subjective perception of "sharpness" in an image. In general terms, the acutance is calculated as the mean squared difference between a center pixel and its eight neighboring pixels in a 3x3 window iteratively calculated across an entire image. As the difference of pixel intensities increase, the perceived sharpness of the objects represented in the image also increase. This perceived increase of sharpness is represented by a higher acutance index value. As the acutance index value decreases, the perceived sharpness of the image decreases resulting in lower contrast which may interfere with accurate detection of friction ridge detail. The acutance variable calculation routine is illustrated in Figures 2-2a and 2-2b and stated in equation 2-2 (adapted from Choong et al. [54]).

I_1	I_2	I_3
I_8	I_c	I_4
I_7	I_6	I_5

Figure 2-2a: The 3x3 window representing a neighborhood of pixel values (the center pixel surrounding by its 8 contiguous neighbors).

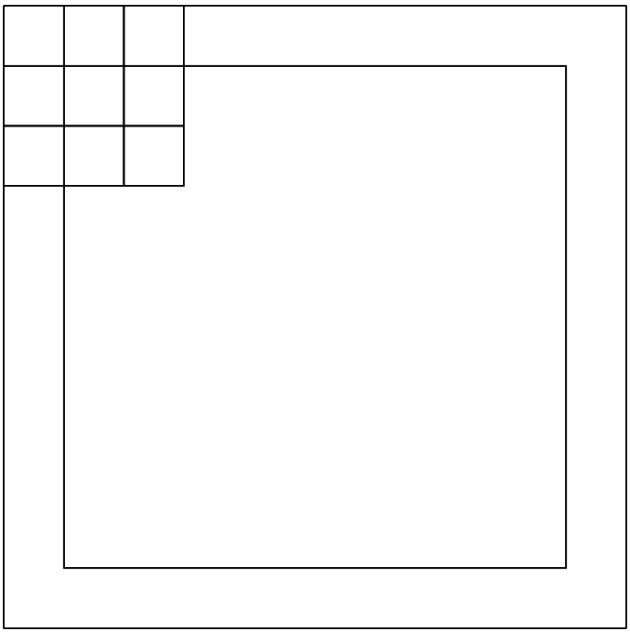


Figure 2-2b: The external box is a simplistic illustration representing the entire ROI containing p x p pixels (e.g. for an image resolution of 500 pixels per inch, p = 50 pixels. The inner box is a simplistic illustration representing the inner window of p-1 x p-1 pixels for the ROI in which every pixel serves as the center pixel of the scrolling 3x3 pixel window. The 3x3 window at the top left is a simplistic illustration of the 3x3 window represented in Figure 2-2a.

$$x = ln\left(\frac{\sum(\sum_{n=1}^{8}(I_c - I_n)^2)}{8(p-2)^2}\right)$$

Equation 2-2: The formula for which acutance is calculated for each ROI.

- Mean Object Width (MOW): This variable calculates the mean width of objects segmented as "signal" in the ROI. The term "objects" refers to a set of contiguously thresholded pixels within the "signal." The width of each object is calculated by fitting an ellipse and measuring the width of the minor axis. In the context of friction ridge impressions having perfect quality, those pixels thresholded as "signal" would correspond to separate and distinct "objects" in the image, representing separate friction ridges having nearly uniform and predictable widths. As the values for the mean object width deviate from the expected

width of friction ridges in one direction or another, it suggests there are distorting artifacts in the ROI that may interfere with accurate detection of friction ridge detail. The manner in which the mean object width variable is calculated is illustrated in Figures 2-3a and 2-3b.

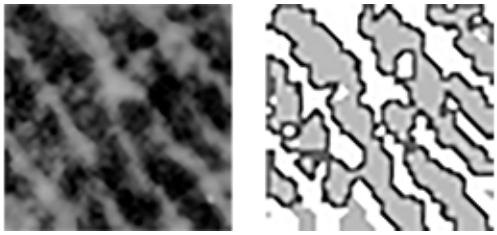


Figure 2-3a: The image on the left represents the original ROI (the darker color pixels correspond to friction ridges). The image on the right represents the mask of the segmented ROI for which the light grey areas correspond to pixels thresholded as "signal." The dark grey borders represent the borders around groups of contigous pixels represenging the various "objects" in the impression. NOTE: Actual size of images are 2.54mm x 2.54mm. Images are enlarged and pixels interpolated for illustration.

Figure 2-3b: An ellipse is fit to each distinct "object" in the image (ellipses overlaid on the origial image of friction ridges). The object width is calculated by measuing the width of the minor axis of each ellipse. In this example, two ridges appear connected together due to smudging in the impression resulting in a larger mean object width for the ROI; thus indicating the presence of distorting factors which may interfere with accurate interpretation of friction ridge detail. NOTE: Actual size of image is 2.54mm x 2.54mm. Images are enlarged and pixels interpolated for illustration.

- Spatial Frequency (SF): This variable calculates the spatial frequency of the ridges in the non-thresholded ROI using the two-dimensional discrete Fourier transform. For high-quality impressions of friction ridges, the ridges have been shown to have a predictable spatial frequency of approximately 2.1 ridges per millimeter for males and 2.4 ridges per millimeter for females [55] (combined mean of approximately 2.25 ridges per millimeter). As the spatial frequency values deviate from the expected output of approximately 2.25 ridges per millimeter in one direction or another, it suggests there are distorting artifacts in the ROI that may interfere with accurate detection of friction ridge detail. The two-dimensional discrete Fourier transform for a sample ROI is shown in Figures 2-4a and 2-4b to illustrate how the system calculates this variable.

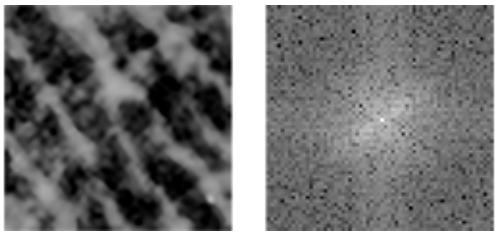


Figure 2-4a: The image on the left represents the original ROI (the darker color pixels correspond to friction ridges). The image on the right represents the discrete two dimensional Fourer transform of the image on the left. NOTE: Actual size of images are 2.54mm x 2.54mm. Images are enlarged and pixels interpolated for illustration.

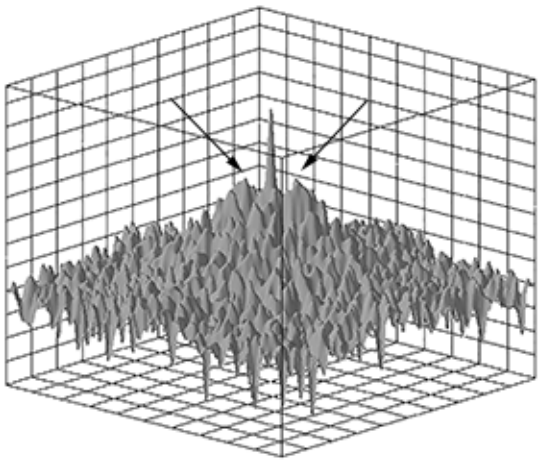


Figure 2-4b: A three-dimensional representation of the pixel intensity values of the discrete two-dimensional Fourier transform image in Figure 2-4a. The vertical axis represents the pixel intensity values corresponding to lighter colored pixels in the Fourier transform image in Figure 2-4a. The tallest point on the vertical axis in the middle represents the DC-value for the image. The second two tallest points on each side of the DC-value represent the spatial frequency of the ridges in the image (indicated by the arrows).

Local Quality Score

As described earlier, the five variable values are calculated for each ROI in an image. Let x_*i* denote the i^{th} variable, with $i = 1\ldots4$ corresponding to S3PG, BS, MOW, and SF, respectively, and x_5 denote ACUT. The raw variable values for S3PG, BS, MOW, and SF are each normalized and scored using a symmetrical distribution scaled between 0 and 1 as provided by $f(x)$ in equation 2-3 below. A symmetrical distribution is used for these variables since a value that deviates too far on either side of the expected value indicates the presence of distorting artifacts in the ROI that may interfere with accurate detection of friction ridge detail.

$$f(x) = e^{\frac{-(x-\hat{\mu})^2}{2\hat{\sigma}^2}}$$

Equation 2-3: Scoring function for raw variable values S3PG, Bimodal Separation, Mean Object Width, and Spatial Frequency. The scoring function provides a maximum value of 1 if the raw variable value is equal to the expected value ($\hat{\mu}$) (i.e. location parameter). As the raw variable value deviates from the expected value on either side, the score is reduced and trends toward 0 at a rate determined by the scale parameter ($\hat{\sigma}$).

The raw variable value for ACUT is provided by a simple logistic cumulative distribution (scaled between 0 and 1) as provided by $g(x)$ in equation 2-4 below. A cumulative distribution is used for this variable since only values that are less than the expected value indicates the presence of lower sharpness and contrast of ridges in the ROI that may interfere with accurate detection of friction ridge detail.

$$g(x) = \frac{1}{1 + e^{-\frac{x-\hat{\mu}}{\hat{s}}}}$$

Equation 2-4: Scoring function for the raw variable value Acutance. The scoring function provides a maximum value of 1 if the raw variable value is equal to the expected value ($\hat{\mu}$) (i.e. location parameter). As the raw variable value deviates from the expected value (lower acutance values), the score is reduced and trends toward 0 at a rate determined by the scale parameter ($\hat{s}$). The Acutance is scored on a cumulative distribution since lower quality is only manifest with lower acutance values.

The input parameters for the scoring functions for each variable consist of the location parameter (i.e. mean raw variable value) and scale parameter (e.g. standard deviation of the raw variable value) empirically estimated from a reference dataset. The reference dataset consisted of 1,373 ROIs selected from pristine quality exemplar impressions. The impressions in this dataset were deposited under controlled conditions using a mixture of traditional ink and Livescan device. Table 2-1 provides the input parameters for the scoring function related to each variable.

Variable	Location Parameter ($\hat{\mu}$)	Scale Parameter ($\hat{\sigma}$ or $\hat{s}$)
S3PG	51.408	4.134
Bimodal Separation	0.843	0.147
Acutance	6.869	0.532
Mean Object Width	1.383	0.391
Spatial Frequency	2.078	0.397

Table 2-1: Input parameters for the scoring functions for each variable.

The five normalized variable values are then combined to create a mean univariate quantitative score summarizing the clarity and quality of the feature represented in the ROI on a scale from 0 to 1 (higher values indicate higher clarity and quality of the friction ridges in the ROI). This ROI score (i.e. Local Quality Score, or "LQS") provides a proxy estimate of the quality of the feature contained within the ROI on the basis of the clarity of the friction ridge detail immediately surrounding it. The LQS is calculated using the formula below:

$$LQS = \frac{\sum_{i=1}^{5} f(x)_i}{5}$$

Equation 2-5: Local Quality Score (LQS) function – calculated for each ROI as the mean of the normalized variable scores, where $f(x)_i$ is the normalized variable score for i-th function in the set containing the normalized variable scores for all 5 variables.

The LQS value is then used as a basis to categorize and color-code the quality of the feature as a graphical output to the user (e.g. high, medium, and low) in terms that align with subjective determinations by human analysts, such as that proposed by Langenburg & Champod (2011) [56]. Features color-coded as green generally indicate areas of high quality, features color-coded as yellow generally indicate areas of medium quality, and features color-coded as red generally indicate areas of low quality.

Global Quality Scores

Three different Global Quality Score (GQS) values are calculated, each of which represent a summary of the overall quality of the impression for different purposes: to predict analysts' determinations of "value," "complexity," and "difficulty" as proposed by Eldridge et al. (2020) [57] and as part of the Analysis phase of the examination methodology. For all three prediction categories (value, complexity, and difficulty), the GQS is calculated as a multinomial combination of two variables: (a) LQS_{sum} – the sum of all LQS values, and (b) *n*FEAT – the total quantity of features identified in the impression. Taken together, these provide explainable quantitative representations and variables of the overall quality of the impression for manual comparison purposes.

The multinomial coefficients for each outcome class (value, complexity, and difficulty) were derived using a multinomial regression model provided by the *nnet* package in R [58] against a training/test-dataset of feature measurements from impressions for which latent print examiners previously analyzed and categorized based on their "value," "complexity," and "difficulty" for comparison. The multinomial model was selected after testing a range of machine learning techniques with the variables LQS_{sum} and *n*FEAT (naïve based classifier, tree-based classifiers, discriminant analysis techniques, neural networks and support vector machines). Overall, the multinomial regression offers a competitive accuracy while maintaining easy explainability (see Appendix B-1 for raw model diagnostics and uncertainty values). The training-dataset was derived as a random 50/50 training-test split obtained from the full dataset provided by Eldridge et al. (2020) [57]. The full dataset consisted of a total of 3,241 determinations made by 116 analysts rendering "value," "complexity," and "difficulty" decisions for each image they viewed from a set of 100 different latent print impressions – each participant was provided a set of approximately 30 impressions to analyze, resulting in each impression being analyzed by between 26 and 41 different analysts. The impressions were generated during the course of normal casework at a large metropolitan police laboratory using standard powder processing and lifting techniques. All participants were practicing latent print examiners recruited by several outreach methods, such as email distribution lists, presentations given at professional educational meetings, and professional contacts. Half of this dataset was used to train the models (1,621 responses) and the other half of this dataset was used to test the models (1,620 responses) described by GQS Test-Dataset 1 below.

It should be noted that the model was trained and tested using the results of each examiner's individual observations and judgments of the impressions rather than artificially combining them. Ground truth for these types of judgments is non-existent. Although consensus judgments could be declared as a surrogate to ground truth for each *image*, the examiners' observations for which their subjective judgments are based are variable which would require artificially aggregating examiners' judgments and disconnecting their individual observations from their individual judgments. As a result, the authors believe a model that is trained using individual examiners' observations and resulting judgments is appropriate in this context. The output of the model, effectively, then reflects a proxy consensus of examiners' judgments for a given input in a specific case impression. Tables 2-2a through 2-2c provide the coefficients related to the multinomial models from the training partition (see Appendix B-1 for raw model diagnostics and uncertainty values on the coefficients). Each multinomial model provides a probability of class inclusion (ranging from 0.00 to 1.00) for each outcome class (e.g., for the Value determination the three outcome classes are no-value, value for exclusion only, and value for identification).

"Value" coefficients	**Intercept**	**LQS_{sum}**	***n*FEAT**
No Value	0.000	0.000	0.000
Value for Exclusion	-1.736	-0.051	0.277
Value for Identification	-6.042	0.495	0.726

Table 2-2a: Multinomial coefficients for each outcome class probability (no-value, value for exclusion only, value for identification) of the "value" determination. Note: In Eldridge et al. [57], participants were given the following response choices: "no value," "some probative or investigative value but insufficient for identification or exclusion," "value for exclusion only," "value for identification only," "value for both identification and exclusion." Responses of "some probative or investigative value but insufficient for identification or exclusion" were categorized as "value for exclusion" to represent the middle bin of the value spectrum. Responses of "value for both identification and exclusion" and "value for identification only" were categorized as "value for identification."

"Complexity" coefficients	**Intercept**	**LQS_{sum}**	***n*FEAT**
Highly Complex	3.325	-0.100	-0.459
Complex	0.000	0.000	0.000
Non-Complex	-1.781	0.741	-0.025

Table 2-2b: Multinomial coefficients for each outcome class probability (highly complex, complex, non-complex) of the "complexity" determination. Note: In Eldridge et al. [57], participants were given the following response choices: "no value," "of value, complex," "of value, non-complex; requiring documentation," and "of value, non-complex; self-evident." Responses of "of value, non-complex; requiring documentation" and "of value, non-complex; self evident" were both categorized as "non-complex." Responses of "no value" were re-labeled "highly complex" to represent the extreme end of the complexity spectrum.

"Difficulty" coefficients	Intercept	LQS_{sum}	*n*FEAT
High	0.000	0.000	0.000
Medium	-1.896	0.289	0.125
Low	-3.071	0.965	-0.004

Table 2-2c: Multinomial coefficients for each outcome class probability (high, medium, low) of the "difficulty" determination.

Recognizing each class represents an outcome along a spectrum (e.g. for the "value" determination: No Value represents the left-most extreme and Value for Identification represents the right-most extreme) and the sum across all classes equals 1.00, we can combine to create single values representing the GQS for each determination (value, complexity, difficulty) by subtracting the probability of class inclusion representing the left-most extreme from the probability of class inclusion representing the right-most extreme to produce a number ranging from -1.00 to 1.00, where higher values indicate stronger support for "value for identification," "non-complex," and "low difficulty" and lower values indicate stronger support for "no value," "highly complex," and "high difficulty." The GQS values for each determination are calculated using the formulae below:

$$Value_{GQS} = p(VID) - p(NV)$$

Equation 2-6: GQS function for Value determination – calculated by subtracting the probability of class inclusion for No Value outcome (NV) from the probability of class inclusion for Value for Identification outcome (VID). Values near -1.00 indicate no-value determinations, values near 1.00 indicate value for identification determinations, and values near 0 indicate value for exclusion only determinations (or inconclusive determinations in lieu of value for exclusion only).

$$Complexity_{GQS} = p(NC) - p(HC)$$

Equation 2-7: GQS function for Complexity determination – calculated by subtracting the probability of class inclusion for Highly Complex outcome (HC) from the probability of class inclusion for Non-Complex outcome (NC). Values near -1.00 indicate no-value determinations, values near 1.00 indicate non-complex determinations, and values near 0 indicate complex determinations.

$$Difficulty_{GQS} = p(L) - p(H)$$

Equation 2-8: GQS function for Difficulty determination – calculated by subtracting the probability of class inclusion for High difficulty outcome (H) from the probability of class inclusion for Low difficulty outcome (L). Values near -1.00 indicate high difficulty determinations, values near 1.00 indicate low difficulty determinations, and values near 0 indicate medium difficulty determinations.

ROC curves will be used to illustrate model performance. The associated areas under the curve (AUC), and confidence intervals have been computed taking advantage of the pROC package [59].

Method Performance

The performance of the method was evaluated in different conditions capturing performance characteristics both locally and globally. The local performance characteristics were

evaluated in terms of (i) the ability of the LQS value to accurately distinguish between the extreme conditions of "good" and "bad" quality ROIs and (ii) the ability of the LQS value to predict analysts' subjective determinations of feature quality according to the GYRO annotation scheme proposed by Langenburg & Champod (2011) [56]. The global performance characteristics were evaluated in terms of the ability of the GQS values to distinguish between analysts' subjective determinations of "value," "complexity," and "difficulty" from test-datasets of feature measurements from impressions for which latent print examiners previously analyzed and categorized.

Local Performance Characteristics:

The local performance characteristics were evaluated to understand the behavior of the system as the clarity of friction ridge detail within the ROIs change. This was evaluated using measurements from two different test-datasets:

(1) LQS-Test-Dataset-1: This dataset consists of 867 "good" quality ROIs selected from high quality regions of exemplar friction ridge impressions and a dataset of 3,816 "bad" quality ROIs selected from low quality regions of latent lift cards submitted under operational conditions as attempts to lift latent images from a variety of different surfaces during normal forensic casework. The "bad" quality ROIs represented impressions with excessive smudging, indiscernible ridge detail, background interference and artifacts, and related factors impacting reliable interpretation of friction ridges, yet still having artifacts present bearing reasonable contrast and clarity but lacking morphological representations of friction ridge detail. The purpose of this dataset is to evaluate how well the LQS values distinguish between the extremes of "good" and "bad" quality ROIs collected under operational conditions.

(2) LQS-Test-Dataset-2: This dataset consists of 4,480 ROIs containing features annotated as "high confidence" (i.e. green) and 920 ROIs containing features annotated as "medium confidence" (i.e. yellow) by practicing latent print examiners according to the GYRO annotation scheme proposed by Langenburg & Champod (2011) [56] across 300 different impressions deposited using normal handling of objects and developed using common latent print processing methods representative of typical casework. This dataset was obtained from John & Swofford (2020) [60]. The purpose of this dataset is to evaluate how well the LQS color-coded quality categories correspond to fingerprint experts' subjective assessment of feature confidence ("high" confidence vs. "medium" confidence).

Global Performance Characteristics:

The global performance characteristics were evaluated to understand the ability of the method to predict human analysts' subjective assessments of whether impressions are considered "suitable" or "of value" as well as assessments of "complexity" and "difficulty" for comparison purposes. These were evaluated using measurements from two different test-datasets:

(1) GQS-Test-Dataset-1: This dataset represents the test fraction derived as a random 50/50 training-test split of the full dataset obtained from Eldridge et al. (2020) [57]. The full dataset consisted of a total of 3,241 analysts' determinations of "value," "complexity," and "difficulty" and documentation of features across a set of 100 different latent print impressions by approximately 116 different participants – each participant was provided a set of approximately 30 impressions to analyze resulting in each impression being analyzed by between 26 and 41 different analysts. The impressions were generated during the course of normal casework at a large metropolitan police laboratory using a variety of standard processing techniques. All participants were practicing latent print examiners recruited by several outreach methods, such as email distribution lists, presentations given at professional educational meetings, and professional contacts. Half of this dataset was used to train the models (1,621 responses) and the other half of this dataset was used to test the models (1,620 responses). The purpose of this dataset is to evaluate how well the GQS values correspond to subjective determinations of value, complexity, and difficulty when examined under pseudo-operational conditions.

(2) GQS-Dataset-2: This dataset consists of 605 latent impressions collected from casework during the course of routine operations by fingerprint experts in a federal crime laboratory in the United States for which fingerprint experts conducted examinations and identified the impressions to corresponding reference standards. All impressions in this dataset were determined to be "suitable" or "of value" for identification purposes. The purpose of this dataset is to evaluate the distribution of GQS values and implications thereof when applied to impressions derived from actual casework and assessed under normal operational conditions.

2.1.4 Results & Discussion

Local Performance Characteristics

The local performance characteristics were evaluated on the basis of how well the LQS values were able to distinguish between the extremes of "good" and "bad" quality ROIs collected under operational conditions using LQS-Test-Dataset-1 and how well the LQS color-coded quality categories correspond to fingerprint experts' subjective assessment of feature confidence ("high" confidence vs. "medium" confidence) using LQS-Test-Dataset-2. Figures 2-5a and 2-5b illustrates the degree of separation observed between the extremes of "Good" and "Bad" quality ROIs using the LQS value.

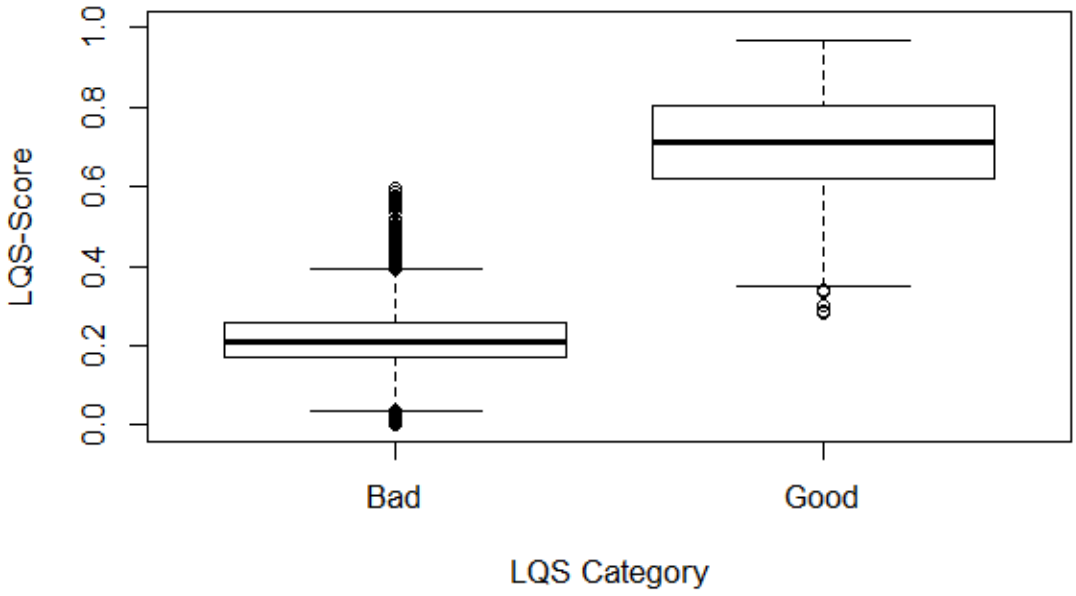


Figure 2-5a: Boxplot of LQS values for "Bad" (n = 3,816) and "Good" (n = 867) quality ROIs from LQS-Test-Dataset-1.

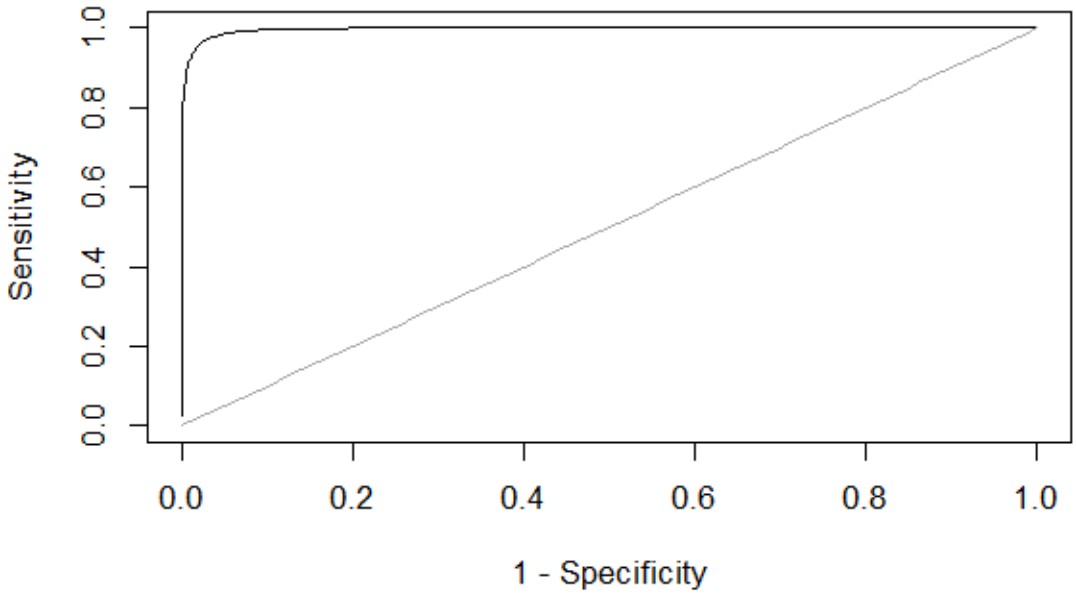


Figure 2-5b: Receiving Operating Characteristic (ROC) curve of LQS values for "Bad" (n = 3,816) and "Good" (n = 867) quality ROIs from LQS-Test-Dataset-1. The area under the curve (AUC) is 99.7% with a 95% confidence interval of (99.6% - 99.8%).

From Figures 2-5a and 2-5b, we see remarkable separation between the two extremes of "Good" and "Bad" quality ROIs. Although these results may be expected for this dataset since they represent the extreme ends of the spectrum, it establishes an important baseline which validates the relevance of the input variables which comprise the LQS value and its ability to distinguish between high-quality friction ridge impressions and low-quality non-friction ridge artifacts. Further, from these data, we can establish thresholds for distinguishing between "high," "medium," and "low" color-coded bins categorizing ROI quality as an overlay output to the user. For this purpose, LQS values between 0.35 and 1.00 are color-coded green (high quality), LQS values between 0.20 and 0.35 are color coded yellow (medium quality), and LQS values between 0.00 and 0.20 are color-coded red (low quality). Using this color-coding scheme, Table 2-3 provides the distribution of "Good" and "Bad" quality ROIs categorized as green, yellow, and red.

ROI Quality / LQS Color Code	Good	Bad	Total
Green	862	318	1,180
Yellow	5	1,892	1,897
Red	0	1,606	1,606
Total	867	3,816	4,683

Table 2-3: Number of LQS values color-coded as green, yellow, and red compared for "Good" and "Bad" quality ROIs using LQS-Test-Dataset-1. LQS values between 0.35 and 1.00 are color-coded green (high quality), LQS values between 0.20 and 0.35 are color coded yellow (medium quality), and LQS values between 0.00 and 0.20 are color-coded red (low quality).

Having established the baseline performance of the LQS values to distinguish between "Good" and "Bad" quality ROIs and a threshold for categorizing as "high," "medium," or "low" quality (i.e. green, yellow, red), we can use LQS-Test-Dataset-2 to evaluate how well the color-coding output correspond to fingerprint experts' subjective assessment of feature quality ("high" quality vs. "medium" quality due to insufficient annotations of "low" quality features in the dataset). Table 2-4 demonstrates the consistency between automated predictions of quality using the LQS color-code scheme and experts' subjective judgments.

Expert Judgement / LQS Color Code	Green	Yellow	Total
Green	3,077	450	3,527
Yellow	1,119	348	1,467
Red	284	122	406
Total	4,480	920	5,400

Table 2-4: Number of LQS values color-coded as green, yellow, and red compared to experts' subjective judgments of feature quality / confidence using GYRO [56] using LQS-Test-Dataset-2. LQS values between 0.35 and 1.00 are color-coded green (high quality), LQS values between 0.20 and 0.35 are color coded yellow (medium quality), and LQS values between 0.00 and 0.20 are color-coded red (low quality). NOTE: As discussed by John & Swofford (2020) [60] from which this dataset was obtained, experts mostly only annotated features as green and yellow. Experts rarely annotated features as low quality (red), thus those data were insufficient for this assessment.

From Table 2-4, we see that approximately 94% of the features annotated by experts as green (high quality) were categorized by the LQS color-code scheme as either green (69%) or yellow (25%). Approximately 6% of the features annotated by experts as green were categorized by the LQS color-code scheme as red. Of the features annotated by experts as yellow (medium quality), approximately 87% were categorized by the LQS color-code scheme as green (49%) or yellow (38%). Approximately 13% of the features annotated by experts as yellow were categorized by the LQS color-code scheme as red. Although not perfect correspondence between green vs. green and yellow vs. yellow (which may be expected given the variable nature of experts' judgements), these results indicate reasonable agreement between experts' subjective assessments of feature quality and LQS color-coded classifications as it relates to general groupings of medium or high-quality features. Taken together, among the 5,400 total features annotated as either green or yellow by experts' subjective judgments, approximately 93% were categorized as either green or yellow by the LQS color-code scheme. Recognizing the variability in subjective judgments of feature quality (e.g. green-yellow or yellow-red), the most significant contribution of the LQS color-code scheme is the ability for it to provide a standardized framework for establishing consistency between examiners related to the relative contribution of features for comparison and flag conditions warranting additional quality assurance review such as those situations where examiners' judgments and the LQS color-code scheme contradict each other on the extreme ends of the spectrum (e.g. green vs. red). While the local performance characteristics are important, the global performance characteristics have the most significant impact on the ultimate outcome of the examination.

Global Performance Characteristics

The global performance characteristics were evaluated on the basis of how well the GQS values correspond to analysts' subjective assessments of "value," "complexity," and "difficulty" using a dataset representing casework-like conditions (GQS-Test-Dataset-1). The implications of applying GQS values to impressions under operational conditions is further explored using a dataset derived directly from casework (GQS-Test-Dataset-2). Each dataset is evaluated separately so that the results can be considered within context of the conditions from which the datasets were obtained (e.g. casework-like conditions vs. casework conditions).

"Value" Determinations

The $Value_{GQS}$ score is calculated by equation 2-6 and can range from -1.0 to 1.0. Values near -1.0 indicate the impression is "not suitable" or "no value" and thus should not proceed for further comparison or should do so with caution and additional quality assurance safeguards in place. Values near 1.0 indicate the impression is "suitable" or "of value for identification" and may proceed for further comparison in accordance with normal operating protocols. Figure 2-6 illustrates how well the $Value_{GQS}$ score correspond to experts' subjective judgments of impressions deemed to be "no value" (n = 252), "value for exclusion only" (n = 227), or "value for identification" (n = 1,141).

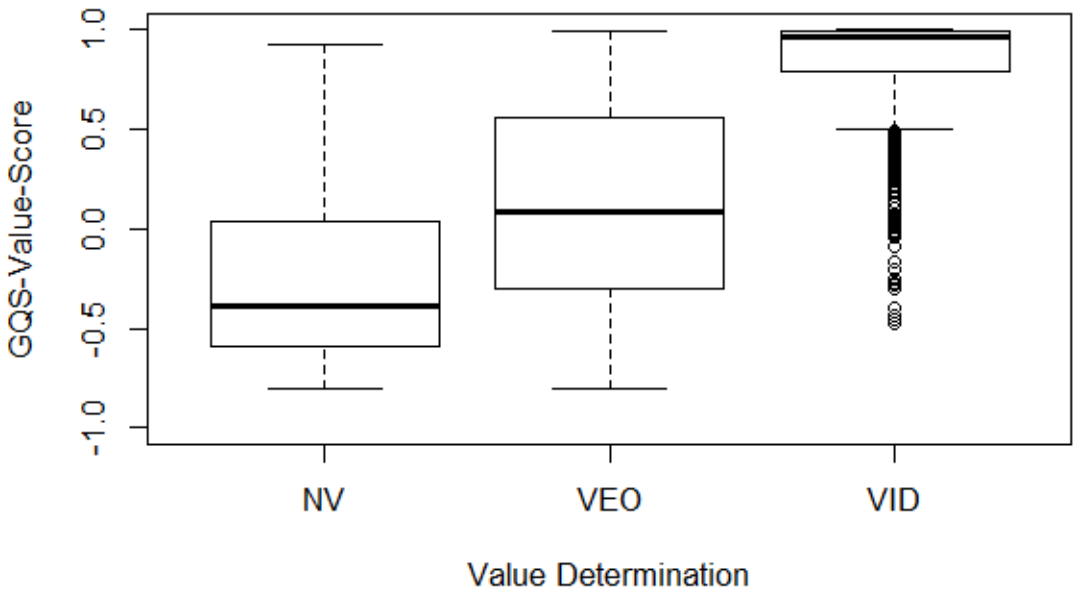


Figure 2-6: Boxplot of $Value_{GQS}$ scores for impressions subjectively judged by experts to be "no value" (NV) (n = 252), "value for exclusion only" (VEO) (n = 227), or "value for identification" (VID) (n = 1,141) from GQS-Test-Dataset-1.

From Figure 2-6, we see the $Value_{GQS}$ score is able to reasonably distinguish between impressions determined to be "no value" and "value for identification," which represent the ends of the value spectrum. There is overlap between the classes; however, the results demonstrate a trend consistent with expectations—the majority of impressions judged as "VID" have higher values compared to those judged as "NV." The impressions deemed "value for exclusion only" represent a broad span of $Value_{GQS}$ scores and are more difficult to predict. This is understandable, however, since the impressions deemed "value for exclusion only" represent the broad category of impressions in the middle of the value spectrum for which disagreement between examiners was most significant. Looking closer at the inter-rater reliability across the full dataset (train and test

partitions combined), *none* of the impressions resulted in consensus determination (defined as two-thirds agreement among participants) for this decision outcome. Consequently, and more practically in an operational setting, the $Value_{GQS}$ score has greater applicability to predicting whether an impression should be categorized as "no value" or "value for identification" and the lack of support for one of those categories should be indicative of the potential for disagreements between experts' interpretations in the middle of the spectrum, thus triggering the impression to be raised for further quality assurance review. Figure 2-7 illustrates the performance of the $Value_{GQS}$ score when distinguishing against those impressions determined to be "no value" and "value for identification" using the receiver operator characteristic (ROC). Table 2-5 demonstrates the performance tradeoff when different threshold values are applied.

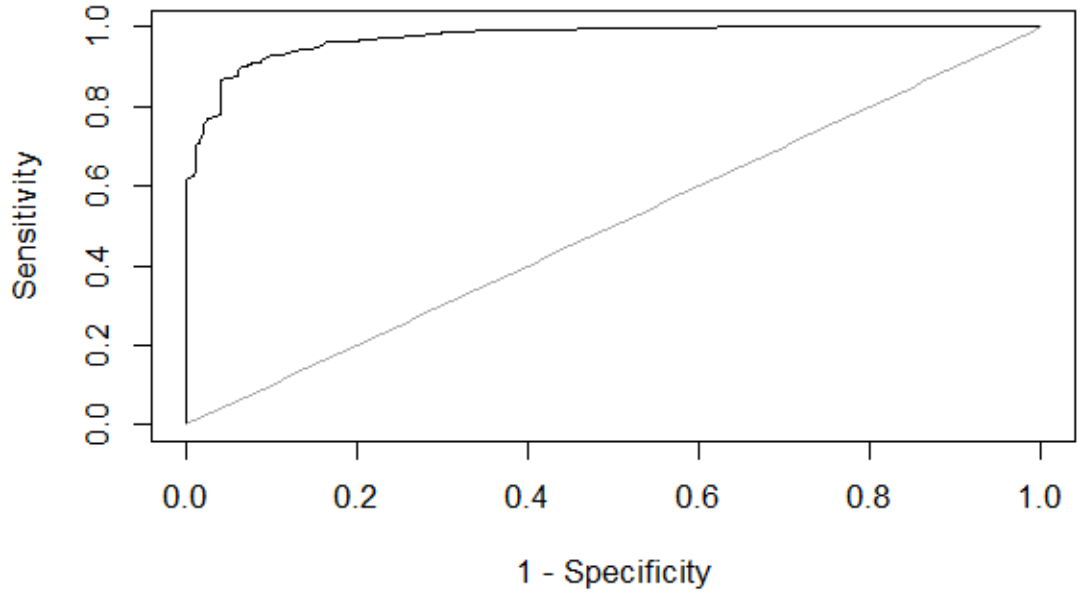


Figure 2-7: Receiving Operating Characteristic (ROC) curve of $Value_{GQS}$ scores for impressions subjectively judged by experts to be "no-value" (n = 252) and "value for identification" (n = 1,141) from GQS-Test-Dataset-1. The area under the curve (AUC) is 97.3% with a 95% confidence interval of (96.5% - 98.2%).

Threshold $Value_{GQS}$	**"No Value"**	**"Value for Identification"**
-0.50	0.623 (0.563-0.683)	1.00 (1.00-1.00)
-0.33	0.484 (0.425-0.548)	0.996 (0.991-0.999)
-0.25	0.405 (0.345-0.464)	0.992 (0.987-0.996)
0.00	0.274 (0.218-0.329)	0.979 (0.97-0.987)
0.25	0.159 (0.115-0.206)	0.954 (0.942-0.966)
0.33	0.127 (0.087-0.171)	0.938 (0.924-0.952)
0.50	0.063 (0.036-0.095)	0.895 (0.876-0.912)

Table 2-5: Proportion of responses resulting in $Value_{GQS}$ score greater than threshold values (-0.50, -0.33, -0.25, 0.00, 0.25, 0.33, 0.50) and assessed as "no-value" (n = 252) and "value for identification" (n = 1,141) by experts from GQS-Test-Dataset-1. Confidence intervals are indicated (lower CI - upper CI).

"Complexity" Determinations

The Complexity_{GQS} score is calculated by equation 2-7 and can range from -1.0 to 1.0. Values near -1.0 indicate the impression is "not suitable" or "highly complex" and thus should only proceed to comparison with caution and additional quality assurance safeguards in place. Values near 1.0 indicate the impression is "non-complex" and may proceed for further comparison in accordance with normal operating protocols. Figure 2-8 illustrates how well the Complexity_{GQS} score corresponds to experts' subjective judgments of impressions deemed to be "no value" or "highly complex" (n = 291), "complex" (n = 452), or "non-complex" (n = 877).

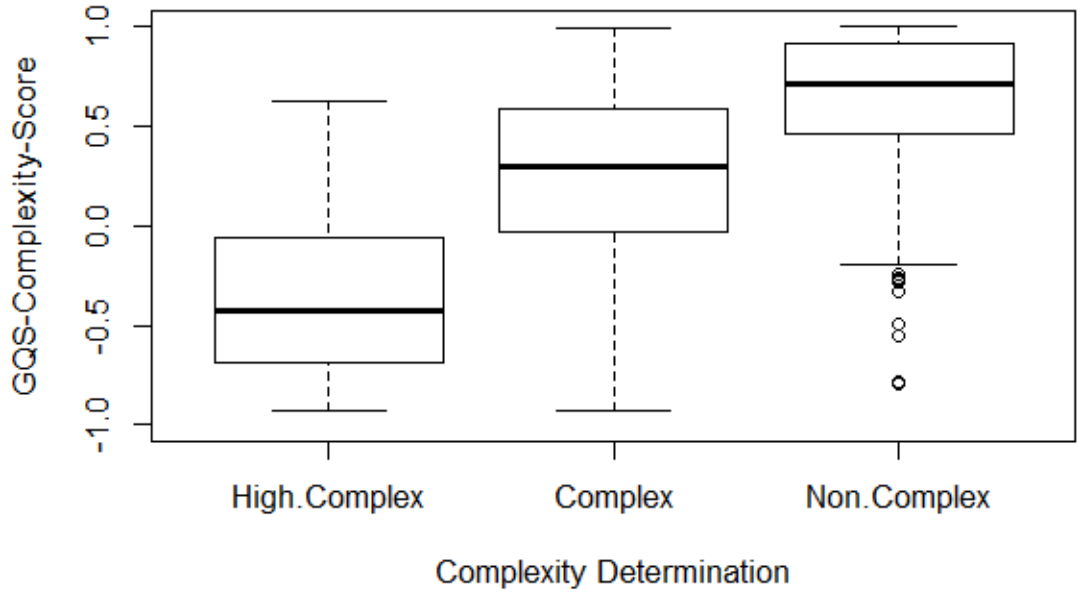


Figure 2-8: Boxplot of Complexity_{GQS} scores for impressions subjectively judged by experts to be "highly complex" (n = 291), "complex" (n = 452), or "non-complex" (n = 877) from GQS-Test-Dataset-1.

It transpires from Figure 2-8, that the Complexity_{GQS} score is able to reasonably distinguish between impressions determined to be "highly complex" and "non-complex," which represent the ends of the complexity spectrum. There is overlap between the classes; however, the results demonstrate a trend consistent with expectations—the majority of impressions judged as "non-complex" have higher values compared to those judged as "highly complex." The impressions deemed "complex" represent a broad span of Complexity_{GQS} scores and are more difficult to predict. Similar to the "value" spectrum, this is understandable since impressions deemed "complex" represent the broad category of impressions in the middle of the complexity spectrum for which disagreement between examiners was most significant. Consequently, and more practically in an operational setting, the Complexity_{GQS} score has greater applicability to predicting whether an impression should be categorized as "highly complex" or "non-complex" and the lack of support for one of those categories should be indicative of the potential for disagreements

between experts' interpretations in the middle of the spectrum, thus triggering the impression to be raised for further quality assurance review. Figure 2-9 illustrates the performance of the $Complexity_{GQS}$ score when distinguishing against those impressions determined to be "highly complex" and "non-complex" using the receiver operator characteristic (ROC). Table 2-6 demonstrates the performance tradeoff when different threshold values are applied.

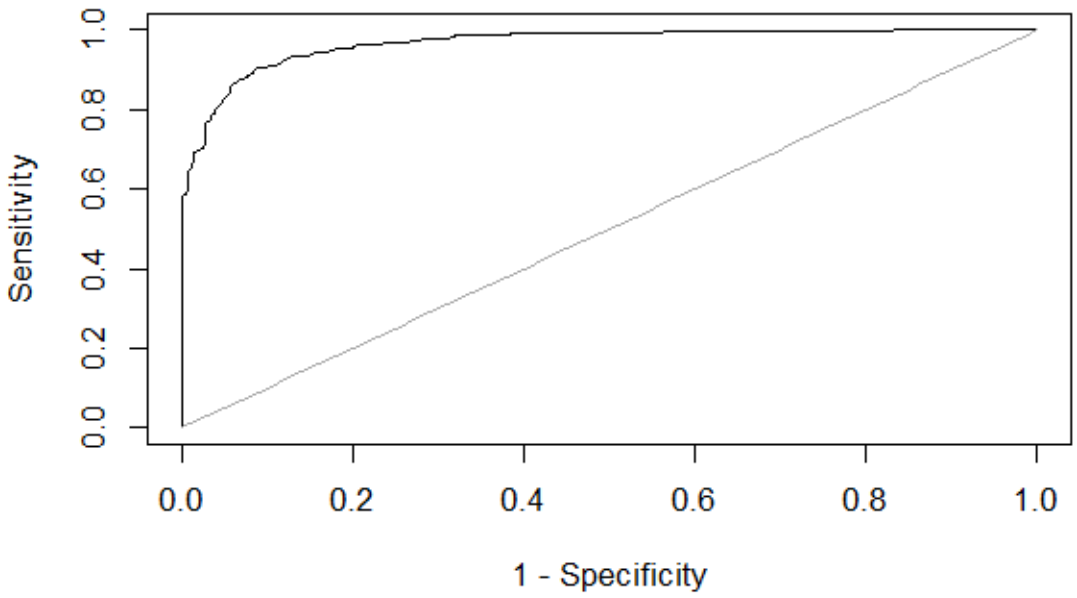


Figure 2-9: Receiving Operating Characteristic (ROC) curve of $Complexity_{GQS}$ scores for impressions subjectively judged by experts to be "highly complex" (n = 291) and "non-complex" (n = 877) from GQS-Test-Dataset-1. The area under the curve (AUC) is 96.8% with a 95% confidence interval of (95.9% - 97.7%).

Threshold $Complexity_{GQS}$	"Highly Complex"	"Non-Complex"
-0.50	0.570 (0.512-0.625)	0.997 (0.992-1.00)
-0.33	0.419 (0.364-0.478)	0.994 (0.989-0.999)
-0.25	0.378 (0.323-0.433)	0.989 (0.981-0.995)
0.00	0.206 (0.162-0.254)	0.962 (0.950-0.975)
0.25	0.076 (0.048-0.107)	0.886 (0.864-0.906)
0.33	0.055 (0.031-0.082)	0.854 (0.830-0.877)
0.50	0.027 (0.010-0.048)	0.717 (0.688-0.747)

Table 2-6: Proportion of responses resulting in $Complexity_{GQS}$ score greater than threshold values (-0.50, -0.33, -0.25, 0.00, 0.25, 0.33, 0.50) and assessed as "highly complex" (n = 291) and "non-complex" (n = 877) by experts from GQS-Test-Dataset-1. Confidence intervals are indicated (lower CI - upper CI).

"Difficulty" Determinations

The Difficulty_{GQS} score is calculated by equation 2-8 and can range from -1.0 to 1.0. Values near -1.0 indicate the impression is "high difficulty" and thus should only proceed to comparison with caution and additional quality assurance safeguards in place. Values near 1.0 indicate the impression is "low difficulty" and may proceed for further comparison in accordance with normal operating protocols. Figure 2-10 illustrates how well the Difficulty_{GQS} score corresponds to experts' subjective judgments of impressions deemed to be "high difficulty" (n = 487), "medium difficulty" (n = 556), or "low difficulty" (n = 577).

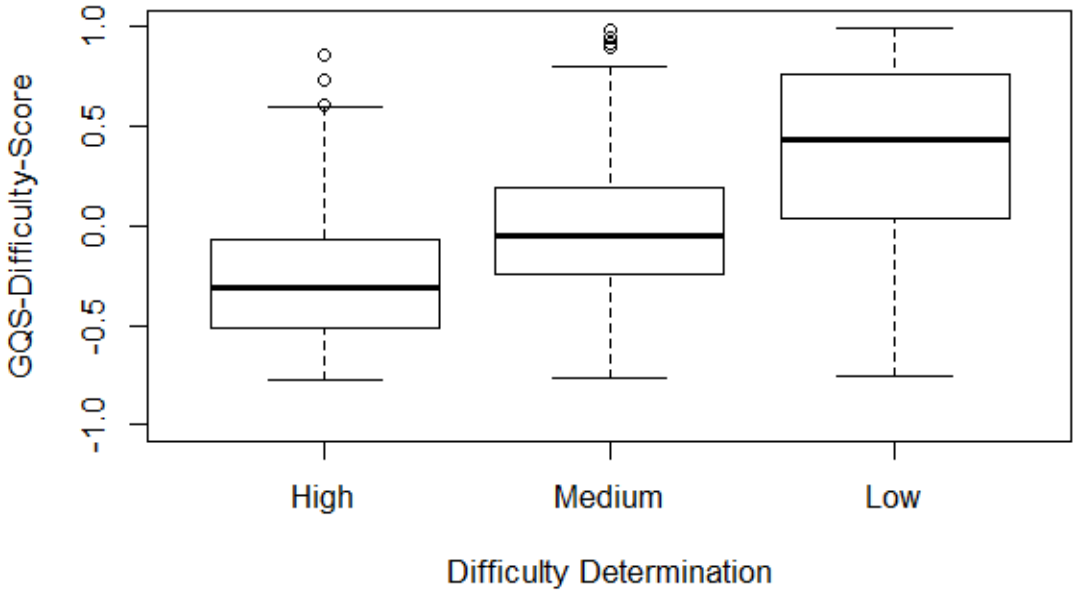


Figure 2-10: Boxplot of Difficulty_{GQS} scores for impressions subjectively judged by experts to be "high difficulty" (n = 487), "medium difficulty" (n = 556), or "low difficulty" (n = 577) from GQS-Test-Dataset-1.

From Figure 2-10, we see the Difficulty_{GQS} score is able to generally distinguish between impressions determined to be "high difficulty" and "low difficulty," which represent the ends of the difficulty spectrum. There is overlap between the classes; however, the results demonstrate a trend consistent with expectations—the majority of impressions judged as "low" difficulty have higher values compared to those judged as "high" difficulty. The impressions deemed "medium difficulty" represent a broad span of Difficulty_{GQS} scores and are more difficult to predict. Similar to the "value" and "complexity" spectrums, this is understandable since impressions deemed "medium difficulty" represent the broad category of impressions in the middle of the spectrum for which disagreement between examiners was most significant. Consequently, and more practically in an operational setting, the Difficulty_{GQS} score has greater applicability to predicting whether an impression should be categorized as "high difficulty" or "low difficulty" and the lack of support

for one of those categories should be indicative of the potential for disagreements between experts' interpretations in the middle of the spectrum, thus triggering the impression to be raised for further quality assurance review. Figure 2-11 illustrates the performance of the $Difficulty_{GQS}$ score when distinguishing against those impressions determined to be "high difficulty" and "low difficulty" using the receiver operator characteristic (ROC). Table 2-7 demonstrates the performance tradeoff when different threshold values are applied.

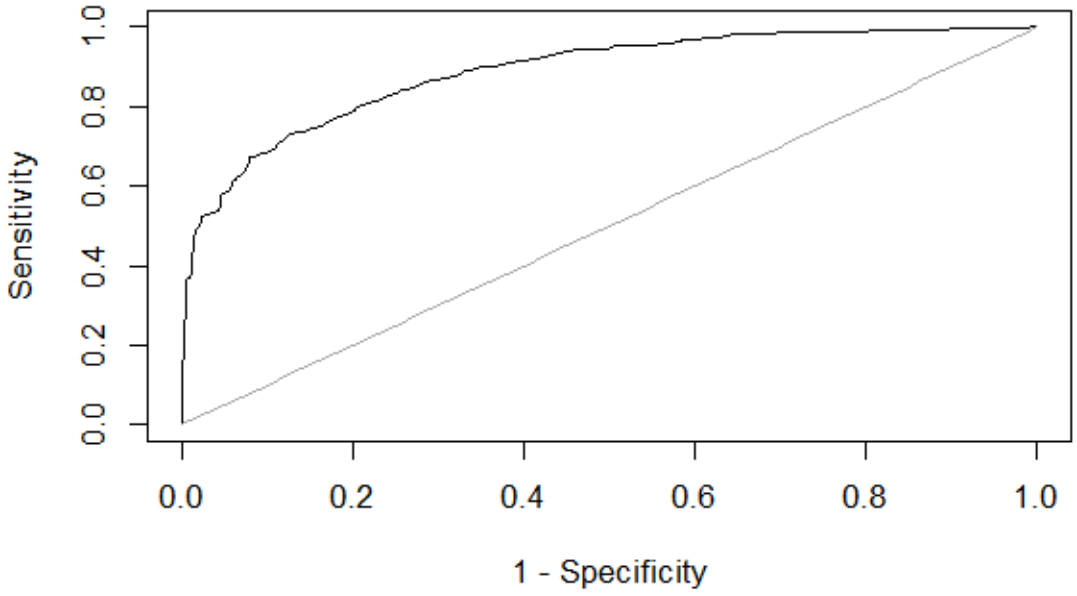


Figure 2-11: Receiving Operating Characteristic (ROC) curve of $Difficulty_{GQS}$ scores for impressions subjectively judged by experts to be "high difficulty" (n = 487) and "low difficulty" (n = 577) from GQS-Test-Dataset-1. The area under the curve (AUC) is 88.8% with a 95% confidence interval of (86.9% - 90.7%).

Threshold $Difficulty_{GQS}$	**"High Difficulty"**	**"Low Difficulty"**
-0.50	0.729 (0.690-0.768)	0.986 (0.976-0.995)
-0.33	0.515 (0.470-0.561)	0.953 (0.936-0.969)
-0.25	0.415 (0.372-0.458)	0.922 (0.899-0.943)
0.00	0.193 (0.158-0.228)	0.782 (0.747-0.815)
0.25	0.057 (0.037-0.080)	0.610 (0.570-0.650)
0.33	0.045 (0.029-0.064)	0.555 (0.515-0.594)
0.50	0.012 (0.004-0.023)	0.449 (0.409-0.490)

Table 2-7: Proportion of responses resulting in $Difficulty_{GQS}$ score greater than threshold values (-0.50, -0.33, -0.25, 0.00, 0.25, 0.33, 0.50) and assessed as "high difficulty" (n = 487) and "low difficulty" (n = 577) by experts from GQS-Test-Dataset-1. Confidence intervals are indicated (lower CI - upper CI).

Casework Evaluation

From GQS-Test-Dataset-1, we see that the GQS values are able to reasonably distinguish between impressions on the ends of the value, complexity, and difficulty spectra thus indicating those impressions which may proceed to further comparison in accordance with normal operational protocols versus those impressions which may be flagged for further quality assurance review and additional safeguards. Having established the baseline performance characteristics under case-work like conditions, we can consider the implications if this quality metric were to be applied in an operational setting on actual casework to demonstrate the distribution of GQS values and potentially indicate the need for intervention from a quality assurance perspective when GQS values fall below an established threshold. To evaluate this, we use GQS-Test-Dataset-2, which consists of 605 impressions that were deemed "value for identification" by experts' subjective judgements (and subsequently identified to exemplar impressions). Although this dataset does not include those impressions deemed to be "no value" since operational procedures did not require retention of annotated images for that outcome category, we can consider the proportion of impressions for which the determination of "value for identification" was supported. Similarly, despite the impressions not being pre-categorized against the complexity spectrum or difficulty spectrum, we can visualize the distribution of the impressions against each metric for general context.

Figure 2-12 illustrates the distribution of $Value_{GQS}$ scores, $Complexity_{GQS}$ scores, and $Difficulty_{GQS}$ scores for the GQS-Test-Dataset-2.

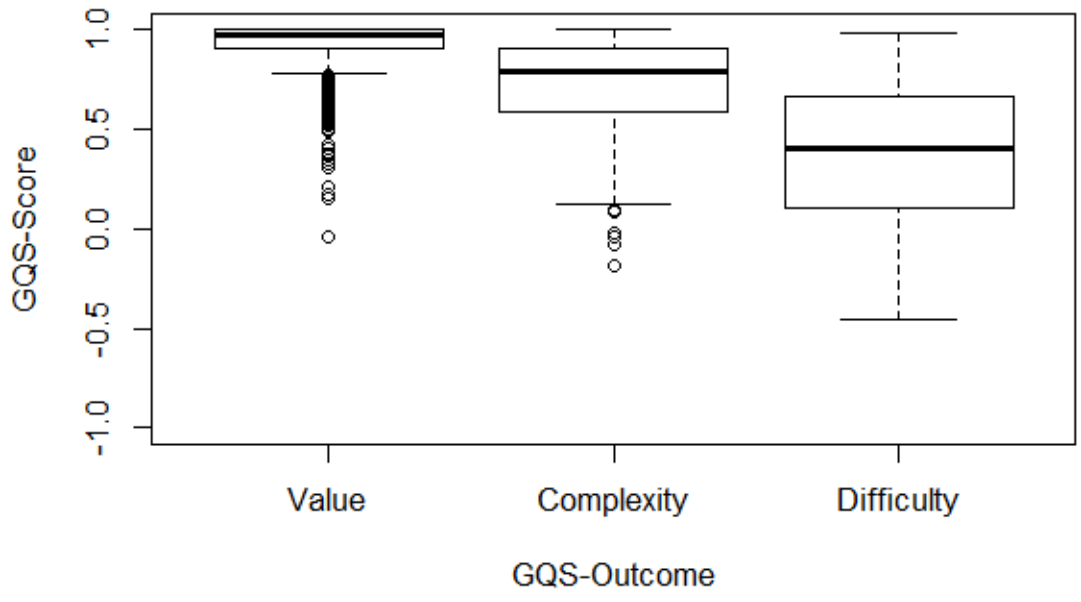


Figure 2-12: Boxplot of $Value_{GQS}$, $Complexity_{GQS}$, and $Difficulty_{GQS}$ scores for 605 impressions subjectively judged by experts to be "value for identification" and subsequently identified to exemplar impressions during normal casework conditions from GQS-Test-Dataset-2.

If we were to apply threshold values to the GQS metrics to evaluate how often the experts' assessment of "value for identification" was supported or to indicate circumstances in which the impressions may be flagged for additional quality assurance review, we can consider the implications to practice more clearly. For the Value determination, Table 2-5 suggests a $Value_{GQS}$ score of 0.50 is a reasonable threshold. For the Complexity determination, Table 2-6 suggests a $Complexity_{GQS}$ score of 0.33 is a reasonable threshold. For the Difficulty determination, Table 2-7 suggests a $Difficulty_{GQS}$ score of 0.00 is a reasonable threshold. Table 2-8 provides the proportion of impressions for which normal procedures are sufficient and those for which additional quality assurance review may be considered based on the results of the GQS metrics. From these data, we see reasonably strong support for experts' subjective judgement on the casework sample (GQS-Test-Dataset-2) and only a small percentage of impressions for which additional quality assurance review might be considered (~2% lacking support for value, ~6% categorized as complex, and ~16% categorized as difficult).

GQS Metric	Proportion of Cases with Normal Procedures Warranted	Proportion of Cases to Consider Additional Quality Assurance Review
Value	0.977	0.023
Complexity	0.942	0.058
Difficulty	0.843	0.157

Table 2-8: Proportion of impressions for which normal procedures are warranted and those for which additional quality assurance review may be considered based on the results of the GQS metrics from GQS-Test-Dataset-2 (n = 605) and the following thresholds: $Value_{GQS}$ scores less than 0.50, $Complexity_{GQS}$ scores less than 0.33, and $Difficulty_{GQS}$ scores less than 0.00. Note: GQS-Test-Dataset-2 is a dataset of impressions taken from a single federal laboratory in the United States which were considered "value for identification" and subsequently identified to exemplar impressions. Given the lack of quantifiable standards for "value for identification" at the time these impressions were examined, the extent to which these results are generalizable is unclear.

General Discussion

The method proposed provides three different quality metrics which can be used as a means to provide empirical support to experts' subjective assessments and a framework for establishing policies and procedures to flag impressions warranting further quality assurance review. Determinations of "value" have been considered by the friction ridge discipline for decades and are familiar to all practicing examiners. Determinations of "complexity" and "difficulty," however, are more recent terms to categorize impressions which tend to have lower quality and quantity of features and are therefore more susceptible to erroneous outcomes. With limited time and resources due to growing backlogs and operational demands, it is critical to have a means of focusing efforts on those impressions most vulnerable to errors or may require additional quality control measures. This method provides a means of accomplishing this goal in a more objective, transparent, and consistent fashion grounded by empirical validation. Although ground truth is non-existent for determinations of "value," "complexity," and "difficulty," the results demonstrate

reasonable agreement to experts' subjective assessments and illustrate a consistent general trend of increasing GQS values across the ordinal scale of "value," "complexity," and "difficulty" determinations. Having these quantitative outputs along ordinal scales, further work could enable a visual illustration and representation of the overall quality of an impression in three-dimensional space based on axes of "value," "complexity," and "difficulty."

Two important limitations for this method remain. First, the LQS and GQS values are dependent upon the subjective detection and annotation of friction ridge skin features (minutiae) by the human expert. Second, the method relies on clarity attributes of friction ridge minutiae and does not consider all of the attributes that experts may consider when making subjective determinations, such as pattern type, feature type, rarity of features and their configurations, continuity of ridge detail between features, and other types of features (non-minutiae) available.

To attenuate these limitations, two general recommendations for policy and procedure could be considered. First, the method should be used *after* the expert has visually analyzed, detected, and annotated friction ridge skin features for which the expert has reasonably high confidence of their presence. Second, the method should be used as a framework for flagging impressions which may require additional quality assurance review. Although the method demonstrates reasonable consistency with experts' judgements, it should not be considered a replacement for the experts' interpretation. This method is a step toward greater transparency and objectivity, but is not designed or intended to supplant the careful interpretation of experts.

This method provides fingerprint experts the capability to provide an empirical foundation to support their subjective interpretations following *Analysis*. It also offers a framework for organizations to establish transparent, measurable, and demonstrable criteria for Value determinations and a means of flagging impressions that are vulnerable to erroneous outcomes or inconsistency between experts (e.g. higher Complexity and/or Difficulty). Finally, it provides a means for quantitatively summarizing the overall quality of the impression in terms of Value, Complexity, and Difficulty for ensuring representative distributions in samples used for research designs, proficiency testing, error rate testing, and other applications by forensic science stakeholders. As a stand-alone application, this method enables the forensic science community to take a step toward greater transparency and empiricism – particularly as it relates to Value and Complexity determinations during casework examinations and assessments of Difficulty for research, training, and testing purposes. Further, because this method provides quality assessments at both the local and global levels (LQS and GQS), its development lends the possibility of integrating with other quality assessment and statistical evaluation software applications, such as *FRStat* [50], to provide a complete tool-pack to ensure experts' interpretations are empirically supported for all major decisions throughout the entire examination methodology.

2.1.5 Conclusion

Over the years, the forensic science community has faced increasing criticism by scientific and legal commentators, challenging the validity and reliability of many forensic examination methods that rely on subjective interpretations by forensic practitioners. Among those concerns is the lack of an empirically demonstrable basis to assess the quality of fingerprint evidence for a given case.

In this paper, a method is presented which measures the clarity of friction ridge features and evaluates the quality of impression across three different scales: Value, Complexity, and Difficulty. The local quality scores (LQS) provide a quantitative assessment of the quality of individual features based on the clarity of the local region of friction ridge detail immediately surrounding each feature. Individual features are then color-coded green, yellow, or red indicating high, medium, or low quality. The results demonstrate remarkable separation between regions representing the extreme ends of "good" and "bad" quality of friction ridge detail and general agreement with experts' subjective assessments of feature quality based on features categorized as "high" or "medium" quality. While quality assessments at localized regions are important, quality assessments for the overall impression have the most significant impact on the ultimate outcome of the examination. The global quality scores (GQS) provide quantitative assessments of the quality of the entire impression against different outcome scales (value, complexity, and difficulty) based on the quality and quantity of individual features. The results demonstrate reasonable consistency between automated predictions and experts' subjective assessments. In an operational environment, the tool is intended to provide an empirical foundation to support experts' subjective judgments, provide transparency to the overall quality of the impression for a given outcome (e.g. determination of value, complexity, or difficulty), and provide a framework to establish policies and procedures for examination decisions geared toward flagging impressions that are generally lower quality and more vulnerable to disagreements between experts or potentially erroneous interpretations.

As with any method, there are limitations to consider. The most significant is that this method relies on the features annotated by the expert and does not take into account all aspects of the friction ridge detail. Consequently, the system should not be considered as a means of supplanting expert interpretation and judgement when analyzing friction ridge detail. Rather, the method should be considered a tool to support experts' judgements or detect potentially problematic impressions necessitating further quality assurance review.

Although various aspects of this method may be further optimized, the performance characteristics described are proposed as a sufficient basis to demonstrate the foundational validity of the method to perform within the scope of its intended purpose – as a means of providing a quantitative measure of the quality of a fingerprint. Further optimizations which may improve upon the method's performance are encouraged for future works.

2.2 Comparison with Other Methods

The preceding section presents the published manuscript [49] that describes the development and validation of DFIQI as a stand-alone software application. This section is supplemental to the published manuscript [49] and explores the performance of DFIQI compared to other available methods. The results show that the DFIQI provides comparable performance to other available methods and that no added value is obtained when machine learning (ML) techniques are leveraged to combine multiple solutions.

2.2.1 Background

The DFIQI algorithm presented in this chapter provides a measure of the overall quality of the impression for further examination purposes as well as an objective measure of the clarity of friction ridge features identified by the analyst. The DFIQI algorithm, however, is not the only algorithm that has been proposed for such purposes. Among several other methods that have been proposed, LFIQ [44], LQMetric [14], and ESLR [61] algorithms have been made accessible for evaluation. Each algorithm accounts for different variables and provides a distinct approach to the calculation of its quality measure. Taking into account that the DFIQI was developed with consideration of balancing performance against computational complexity and algorithmic transparency as it relates to the selection of variables and the machine learning classifier, it seems prudent to conduct an exploratory comparison of the DFIQI against these other methods to better understand the impact of this tradeoff. Rather than comparing the performance the DFIQI as a standalone application as described in [49] with each method individually, we can explore the value of a multidimensional combination of the various algorithmic approaches across several different machine learning classifiers. This combined approach would enable us to measure the extent to which the performance of a new integrated algorithm for which the input variables are defined by the outputs and related meta-data from each of the individual methods improves over the performance of the DFIQI alone for predicting the quality of impressions as it relates to value, complexity, and difficulty determinations across all of the methods. This evaluation is useful as it provides a more objective means of assessing the optimal approach, amongst these four distinct algorithms (DFIQI [49], LFIQ [44], LQMetric [14], and ESLR [61]) or a combination thereto through more sophisticated machine learning classifiers, for measuring the overall quality of the impressions for further examination purposes.

2.2.2 Materials & Methods

This exploratory comparison was conducted using a dataset provided by Eldridge et al. (2020) [57]. The full dataset consisted of a total of 3,241 determinations made by 116 analysts rendering "value," "complexity," and "difficulty" decisions for each image they viewed from a set of 100 different marks – each participant was provided a set of approximately 30 impressions to analyze, resulting in each impression being analyzed by between 26 and 41 different analysts. The impressions were obtained during the course of normal casework at a large metropolitan police laboratory using standard powder processing and lifting techniques. All participants were practicing friction ridge examiners recruited by several outreach methods, such as email distribution lists, presentations given at professional educational meetings, and professional contacts. As described in Swofford et al. [49], the DFIQI was developed and tested using a random 50/50 training-test split obtained from this dataset as it relates to individual examiners' observations and judgments. For purposes of this evaluation, however, additional models were developed and tested using cross validation schemes described below.

The baseline performance of the DFIQI to which the performance of the other methods is compared against was established in two distinct ways: First, the baseline performance of the actual DFIQI algorithm (as it is originally proposed in [49]) is provided by the raw model diagnostics discussed in Appendix B-1. This is based on a multinomial regression model using

two variables (LQS_{sum} and *n*FEAT) and a random 50/50 training-test split using individual examiners' observations and judgments of "value," "complexity," and "difficulty" from the full dataset provided by Eldridge et al. [57]. For purposes of this exploratory comparison, these data are referred to as the baseline performance of the "*actual* DFIQI" algorithm. Although these data provide the raw model diagnostics of the actual DFIQI algorithm as it was originally proposed in [49], it does not provide a straightforward means of exploring the utility of a multidimensional combination of the various algorithmic approaches across several different machine learning classifiers and measuring the extent to which the performance of a new integrated algorithm (for which the input variables are defined by the outputs and related meta-data from each of the individual methods – LFIQ [44], LQMetric [14], and ESLR [61]) improves over the performance of the DFIQI variables alone. This is because all of the other methods available for this exploratory comparison (i.e., LFIQ [44], LQMetric [14], and ESLR [61]) do not rely on user inputs and therefore only have a single set of quality measures output per image (independent of user inputs). This contrasts with the DFIQI in that, as noted above, the DFIQI is dependent on users to identify the features in the impression. Variations between examiners in the specific quantities and locations of features identified in the impression will result in different quality measures output for the same image. Thus, for purposes of this evaluation, we propose a second way of establishing the baseline performance of the DFIQI and permit a more straightforward comparison by creating a *new* derivative algorithm trained using the *mean* values of the variables and related meta-data from the DFIQI across all examiners for each image. This would create a single set of consensus derived values as it relates to the DFIQI variables alone for each image independent of the input from a specific user. Differences between examiners in the outcome judgment (i.e., determinations of "value," "complexity," and "difficulty") are assigned (for purposes of the training) based on majority consensus among all examiners for each image. For purposes of this exploratory comparison, model diagnostics related to this approach are referred to as the baseline performance of the "*derivative* DFIQI" algorithm. Having a single set of input values representing the mean values of the DFIQI variables then enable these data to be combined with the variables from the other algorithmic approaches (i.e., LFIQ [44], LQMetric [14], and ESLR [61]) to create the new integrated algorithm for which we can measure the extent to which performance is improved compared to a model using the DFIQI variables alone.

For purposes of this evaluation, three sets of models were developed (for each type of judgment – "value," "complexity," and "difficulty") using the applicable variables and related meta-data from the DFIQI, LFIQ, LQMetric, and ESLR. One set of models established the baseline performance using the DFIQI variables alone (referred to as "*derivative* DFIQI" baseline). Variables and related meta-data used in this set of models included: SP3G, Bimodal Separation, Acutance, Mean Object Width, Spatial Frequency, LQS_{sum}, *n*FEAT, and GQS (as applicable for each judgment outcome, i.e., $Value_{GQS}$, $Complexity_{GQS}$, $Difficulty_{GQS}$). A second set of models established the baseline performance of all other methods combined (LFIQ, LQMetric, and ESLR) (referred to as "LFIQ-LQM-ESLR" baseline). Variables and related meta-data used in this set of models included: LFIQ-1, LQM-overall-quality, LQM-overall-clarity, LQM-VCMP, LQM-VID, LQM-area-of-ridge-flow, LQM-area-of-good-ridge-flow, LQM-largest-contiguous-area-of-ridge-flow, LQM-largest-contiguous-area-of-good-ridge-flow, LQM-FA, and ESLR. A third set of models (i.e., integrated algorithm) provided a measure of the extent to which the performance is improved by combining the variables from all algorithmic approaches (DFIQI, LFIQ, LQMetric, and ESLR) compared to the performance provided by the DFIQI alone (referred to as "Integrated

Algorithm – All Methods Combined"). Variables and related meta-data used in this set of models included those which were used in the prior two sets of models combined. All models were developed using the *caret* package in R [62] using a range of machine learning techniques (naïve based classifier, tree-based classifiers, discriminant analysis techniques, neural networks and support vector machines) called directly from the *caret* package, specifically: Classification and Regression Tree (CART), Random Forest (RF), Multinomial (MultiNom), K-Nearest Neighbors (KNN), Naïve Bayes (NB), C5.0, and Support Vector Machine (SVM). The machine learning techniques were applied using a "leave-one-out" cross validation for which training was conducted using the mean values for the input variables across all examiners for 99 of the 100 images, then tested using the individual examiners' responses for the image left out. This process was repeated 100 times such that with each iteration a different image was left out. For purposes of the training, the judgment outcomes were assigned based on majority consensus of examiners' responses for each image. This approach creates models that are intended to predict the judgment outcome that is most likely supported by a consensus of examiners. Testing, however, was done in two distinct ways: (1) the values for the input variables were based on the examiners' individual responses and the judgment outcomes assigned were based on the majority consensus of examiners' responses for each image, and (2) the values for the input variables *and* judgment outcomes assigned were both based on the examiners' individual responses (noting, however, that values for the input variables for LFIQ, LQM, and ESLR were the same across all examiners since those methods do not require examiner input). The first approach evaluates how well the models predicted the consensus judgment outcomes. The second approach evaluates how well the models predicted the individual examiners' judgment outcomes, which mirrors how the "*actual* DFIQI" algorithm was tested.

2.2.3 Results & Discussion

The performance of each method was evaluated in terms of classification accuracy[6] based on the raw model diagnostics provided by the *nnet* package in R [58]. Although these raw model diagnostics do not necessarily correspond to the performance of the GQS scores output by the *actual* DFIQI when applied in the context of a binary decision to flag an impression for further quality assurance review (consistent with how the DFIQI is intended to be applied in practice), they are useful for purposes of this evaluation as they provide a consistent means of measuring the extent to which the performance of the various models can be expected to change when different variables and machine learning techniques are used.

The baseline performance of the "*actual* DFIQI" for each judgment outcome is provided by Appendix B-1 and reproduced below in Table 2-9. These data provide the nexus between the actual performance of the DFIQI, in terms of the classification accuracy based on the raw model diagnostics as originally proposed by [49], and the new models developed for purposes of this exploratory comparison.

[6] Ground truth is non-existent for determinations of "value," "complexity," and "difficulty." Although the term "precision" is technically more appropriate, the term "accuracy" in this context refers to the extent to which the output classification from the model corresponds to examiners' determinations and is used in this section to be consistent with the discussion in the supplemental appendix of the published manuscript reflected in this chapter.

Outcome Judgment	Overall Accuracy	95% Confidence Interval (lower bound – upper bound)
Value	0.805	(0.785 - 0.824)
Complexity	0.674	(0.650 - 0.696
Difficulty	0.580	(0.555 - 0.604)

Table 2-9: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "actual DFIQI" for each judgment class (Value, Complexity, Difficulty). Table reproduced from [49].

"Value" Determinations

The overall classification accuracy of the "*derivative* DFIQI" baseline set of models is provided in Tables 2-10a and 2-10b for each of the six machine learning techniques used for this evaluation. These data provide the baseline performance of models trained using the mean values of the DFIQI variables and assignment of the outcome judgment for each image based on majority consensus among all examiners. Table 2-10a provides the baseline performance of the models when tested using the examiners' individual values of the DFIQI variables and outcome judgments assigned based on majority consensus among all examiners for each image. Table 2-10b provides the baseline performance of the models when tested using the examiners' individual values of the DFIQI variables and outcome judgments assigned based on the examiners' individual responses.

Value Determinations (*Derivative* DFIQI Baseline) *Consensus Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.816	0.795 – 0.836
RF	0.823	0.804 – 0.843
MultiNom	0.833	0.814 – 0.852
KNN	0.828	0.809 – 0.847
NB	0.805	0.785 – 0.825
C5.0	0.838	0.820 – 0.857
SVM	0.836	0.818 – 0.853

Table 2-10a: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "derivative DFIQI baseline" for each machine learning technique for Value determinations (tested using outcome judgments assigned based on majority consensus among all examiners for each image).

Value Determinations (*Derivative* DFIQI Baseline) *Individual Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.809	0.787 – 0.832
RF	0.812	0.790 – 0.834
MultiNom	0.809	0.787 – 0.831
KNN	0.809	0.786 – 0.831
NB	0.789	0.769 – 0.810
C5.0	0.811	0.788 – 0.833
SVM	0.798	0.774 – 0.822

Table 2-10b: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "derivative DFIQI baseline" for each machine learning technique for Value determinations (tested using outcome judgments assigned based on the examiners' individual responses).

The overall classification accuracy of the "LFIQ-LQM-ESLR" baseline set of models is provided in Tables 2-11a and 2-11b for each of the six machine learning techniques used for this evaluation. These data provide the baseline performance of models trained using the LFIQ, LQM, and ESLR variables and assignment of the outcome judgment for each image based on majority consensus among all examiners. Table 2-11a provides the baseline performance of the models when tested using the values of the LFIQ, LQM, and ESLR variables and outcome judgments assigned based on majority consensus among all examiners for each image. Table 2-11b provides the baseline performance of the models when tested using the values of the LFIQ, LQM, and ESLR variables and outcome judgments assigned based on the examiners' individual responses.

Value Determinations (LFIQ-LQM-ESLR Baseline) *Consensus Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.849	0.841 – 0.858
RF	0.835	0.825 – 0.845
MultiNom	0.754	0.739 – 0.768
KNN	0.823	0.810 – 0.835
NB	0.556	0.532 – 0.579
C5.0	0.698	0.676 – 0.720
SVM	0.825	0.816 – 0.835

Table 2-11a: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "LFIQ-LQM-ESLR baseline" for each machine learning technique for Value determinations (tested using outcome judgments assigned based on majority consensus among all examiners for each image).

Value Determinations (LFIQ-LQM-ESLR Baseline) *Individual Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.724	0.695 – 0.753
RF	0.704	0.676 – 0.732
MultiNom	0.674	0.649 – 0.698
KNN	0.701	0.671 – 0.732
NB	0.548	0.524 – 0.573
C5.0	0.601	0.578 – 0.624
SVM	0.710	0.679 – 0.740

Table 2-11b: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "LFIQ-LQM-ESLR baseline" for each machine learning technique for Value determinations (tested using outcome judgments assigned based on the examiners' individual responses).

The overall classification accuracy of the "Integrated Algorithm – All Methods Combined" set of models is provided in Tables 2-12a and 2-12b for each of the six machine learning techniques used for this evaluation. These data provide the performance of all methods combined – models trained using the mean values of the DFIQI variables, combined with the values of the LFIQ, LQM, and ESLR variables, and assignment of the outcome judgment for each image based on majority consensus among all examiners. Table 2-12a provides the performance of the models when tested using the examiners' individual values of the DFIQI variables, combined with the values of the LFIQ, LQM, and ESLR variables, and outcome judgments assigned based on majority consensus among all examiners for each image. Table 2-12b provides the baseline performance of the models when tested using the examiners' individual values of the DFIQI variables, combined with the values of the LFIQ, LQM, and ESLR variables, and outcome judgments assigned based on the examiners' individual responses. This set of models provide the extent to which performance is improved when combining variables from all methods compared to a model using the DFIQI variables alone.

Value Determinations (Integrated Algorithm – All Methods Combined) *Consensus Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.816	0.795 – 0.836
RF	0.832	0.812 – 0.851
MultiNom	0.764	0.745 – 0.782
KNN	0.833	0.821 – 0.845
NB	0.747	0.728 – 0.767
C5.0	0.837	0.818 – 0.856
SVM	0.837	0.822 – 0.852

Table 2-12a: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "Integrated Algorithm – All Methods Combined" for each machine learning technique for Value determinations (tested using outcome judgments assigned based on majority consensus among all examiners for each image).

Value Determinations (Integrated Algorithm – All Methods Combined) *Individual Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.809	0.787 – 0.832
RF	0.810	0.788 – 0.832
MultiNom	0.775	0.754 – 0.796
KNN	0.710	0.680 – 0.741
NB	0.737	0.717 – 0.758
C5.0	0.810	0.788 – 0.832
SVM	0.798	0.775 – 0.821

Table 2-12b: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "Integrated Algorithm – All Methods Combined" for each machine learning technique for Value determinations (tested using outcome judgments assigned based on the examiners' individual responses).

"Complexity" Determinations

The overall classification accuracy of the "*derivative* DFIQI" baseline set of models is provided in Tables 2-13a and 2-13b for each of the six machine learning techniques used for this evaluation. These data provide the baseline performance of models trained using the mean values of the DFIQI variables and assignment of the outcome judgment for each image based on majority consensus among all examiners. Table 2-13a provides the baseline performance of the models when tested using the examiners' individual values of the DFIQI variables and outcome judgments assigned based on majority consensus among all examiners for each image. Table 2-13b provides the baseline performance of the models when tested using the examiners' individual values of the DFIQI variables and outcome judgments assigned based on the examiners' individual responses.

Complexity Determinations (*Derivative* DFIQI Baseline) *Consensus Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.674	0.653 – 0.695
RF	0.668	0.647 – 0.688
MultiNom	0.667	0.646 – 0.688
KNN	0.635	0.613 – 0.657
NB	0.645	0.622 – 0.667
C5.0	0.654	0.632 – 0.676
SVM	0.655	0.632 – 0.678

Table 2-13a: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "derivative DFIQI baseline" for each machine learning technique for Complexity determinations (tested using outcome judgments assigned based on majority consensus among all examiners for each image).

Complexity Determinations (*Derivative* DFIQI Baseline) *Individual Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.682	0.658 – 0.706
RF	0.677	0.654 – 0.700
MultiNom	0.670	0.647 – 0.693
KNN	0.666	0.643 – 0.688
NB	0.660	0.636 – 0.685
C5.0	0.668	0.645 – 0.692
SVM	0.663	0.639 – 0.687

Table 2-13b: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "derivative DFIQI baseline" for each machine learning technique for Complexity determinations (tested using outcome judgments assigned based on the examiners' individual responses).

The overall classification accuracy of the "LFIQ-LQM-ESLR" baseline set of models is provided in Tables 2-14a and 2-14b for each of the six machine learning techniques used for this evaluation. These data provide the baseline performance of models trained using the LFIQ, LQM, and ESLR variables and assignment of the outcome judgment for each image based on majority consensus among all examiners. Table 2-14a provides the baseline performance of the models when tested using the values of the LFIQ, LQM, and ESLR variables and outcome judgments assigned based on majority consensus among all examiners for each image. Table 2-14b provides the baseline performance of the models when tested using the values of the LFIQ, LQM, and ESLR variables and outcome judgments assigned based on the examiners' individual responses.

Complexity Determinations (LFIQ-LQM-ESLR Baseline) *Consensus Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.590	0.576 – 0.604
RF	0.578	0.562 – 0.593
MultiNom	0.610	0.593 – 0.626
KNN	0.658	0.644 – 0.672
NB	0.600	0.585 – 0.616
C5.0	0.637	0.622 – 0.651
SVM	0.623	0.606 – 0.639

Table 2-14a: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "LFIQ-LQM-ESLR baseline" for each machine learning technique for Complexity determinations (tested using outcome judgments assigned based on majority consensus among all examiners for each image).

Complexity Determinations (LFIQ-LQM-ESLR Baseline) *Individual Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.554	0.530 – 0.578
RF	0.554	0.532 – 0.575
MultiNom	0.558	0.536 – 0.580
KNN	0.542	0.519 – 0.565
NB	0.514	0.492 – 0.536
C5.0	0.562	0.540 – 0.584
SVM	0.555	0.527 – 0.583

Table 2-14b: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "LFIQ-LQM-ESLR baseline" for each machine learning technique for Complexity determinations (tested using outcome judgments assigned based on the examiners' individual responses).

The overall classification accuracy of the "Integrated Algorithm – All Methods Combined" set of models is provided in Tables 2-15a and 2-15b for each of the six machine learning techniques used for this evaluation. These data provide the performance of all methods combined – models trained using the mean values of the DFIQI variables, combined with the values of the LFIQ, LQM, and ESLR variables, and assignment of the outcome judgment for each image based on majority consensus among all examiners. Table 2-15a provides the performance of the models when tested using the examiners' individual values of the DFIQI variables, combined with the values of the LFIQ, LQM, and ESLR variables, and outcome judgments assigned based on majority consensus among all examiners for each image. Table 2-15b provides the baseline performance of the models when tested using the examiners' individual values of the DFIQI variables, combined with the values of the LFIQ, LQM, and ESLR variables, and outcome judgments assigned based on the examiners' individual responses. This set of models provide the extent to which performance is improved when combining variables from all methods compared to a model using the DFIQI variables alone.

Complexity Determinations (Integrated Algorithm – All Methods Combined) *Consensus Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.674	0.653 – 0.695
RF	0.679	0.657 – 0.701
MultiNom	0.636	0.626 – 0.656
KNN	0.628	0.616 – 0.640
NB	0.646	0.627 – 0.664
C5.0	0.677	0.656 – 0.697
SVM	0.670	0.651 – 0.690

Table 2-15a: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "Integrated Algorithm – All Methods Combined" for each machine learning technique for Complexity determinations (tested using outcome judgments assigned based on majority consensus among all examiners for each image).

Complexity Determinations (Integrated Algorithm – All Methods Combined) *Individual Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.682	0.658 – 0.706
RF	0.681	0.658 – 0.703
MultiNom	0.653	0.631 – 0.675
KNN	0.539	0.516 – 0.563
NB	0.632	0.611 – 0.653
C5.0	0.676	0.653 – 0.699
SVM	0.653	0.631 – 0.676

Table 2-15b: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "Integrated Algorithm – All Methods Combined" for each machine learning technique for Complexity determinations (tested using outcome judgments assigned based on the examiners' individual responses).

"Difficulty" Determinations

The overall classification accuracy of the "*derivative* DFIQI" baseline set of models is provided in Tables 2-16a and 2-16b for each of the six machine learning techniques used for this evaluation. These data provide the baseline performance of models trained using the mean values of the DFIQI variables and assignment of the outcome judgment for each image based on majority consensus among all examiners. Table 2-16a provides the baseline performance of the models when tested using the examiners' individual values of the DFIQI variables and outcome judgments assigned based on majority consensus among all examiners for each image. Table 2-16b provides the baseline performance of the models when tested using the examiners' individual values of the DFIQI variables and outcome judgments assigned based on the examiners' individual responses.

Difficulty Determinations (*Derivative* DFIQI Baseline) *Consensus Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.540	0.514 – 0.567
RF	0.594	0.571 – 0.617
MultiNom	0.590	0.565 – 0.614
KNN	0.563	0.539 – 0.587
NB	0.598	0.575 – 0.621
C5.0	0.550	0.528 – 0.572
SVM	0.584	0.559 – 0.609

Table 2-16a: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "derivative DFIQI baseline" for each machine learning technique for Difficulty determinations (tested using outcome judgments assigned based on majority consensus among all examiners for each image).

Difficulty Determinations (*Derivative* DFIQI Baseline) *Individual Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.532	0.502 – 0.562
RF	0.552	0.522 – 0.582
MultiNom	0.552	0.524 – 0.580
KNN	0.525	0.493 – 0.556
NB	0.554	0.523 – 0.584
C5.0	0.530	0.501 – 0.559
SVM	0.552	0.523 – 0.582

Table 2-16b: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "derivative DFIQI baseline" for each machine learning technique for Difficulty determinations (tested using outcome judgments assigned based on the examiners' individual responses).

The overall classification accuracy of the "LFIQ-LQM-ESLR" baseline set of models is provided in Tables 2-17a and 2-17b for each of the six machine learning techniques used for this evaluation. These data provide the baseline performance of models trained using the LFIQ, LQM, and ESLR variables and assignment of the outcome judgment for each image based on majority consensus among all examiners. Table 2-17a provides the baseline performance of the models when tested using the values of the LFIQ, LQM, and ESLR variables and outcome judgments assigned based on majority consensus among all examiners for each image. Table 2-17b provides the baseline performance of the models when tested using the values of the LFIQ, LQM, and ESLR variables and outcome judgments assigned based on the examiners' individual responses.

Difficulty Determinations (LFIQ-LQM-ESLR Baseline) *Consensus Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.575	0.560 – 0.590
RF	0.528	0.514 – 0.543
MultiNom	0.600	0.580 – 0.619
KNN	0.485	0.470 – 0.502
NB	0.628	0.615 – 0.642
C5.0	0.530	0.516 – 0.545
SVM	0.559	0.545 – 0.573

Table 2-17a: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "LFIQ-LQM-ESLR baseline" for each machine learning technique for Difficulty determinations (tested using outcome judgments assigned based on majority consensus among all examiners for each image).

Difficulty Determinations (LFIQ-LQM-ESLR Baseline) *Individual Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.503	0.482 – 0.525
RF	0.485	0.464 – 0.505
MultiNom	0.517	0.495 – 0.540
KNN	0.483	0.459 – 0.507
NB	0.511	0.489 – 0.533
C5.0	0.466	0.446 – 0.485
SVM	0.495	0.475 – 0.514

Table 2-17b: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "LFIQ-LQM-ESLR baseline" for each machine learning technique for Difficulty determinations (tested using outcome judgments assigned based on the examiners' individual responses).

The overall classification accuracy of the "Integrated Algorithm – All Methods Combined" set of models is provided in Tables 2-18a and 2-18b for each of the six machine learning techniques used for this evaluation. These data provide the performance of all methods combined – models trained using the mean values of the DFIQI variables, combined with the values of the LFIQ, LQM, and ESLR variables, and assignment of the outcome judgment for each image based on majority consensus among all examiners. Table 2-18a provides the performance of the models when tested using the examiners' individual values of the DFIQI variables, combined with the values of the LFIQ, LQM, and ESLR variables, and outcome judgments assigned based on majority consensus among all examiners for each image. Table 2-18b provides the baseline performance of the models when tested using the examiners' individual values of the DFIQI variables, combined with the values of the LFIQ, LQM, and ESLR variables, and outcome judgments assigned based on the examiners' individual responses. This set of models provide the

extent to which performance is improved when combining variables from all methods compared to a model using the DFIQI variables alone.

Difficulty Determinations (Integrated Algorithm – All Methods Combined) *Consensus Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.516	0.489 – 0.543
RF	0.607	0.584 – 0.630
MultiNom	0.595	0.572 – 0.618
KNN	0.551	0.536 – 0.566
NB	0.613	0.593 – 0.632
C5.0	0.586	0.563 – 0.608
SVM	0.572	0.551 – 0.592

Table 2-18a: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "Integrated Algorithm – All Methods Combined" for each machine learning technique for Difficulty determinations (tested using outcome judgments assigned based on majority consensus among all examiners for each image).

Difficulty Determinations (Integrated Algorithm – All Methods Combined) *Individual Outcome Judgments*		
Machine Learning Technique	**Overall Accuracy**	**95% Confidence Interval (lower bound – upper bound)**
CART	0.506	0.475 – 0.537
RF	0.562	0.533 – 0.591
MultiNom	0.538	0.509 – 0.566
KNN	0.485	0.462 – 0.509
NB	0.557	0.531 – 0.583
C5.0	0.552	0.525 – 0.579
SVM	0.537	0.512 – 0.562

Table 2-18b: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications of the "Integrated Algorithm – All Methods Combined" for each machine learning technique for Difficulty determinations (tested using outcome judgments assigned based on the examiners' individual responses).

General Discussion

There are three key observations we can make based on these data presented in Tables 2-9 through 2-18b. First, the baseline performance of the "*derivative* DFIQI" (for both individual outcome judgments and consensus outcome judgments) is consistent with the performance of the "*actual* DFIQI" as provided by Appendix B-1 across *all* outcome judgments ("value," "complexity," and "difficulty"). This is relevant as it demonstrates that the modification used in the training of the "*derivative* DFIQI" to permit these comparisons did not have a substantive impact to the performance of the baseline. Second, in terms of the best performing machine

learning techniques, the "derivative DFIQI" (for both individual outcome judgments and consensus outcome judgments) performs similar to the "LFIQ-LQM-ESLR Baseline" for consensus outcome judgments only. The performance the "LFIQ-LQM-ESLR Baseline" for individual outcome judgments, however, was consistently inferior. These findings were observed across all outcome judgments ("value," "complexity," and "difficulty"). Third, the performance of the "Integrated Algorithm – All Methods Combined" did not result in substantive improvements compared to the baseline performance provided by the models developed using the DFIQI variables alone. These results suggest that, for purposes of predicting examiners' judgments of "value," "complexity," and "difficulty," the DFIQI variables are influential variables for predicting individual and consensus outcome judgments and the tradeoff decisions that were made during the development of DFIQI (i.e., reduced computational complexity and algorithmic transparency) did not have a substantive impact to performance compared to the more complex algorithms evaluated. Although these findings do not allow strong inferences as to *why* the models developed using the DFIQI variables resulted in the greatest performance, one key distinction between the DFIQI and the other methods evaluated is that the values of the DFIQI variables rely on examiners' inputs (as it relates to identifying the specific quantities and locations of features in the impression) whereas the other methods do not. Thus, in this context of predicting examiners' outcome judgments for which ground truth is non-existent, it seems that a tradeoff must be made in terms of performance versus complete objectivity. Looking forward, however, objectivity of the DFIQI might be improved by relying on high performance automated feature extraction algorithms to automatically detect the quantity and locations of features as a pre-processing step prior to using the DFIQI (thereby reducing the variability caused by user inputs).

3 Statistical Interpretation Software (FRStat)

This chapter presents a manuscript entitled "A Method for the Statistical Interpretation of Friction Ridge Skin Impression Evidence: Method Development and Validation" (Swofford et al., 2018) [50] published in *Forensic Science International* that describes the development and validation of a publicly accessible algorithm and software application (referred to as the Friction Ridge Statistical Interpretation Software, or FRStat). The FRStat algorithm first calculates the similarity (referred to as the Global Similarity Statistic, or GSS) between two sets of features identified by an analyst on two separate impressions which the analyst believes to correspond. The software then provides two estimates, one indicating how often impressions originating from common sources would result in a GSS that is equal to or less than the calculated GSS and another indicating how often impressions from different sources would result in a GSS that is equal to or greater than the calculated GSS. The two values are then combined as a ratio providing a single summary statistic indicating to what extent the GSS is consistent with impressions originating from a common source compared to different sources. In addition to the published manuscript, this chapter also discusses the performance of FRStat compared to another available methods.

3.1 Method Development and Validation

A Method for the Statistical Interpretation of Friction Ridge Skin Impression Evidence: Method Development and Validation

[1]Swofford, H.J.; [1]Koertner, A.J.; [2]Zemp, F.; [3]Ausdemore, M.; [4]Liu, A.; [1]Salyards M.J.

[1]U.S. Army Criminal Investigation Laboratory, Defense Forensic Science Center, USA
[2]School of Criminal Justice, Forensic Science Institute, University of Lausanne, Switzerland
[3]Department of Mathematics and Statistics, University of South Dakota, USA
[4]Department of Statistics, University of Virginia, USA

3.1.1 Abstract

The forensic fingerprint community has faced increasing amounts of criticism by scientific and legal commentators, challenging the validity and reliability of fingerprint evidence due to the lack of an empirically demonstrable basis to evaluate and report the strength of the evidence in a given case. This paper presents a method, developed as a stand-alone software application, *FRStat*, which provides a statistical assessment of the strength of fingerprint evidence. The performance was evaluated using a variety of mated and non-mated datasets. The results show strong performance characteristics, often with values supporting specificity rates greater than 99%. This method provides fingerprint experts the capability to demonstrate the validity and reliability of fingerprint evidence in a given case and report the findings in a more transparent and standardized fashion with clearly defined criteria for conclusions and known error rate information thereby responding to concerns raised by the scientific and legal communities.

Keywords: Forensic Science; Fingerprints; Strength of evidence; Weight of Evidence; Likelihood Ratio; Probability

3.1.2 Introduction

Over the last several years, the forensic science community has faced increasing amounts of criticism by scientific and legal commentators, challenging the validity and reliability of many forensic examination methods that rely on subjective interpretations by forensic practitioners [1-7]. Of particular concern, noted in 2009 by the National Research Council (NRC) of the National Academies of Science (NAS) [3] as well as the President's Council of Advisors on Science and Technology (PCAST) as recently as September 2016 [7], is the lack of an empirically demonstrable basis to substantiate conclusions from pattern evidence, thus limiting the ability for the judiciary to reasonably understand the reliability of the expert's testimony for the given case. Consistent with several academic commentators, both the NRC and PCAST strongly encouraged the forensic science community to develop tools to evaluate and report the strength of forensic evidence using validated statistical methods [3, 7, 9]. While these concerns apply to nearly every pattern evidence discipline, the forensic fingerprint discipline has received most of the attention because fingerprint analysis is one of the most widely used techniques in the criminal justice system. As a result, over the last several years numerous methods and models have been proposed to provide a statistical estimate of the weight of fingerprint evidence using features that are familiar to forensic practitioners, primarily fingerprint minutiae [29-43].

Prior methods can be classified as either (a) feature-based models, which calculate probability estimates from the random correspondence of feature configurations within a pre-determined tolerance or (b) similarity metric models, which calculate the probability estimates from distributions of similarity scores. Among the feature-based models: Zhu, Dass, and Jain proposed a family of finite mixture models to represent the distribution of fingerprint minutiae, including minutiae clustering tendencies and dependencies in different regions of the fingerprint image domain to calculate the probability of a random correspondence [30]; Su and Srihari proposed a model based on the spatial distribution of fingerprint minutiae, taking into account the dependency of each minutiae on nearby minutiae and the confidence of their presence in the evidence, to calculate the probability of random correspondence [34]; Lim and Dass proposed a simulation model based on the distribution of fingerprint minutiae estimated using a Bayesian MCMC framework [35]; Abraham et al. proposed a model based on support vector machines trained with features discovered via morphometric and spatial analyses of corresponding minutiae configurations for both match and close non-match populations [39]. Among the similarity metric models: Neumann et al. proposed a variety of models based on a similarity metric calculated from feature vectors taking into consideration type, direction, and relative spatial relationships of fingerprint minutiae [29, 32, 37] as well as taking into account general pattern [38]; Egli [31, 33, 41], Choi and Nagar [36], and Leegwater et al. [43] proposed a variety of models based on the distribution of similarity scores from Automated Fingerprint Identification Systems (AFIS). Abraham, deJongh, and Rodriguez evaluate the effect of different types of conditioning on the impact of the results derived from AFIS-based models [40]. Taking a slightly different approach than those discussed above, Neumann et al. proposed a model relying on an AFIS algorithm to estimate the probability distributions of spatial relationships, directions and types of minutiae rather than directly modeling the distribution of AFIS scores [42].

Although each of the proposed models demonstrated promising performance metrics, none have been widely accessible to the forensic community, thus prohibiting their ability to be further evaluated or implemented into routine casework. Consequently, forensic science laboratories throughout the United States have been unable to adequately address the concerns by the NRC and PCAST by demonstrating the reliability of fingerprint evidence *for the case at hand*. In light of this gap, this paper presents a method, developed as a stand-alone software application, *FRStat*, which measures the similarity between two configurations of friction ridge skin features and calculates a similarity metric. Statistical modeling of the distributions of the similarity statistic values from mated and non-mated impressions facilitates a statistical assessment of the strength of the fingerprint evidence. Although this method builds upon the general concepts of similarity-based models described earlier, this method utilizes a novel approach for quantifying the similarity and strength of fingerprint evidence. Further, having been developed as a stand-alone software application by the United States Government, this method is accessible to the forensic community thereby providing the capability to ensure the strength of fingerprint evidence is evaluated with an empirically grounded basis.

This paper provides a brief overview of the similarity calculations performed by the method followed by more detailed discussions regarding its development, performance and validation. Limitations of the method and considerations for policy and procedure when applied to forensic casework are also discussed.

3.1.3 Materials & Methods

Similarity Calculations

In general terms, the method measures the similarity between the configurations of friction ridge skin features (often referred to as level 2 detail or minutiae) from two different fingerprint images. The spatial relationships and angles of the features annotated by a forensic examiner are used to calculate a similarity statistic (i.e. score). The similarity statistic is then evaluated against datasets of similarity statistic values derived from pairs of impressions relevant for forensic casework made by mated (same) and non-mated (different) sources of friction ridge skin to calculate a statistical estimate of the strength of the given comparison. The method consists of three overarching steps: (1) feature pairing, (2) feature measurements, and (3) similarity statistic calculations.

In order to perform the similarity calculations, the features must be paired between the two impressions. Features are paired by initially detecting the Cartesian coordinates and angles of the annotated features on each image, which represent the locations and angles of ridge flow for the features. Using those feature details, a series of transformations are performed by iteratively rotating and translating the feature configurations to identify the optimal overlay of features between the two impressions among all possible overlays. Corresponding features are paired between the two images using a well-established combinatorial optimization algorithm to solve for the "optimal assignment" of features within each configuration [63]. Figure 3-1 illustrates the overlay and pairing of features. Once paired, the features retain their original Cartesian coordinates and angles as they appear on their respective images.

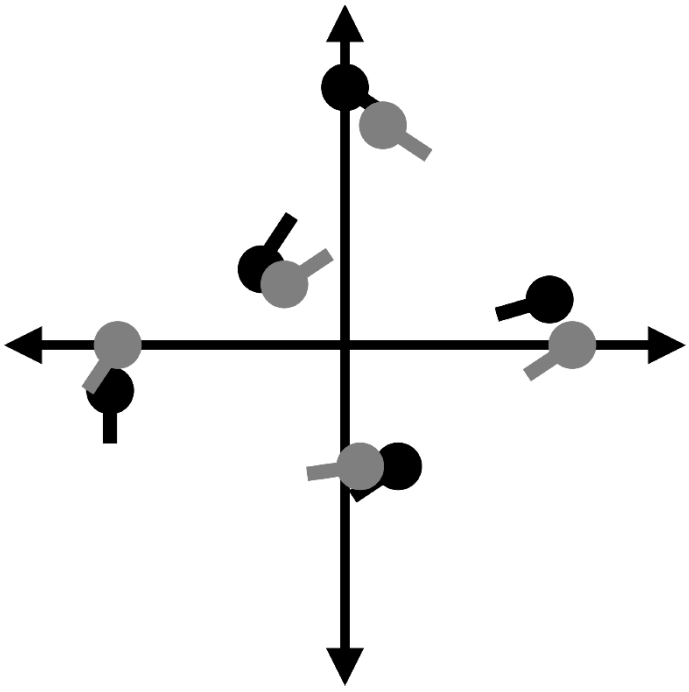


Figure 3-1. Conceptual illustration of the overlay and pairing of features. The grey annotations represent features on one impression and the black annotations represent features on the other.

Feature measurements are performed by applying a series of translation and rotation transformations to the paired features to facilitate anchoring and overlay of feature triplets (sub-configurations of three features). Within the feature triplet, two features serve as primary and secondary anchors while the third feature is measured with respect to the Euclidean distance and angle differences between the paired features. The primary anchor features are aligned on the origin of a coordinate plane and the secondary anchor features are aligned parallel to the x-axis. Figure 3-2 illustrates this concept of anchoring and overlaying a feature triplet. Using the measured differences between paired features, a "weight" is calculated for both the distance difference and angle difference between each feature. This process is repeated such that weights for distance and angle differences are calculated for all features using every possible combination of features in each triplet.

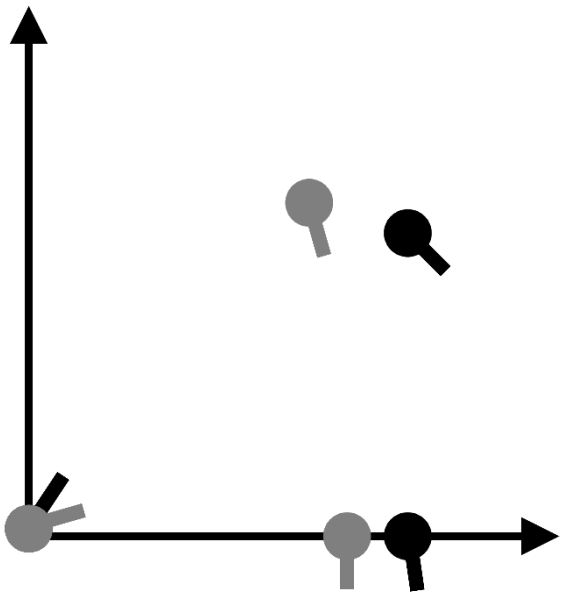


Figure 3-2. Conceptual illustration of the anchoring and overlay of a feature triplet. The primary pair of anchor features are on the origin. The secondary pair of anchor points are parallel to the x-axis. The grey annotations represent features on one impression and the black annotations represent features on the other.

The weight functions exploit subtle variations in the measured differences as well as provide context to the significance of those measurements in terms of the plasticity of friction ridge skin. The weight functions were designed such that the following criteria were met:

a. The weight functions are insensitive to common variations of feature location and angle displacements in mated source impressions due to distortion during friction ridge skin deposition under heavy pressure and movement.

b. The weight functions maximize the separation of similarity statistic values between mated and non-mated impressions for a given quantity of features.

c. The weight functions increase the separation of similarity statistic values between mated and non-mated impressions as the number of features increases.

The rules and parameter values for the weight functions are based on the empirical observations by Fagert & Morris [64]. In their study, Fagert & Morris [64] measured the variations of features commonly observed from repeated impressions of mated source fingers under various conditions of lateral pressure with respect to the distance difference and angle differences of features. Using the observations by Fagert & Morris [64] as an initial starting point, manual optimizations of the rules and parameter values for the weight functions were performed using a subset of mated fingerprint samples representing actual casework conditions. Once the measurements and weights for each feature are calculated they are combined into a single statistic

and transformed to represent the global similarity of the entire configuration of features (once transformed, higher values indicate higher similarity).

As noted above, the similarity statistic is dependent upon the manual selection and annotation of features by fingerprint experts. Consequently, the precision by which features are annotated introduces uncertainty in the calculated value of the similarity statistic. The method accounts for this uncertainty by applying an iterative random sampling scheme for the annotated details resulting in random displacements of the feature annotations in terms of Euclidean distance and angles. The parameters for the random displacements of feature annotations were determined by modeling the variability of feature annotations in latent impressions and reference impressions across multiple practicing fingerprint experts employed by a federal crime laboratory in the United States. Appendix C-1 provides more specific details regarding the evaluation and statistical modeling of the precision of feature annotations by practicing experts.

Following one-hundred iterations of randomly displacing feature annotations and re-calculating the global similarity statistic (using an unseeded random number generator), the final similarity statistic value output to the user is calculated as the lower bound of the 99% confidence interval for the mean. The lower bound of the 99% confidence interval was selected as it provides a conservative estimate of the "true" similarity statistic value for the given annotation.

Empirical Distributions

The empirical distributions of similarity statistic values among mated and non-mated impressions provide the foundation for estimating the strength of the fingerprint evidence. Taking into consideration that this method is intended for use in criminal or civil courts, the empirical distributions are intentionally biased such that the non-mated data are biased to *higher* similarity statistic values and mated data are biased to *lower* similarity statistic values. For non-mated data, this is accomplished by conditioning on (i) the region of friction ridge skin which maximizes the opportunities of observing higher values and (ii) any set of n features determined to be "optimally paired" from a larger set of m possible features with respect to the combinatorial optimization algorithm described in [63] under any condition of rotation and translation such that the similarity statistic values are maximized. For mated data, this is accomplished by conditioning on lateral pressures and other distortions such that the similarity statistic values are minimized and ensuring that the distributions represent the full range of plausible similarity statistic values that could reasonably be observed in casework when impressions are subject to various distortions during deposition. Keeping in mind that the similarity calculations do not take into account pattern type, feature type, specific feature configurations, or other details which may have biological dependencies, the empirical distributions were not conditioned on those specific criteria. However, because the similarity statistic calculations were designed to account for feature quantity, the distributions are calculated separately for each quantity of features (ranging from 5 to 15).

For the non-mated distributions, conditioning on the delta region was determined to maximize the opportunities of observing higher similarity statistic values. Appendix C-2 provides more specific details regarding this determination. The distributions of similarity statistic values

characterizing the broader population of non-mated samples for each quantity of features (ranging from 5 to 15) were generated using a subset of impressions from the National Institute of Standards and Technology (NIST) Special Database (SD) 27 [65], cropped to a standard size of 0.5in x 0.5in (12.7mm x 12.7mm) centered on the delta and randomly paired to non-mates. Features were annotated by practicing fingerprint experts beginning with those closest to the delta. Only *n* number of features under consideration were annotated in "image #1." All visible features, *m*, in "image #2" were annotated, such that $m >> n$ for each comparison. For each quantity of features, a distribution of 2,000 similarity statistic values was calculated and conditioned on any set of *n* features on image #1 determined to be "optimally paired" from the larger set of *m* possible features on image #2 with respect to the combinatorial optimization algorithm described in [63]. The two-sample Kolmogorov-Smirnov (K-S) test was used to evaluate the stability of the distributions. This was accomplished by comparing the distribution from one half of the dataset to the distribution from the other half of the dataset (each half distinct from one another) for each quantity of features. The K-S test was selected for this purpose on the basis of its ubiquitous use as a non-parametric test of the equality of continuous probability distributions. For all distributions, the K-S test resulted in a $p >> 0.05$ and determined to be sufficiently stable to permit parameter estimation and modeling of the population distributions.

For the mated distributions, a sample of fingerprints were collected from 50 different individuals using a livescan device with extreme distortions deliberately produced. This sample was determined to provide distributions representative of those observed in actual casework. Appendix C-3 provides more specific details regarding this determination. For the mated distribution, each individual provided eleven repeated impressions from the right thumb on the livescan device. The thumb was chosen because it results in maximal pliability of skin compared to the other fingers [64]. The repeat impressions consisted of one "non-distorted" impression used as the reference print and the remaining ten were made with lateral distortions applied in the following directions: north, south, east, west, northeast, northwest, southeast, southwest, twist clockwise, and twist counter-clockwise. Pressure was applied in the respective directions until the skin began to lose grip with the livescan surface. Of the 500 pairs obtained (ten distortions each for fifty different individuals), one pair lacked sufficient clarity to permit accurate determination of the corresponding features and therefore was discarded. Fifteen corresponding fingerprint features for the remaining 499 pairs of mated fingerprint impressions were annotated by practicing fingerprint experts in a federal crime laboratory in the United States. The distribution of similarity statistic values for each subset of feature quantities (ranging from 5 to 15) was calculated by randomly selecting (using a random selection algorithm) four combinations of *n* features out of *m* available (where $m = 15$). This resulted in 1,996 similarity statistic values for each quantity of features (ranging from 5 to 14) and 499 similarity statistic values for 15 features. The stability of the distributions were evaluated using a two-sample K-S test comparing the distribution from one half of the dataset to the distribution from the other half of the dataset (each half distinct from one another) for each quantity of features. For all distributions, the K-S test resulted in a $p >> 0.05$ and determined to be sufficiently stable to permit parameter estimation and modeling of the population distributions.

Parameter Estimation and Modeling

The empirical distributions of similarity statistic values described above (non-mated and mated) were modeled to determine plausible probability density functions which may model the similarity statistic values for the relevant populations of non-mated and mated friction ridge skin impressions. Taking into consideration the visual appearance of the empirical distributions and the construct of the weighting functions, the empirical distributions were each modeled using *k*-component (where $k = 2$ or 3) mixtures of Gaussian distributions. Component weights and parameter estimates were determined using maximum likelihood estimation methods within commercially available statistical analysis software (JMP). Although *k*-component Gaussian mixtures are more common, logistic distributions were applied on the basis of their heavier tails compared to Gaussian distributions. The heavier tails provide more conservative estimates of probabilities in the extreme ends of the distributions. The parameters for the logistic distribution were approximated using the estimated parameters of the Gaussian distributions. This was accomplished by setting the location parameter of the logistic distribution equal to the mean parameter of the Gaussian distribution as well as applying a coefficient to the standard deviation parameter of the Gaussian to approximate the scale parameter of the logistic distribution such that the difference between the two densities is minimized. Prior to estimating the component weights and parameter values, the empirical distributions were partitioned into two groups. For each bin of feature quantities, three-fourths of the sample was randomly selected using a random selection algorithm and used to estimate the population distribution parameters. The remainder of the sample was used to evaluate the goodness of fit of the estimated parameters for the population distribution. Once the optimal parameters were estimated, a one-sample K-S test was performed to evaluate the goodness of fit between the estimated theoretical logistic mixture distribution and the empirical distribution of the partition of similarity statistic values that was not used to estimate the theoretical distribution parameters. This process was repeated for each quantity of features (ranging from 5 to 15) for both mated and non-mated samples. The parametric models are proposed as plausible estimations of the population distributions for each quantity of features. Appendix C-4 provides more specific details regarding these determinations. Figures 3-3 and 3-4 illustrate the overlays between the theoretical density distributions and the empirical distributions of similarity statistic values for non-mated and mated datasets, respectively.

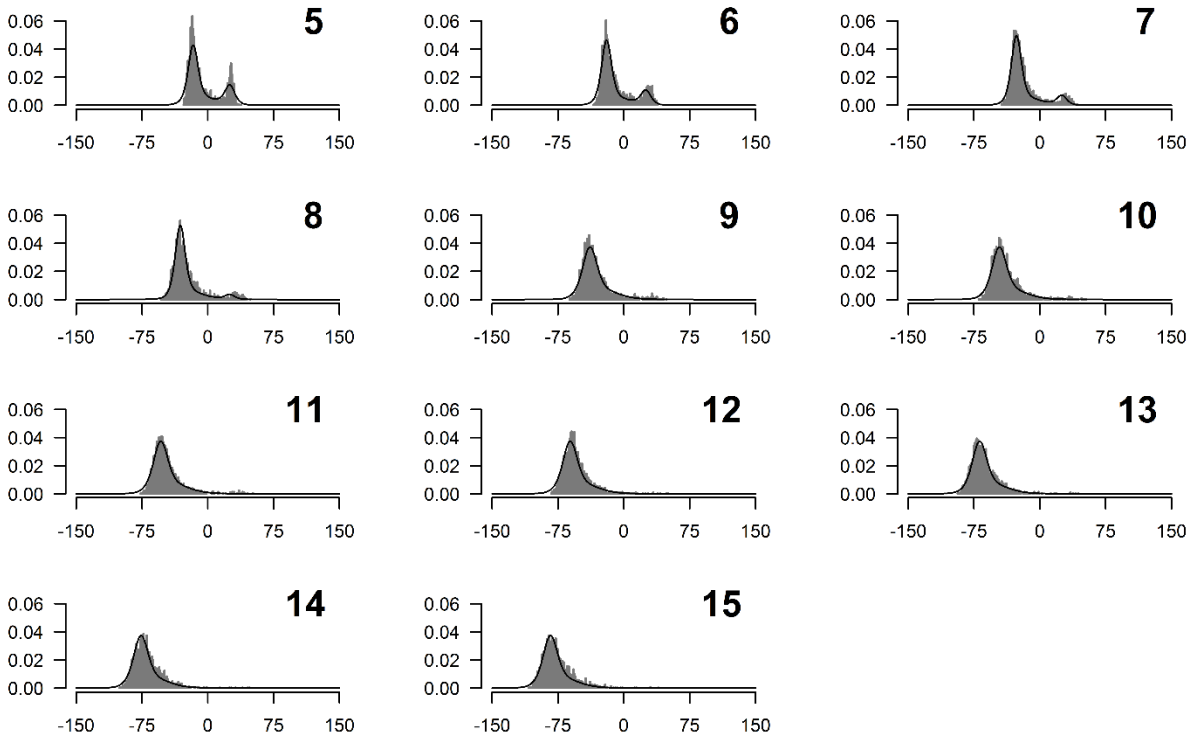


Figure 3-3. Empirical density distributions of the similarity statistic values for the non-mated sample (grey) compared to the theoretical (k-component logistic mixture) distribution (black) for each quantity of features (ranging from 5 to 15). The X-axis represents the global similarity statistic values. The y-axis represents the density.

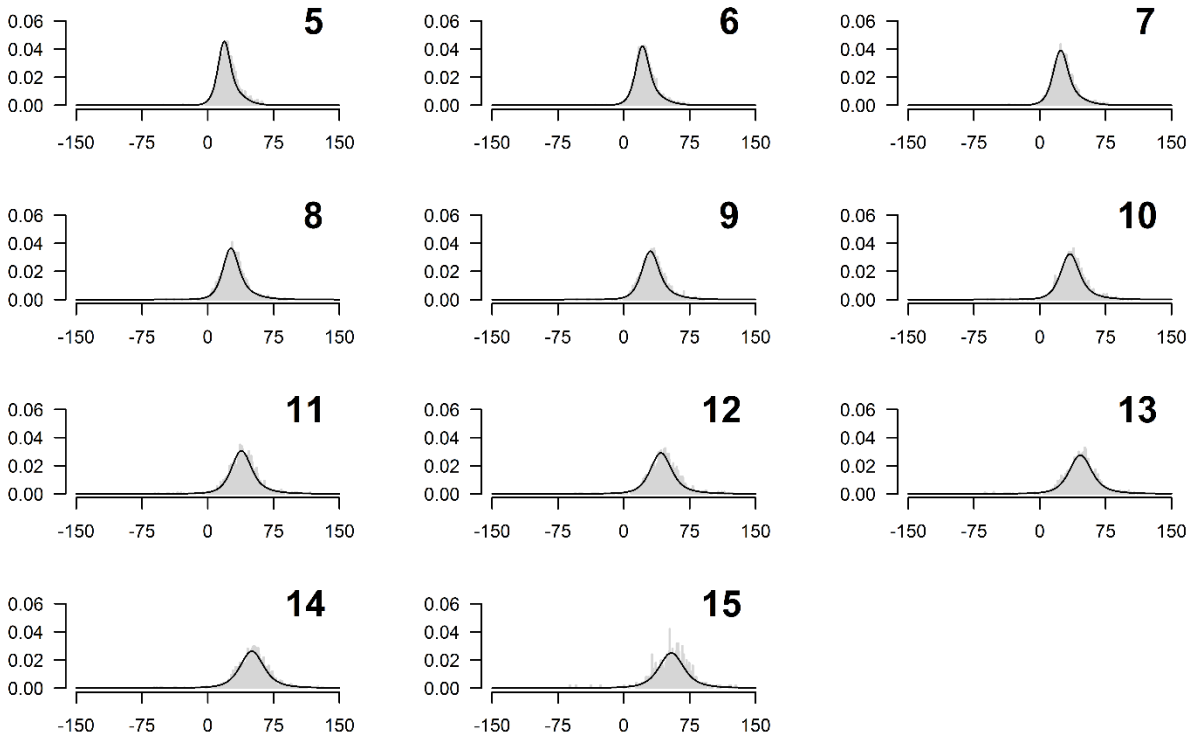


Figure 3-4. Empirical density distributions of the similarity statistic values for the mated sample (grey) compared to the theoretical (k-component logistic mixture) distribution (black) for each quantity of features (ranging from 5 to 15). The X-axis represents the global similarity statistic values. The y-axis represents the density.

Method Performance

The overall performance of the method was evaluated in terms of its sensitivity, specificity, within-sample variability, and between-sample variability. The performance of the method may be evaluated in terms of both the similarity statistic (i.e. global similarity statistic, GSS) values alone as well as in terms of the similarity statistic values in the context of the relevant probability distributions of mated vs. non-mated populations.

In terms of the mated distribution, the value of interest is the left tailed probability (the probability of a specific similarity statistic value *or lower*) as depicted in equation 3-1. In other words, the left tailed probability provides an indication of the proportion of similarity statistic values from mated source impressions which are estimated to be *less* than a specified test statistic value for a given case at hand. In terms of the non-mated distribution, the value of interest is the right tailed probability (the probability of a specific similarity statistic value *or higher*) as depicted in equation 3-2. In other words, the right tailed probability provides an indication of the proportion of similarity statistic values from non-mated source impressions which are estimated to be *greater* than a specified test statistic value for a given case at hand.

$$P\left(GSS_n \leq GSS(t)_n \middle| \theta_{n_{mated}}\right)$$

Equation 3-1: The left-tailed probability of observing a given similarity statistic, GSS(t), value or lower with respect to the distribution of GSS values from mated impressions, where "t" indicates the test statistic, "n" represents the feature quantity, and θ_n represents the parameters characterizing the distribution of values for a given feature quantity.

$$P\left(GSS_n \geq GSS(t)_n \middle| \theta_{n_{non-mated}}\right)$$

Equation 3-2: The right-tailed probability of observing a given similarity statistic, GSS(t), value or higher with respect to the distribution of GSS values from non-mated impressions, where "t" indicates the test statistic, "n" represents the feature quantity, and θ_n represents the parameters characterizing the distribution of values for a given feature quantity.

The values derived from equations 3-1 and 3-2 may be combined as a ratio, such that the estimated proportion of a given similarity statistic value *or lower* among mated sources is considered relative to the estimated proportion of a given similarity statistic value *or higher* among non-mated sources. Equation 3-3 combines equations 3-1 and 3-2 as the numerator and denominator, respectively.

$$\frac{P\left(GSS_n \leq GSS(t)_n \middle| \theta_{n_{mated}}\right)}{P\left(GSS_n \geq GSS(t)_n \middle| \theta_{n_{non-mated}}\right)}$$

Equation 3-3: Ratio of equations 3-1 and 3-2 indicating the relative support of a given similarity statistic, GSS(t), in terms of one proposition (mated) over another (non-mated).

From equation 3-3, values greater than 1 indicate a higher probability of the observed similarity statistic value among mated sources compared to non-mated sources and values less than 1 indicate a higher probability of the observed similarity statistic values among non-mated sources compared to mated sources. Values equal to 1 indicate equal probability of the observed similarity statistic value among mated and non-mated sources.

It is important to note that equations 3-1 and 3-2 are calculated as tail probabilities rather than likelihoods; thus, equation 3-3 is not a true likelihood ratio or Bayes' factor and should not be used as such with the intent of calculating a posterior probability.

Datasets

The performance of the method is evaluated using the following datasets:

1. Mated test dataset #1 (*known* to be mated) – A test dataset of 288 mated latent and reference impressions deposited under semi-controlled, normal handling conditions (to simulate casework) on a variety of different surfaces by 78 different individuals. The purpose of this dataset is to evaluate the performance of the method using latent and reference impressions which are similar to casework in terms of deposition and development, but for which ground truth mated status is known. Latent impressions were developed using a

variety of chemical and physical processing techniques commonly used in casework by fingerprint experts, such as cyanoacrylate ester fuming, fluorescent dye stains, ninhydrin, indanedione, 1-8 diazafluoren-9-one, and fingerprint powders. Each set was visually examined and corresponding features (ranging between 5 and 15) were manually annotated by practicing fingerprint experts in a federal crime laboratory in the United States. The overall quality (clarity) of the latent impressions is considered to be representative of casework impressions. This is based on the subjective evaluation by fingerprint experts as well as a comparison of the empirically measured quality scores using LQMetrics software available in the Universal Latent Workstation. A two-sample K-S test was performed comparing the distribution of LQMetric quality (clarity) scores from this dataset to the distribution of LQMetric clarity scores from the publically available dataset of casework impressions (mated test dataset #2 described below). The value of the K-S test statistic ($D_{288,184} = 0.087$) fails to reject the null hypothesis that the two samples originated from the same distribution ($p > 0.05$) based on a *p*-value decision threshold of 0.01.

2. Mated test dataset #2 (*accepted* to be mated) – A casework dataset of 184 latent and reference impressions publically available by the National Institute of Standards and Technology (NIST) Special Database 27 [65]. Although this dataset is commonly accepted to be mated by the general scientific community, it was collected from adjudicated casework by the Federal Bureau of Investigation and therefore ground truth is not actually known. The purpose of this dataset is to evaluate the performance of the method using latent and reference impressions from actual casework and which has been publically available and commonly used by the general scientific community. Each set was visually examined and corresponding features (ranging between 5 and 15) were manually annotated by practicing fingerprint experts in a federal crime laboratory in the United States. NOTE: The NIST Special Database 27 actually contains 258 latent and reference impressions in total; however, only 184 were able to be evaluated due to a technical issue with the remaining files preventing them from being opened (corrupted image files).

3. Mated test dataset #3 (*believed* to be mated) – A casework dataset of 605 latent and reference impressions collected from casework during the course of routine operations by fingerprint experts in a federal crime laboratory in the United States and reported as "positive associations." The purpose of this dataset is to evaluate the performance of the method using latent and reference impressions from a much larger sample of actual casework impressions as compared to the NIST Special Database 27 alone. The impressions were collected from a wide variety of cases, substrates, and assigned fingerprint experts. The corresponding features (ranging between 7 and 15) were manually annotated by the assigned fingerprint expert during the initial case examination. The selected features were then annotated later in a format suitable for *FRStat* analysis by the same fingerprint expert for purposes of this evaluation.

4. Non-mated test dataset #1 (*known* to be non-mated) – A test dataset of 20 latent print images from the mated test dataset #1 that were selected on the basis of representing the left delta region fingerprint impressions and 25 non-mated reference images obtained from the NIST Special Database 27 [65]. The purpose of this dataset is to evaluate the performance of the method using non-mated impressions for which the impressions were

arbitrarily paired and for which the impressions are publically available and commonly used by the general scientific community. For each latent print image, fifteen features were annotated around the delta region. Each reference print was cropped to a standard size of 0.5in x 0.5in (12.7mm x 12.7mm) centered on the left delta. All features visible in the cropped reference images were manually annotated by practicing fingerprint experts. For each comparison of the 20 latent prints to each of the 25 non-mated reference prints, a configuration of *n* features was randomly selected (using a random selection algorithm) from the latent print and compared against the reference print (each containing *m* annotated features, where $m >> n$) resulting in 500 similarity statistic values for each set of *n* features (ranging from 5 to 15). One similarity statistic value was obtained per image pair. The similarity statistic value was conditioned on any set of *n* features on image #1 determined to be "optimally paired" from the larger set of *m* possible features on image #2 with respect to the combinatorial optimization algorithm described in [63] under any condition of rotation and translation.

5. Non-mated test dataset #2 (*known* to be non-mated; "close non-match" from AFIS database search) – Two separate datasets: (#2a) a test dataset of fingerprint images representing the "delta" region and (#2b) a test dataset of fingerprint images representing the "core" region. The purpose of this dataset is to evaluate the performance of the method using non-mated impressions for which the impressions were paired on the basis of an AFIS similarity algorithm. Each dataset was separated into eleven separate subsets, each containing approximately 100 samples, conditioned on the number of features (*n*) being compared (ranging from 5 features to 15 features). Features were manually annotated by practicing fingerprint experts such that the features closest to the reference point (core or delta depending on the sample) were annotated first and then the remaining *n* features were annotated in a radiating fashion outward. Post annotation, each image was cropped by a bounding rectangle such that only those ridges and features that are part of the annotated configuration remain. These images serve as the "query" print. Each query print was then searched using an AFIS against an operational database containing approximately 100 million different fingerprint impressions from approximately 10 million different individuals. The AFIS ranked the top 20 most similar reference fingerprints to the fingerprint image searched. Of the top 20 results, the fingerprint image in rank 1 was confirmed to be a non-mated source with respect to the query print and used for comparison. Appendix C-2 provides more specific details regarding the development of this dataset.

Sensitivity & Specificity

The sensitivity was measured as the proportion of mated samples which resulted in a probability ratio value greater than a specified threshold ratio value. The specificity was measured as the proportion of non-mated samples which resulted in a ratio value less than a specified threshold ratio. Both the sensitivity and specificity will vary as a function of the ratio value chosen as a threshold. As the threshold ratio value increases, the sensitivity will decrease and the specificity will increase. As the threshold ratio value decreases, the sensitivity will increase and the specificity will decrease. Accordingly, both sensitivity and specificity were measured

separately using threshold ratio values of 1, 10, and 100, respectively. In addition to these threshold values, Receiver Operator Characteristics (ROC) curves illustrate the performance of the method across the full range of potential threshold values.

The sensitivity was evaluated using the mated test dataset #1 (*known* to be mated). Mated test dataset #2 (*accepted* to be mated) and mated test dataset #3 (*believed* to be mated) were also utilized to evaluate the consistency between threshold ratio values and experts' interpretation of mated status. The term "consistency" is used here since it is not a true measure of sensitivity because mated status is not truly known. Each dataset was considered separately. Of the total number of available latent and reference impressions in each dataset, up to ten different configurations of *n* features were randomly selected (using a random selection algorithm) from *m* available for each quantity of features (ranging between 5 and 15) to evaluate the results across the impressions subject to different conditions of distortion. Each configuration is considered as a separate measurement.

The specificity was evaluated using the non-mated test dataset #1 (*known* to be non-mated) as well as the non-mated test datasets #2a and #2b (*known* to be non-mated, "close non-match" from AFIS database search). The use of both datasets provides two different perspectives of the specificity as a result of prints being paired with non-mated impressions selected arbitrarily (non-mated dataset #1) as well prints being paired with the most-similar non-mated impression selected from a database of approximately 100 million others. In the latter context, "most-similar" is defined as the #1 rank candidate response from a large operational AFIS utilizing blackbox fingerprint search and matching algorithms. It is reasonable to consider the distribution of similarity statistic values from the non-mated test dataset #2 as representing the extreme tail of the distribution of values from the non-mated test dataset #1.

Within-Sample Variability & Between-Sample Variability

The variability of the method was evaluated separately in terms of the within-sample variability and between-sample variability of the similarity statistic values. The within-sample variability captures the variation as a result of multiple measurements of the *same* features. The between-sample variability captures the variation as a result of multiple measurements of *different* features and prints. Thus, the within-sample variability accounts for variations due to the imprecision and uncertainty of the specific location and angles of the feature annotations and the between-sample variability accounts for variations due to differences in distortions caused by pressure, substrate, etc. from different measurements across different configurations of features and impressions.

By taking into account the imprecision of feature annotations described in Appendix C-1, repeat measurements of the same features (without manual re-annotation) are subject to variation due to the random resampling scheme built into the method. The within-sample variability captures the variation of the similarity statistic values as a result of multiple measurements of the *same* features. The within-sample variability was evaluated using 92 image replicates from the mated test dataset #1 and mated test dataset #2, each of which contained 15 annotated features. Considering the intended use of this method is on impressions believed to be mated by the

fingerprint expert, the within-sample variability was not evaluated on the non-mated test datasets. For each image replicate, a configuration of *n* features was selected at random. Using the *same* configuration of *n* features for each respective replicate, a series of 25 repeat measurements were taken (where each measurement represents the lower bound of the 99% confidence interval of the *k*-iterations from the random resampling scheme; and where $k = 100$). The standard deviation of the 25 repeat measurements for each of the 92 image replicates was calculated. Using the standard deviations from each of the 92 image replicates, the combined standard deviation was calculated as the within-sample variability. This was repeated for each bin of feature quantities (ranging from 5 to 15).

The between-sample variability captures the variation of the similarity statistic values as a result of multiple (different) measurements of *different* features across different impressions. While variabilities of the similarity measurements as a result of the imprecision of the feature annotation process are taken into account in the similarity statistic calculations, the variabilities of the similarity measurements as a result of different conditions of distortion across different regions of an impression or across different impressions are not since they are not a consequence of repeat attempts to measure the same feature data. Rather, the between-sample variability is expected to represent a much larger range of similarity statistic values similar to the range of values represented by the estimated parameters of the population distributions discussed in further detail in Appendix C-4. The between-sample variability was evaluated using all image replicates from the mated test dataset #1 (*known* to be mated), mated test dataset #2 (*accepted* to be mated), and mated test dataset #3 (*believed* to be mated) combined. Considering the intended use of this method is on impressions believed to be mated by the fingerprint expert, the between-sample variability was not evaluated on the non-mated test datasets. For each of the total number of available latent and reference impressions from each mated test dataset (1,077), up to ten different *k*-configurations of *n* features were randomly selected (using a random selection algorithm) from *m* available for each quantity of features (ranging between 5 and 15) to evaluate the results across the impressions subject to different conditions of distortion. The standard deviation was calculated as the between-sample variability for each bin of feature quantities (ranging from 5 to 15).

The within-sample variability and between sample variability are both illustrated in terms of the similarity statistic value rather than in terms of the probability ratio because the impact to the probability ratio will vary depending on the location of the similarity statistic value within the distributions – subtle variations of the similarity statistic value in the tail of a distribution will cause a more dramatic change to the probability value compared to the other locations, such as the middle region. Thus, representing the variability in terms of the probability ratio itself would be incomplete and potentially misleading.

3.1.4 Results & Discussion

The overall performance of the method was evaluated in terms of its sensitivity, specificity, within-sample variability, and between-sample variability. Initially, the expected performance may be evaluated in terms of comparing the empirical distributions of similarity statistic values between mated and non-mated impressions. These distributions served as the empirical foundation

for the parameter estimations and modeling described in greater detail in Appendix C-4. Figure 3-5 illustrates the empirical distributions in terms of density.

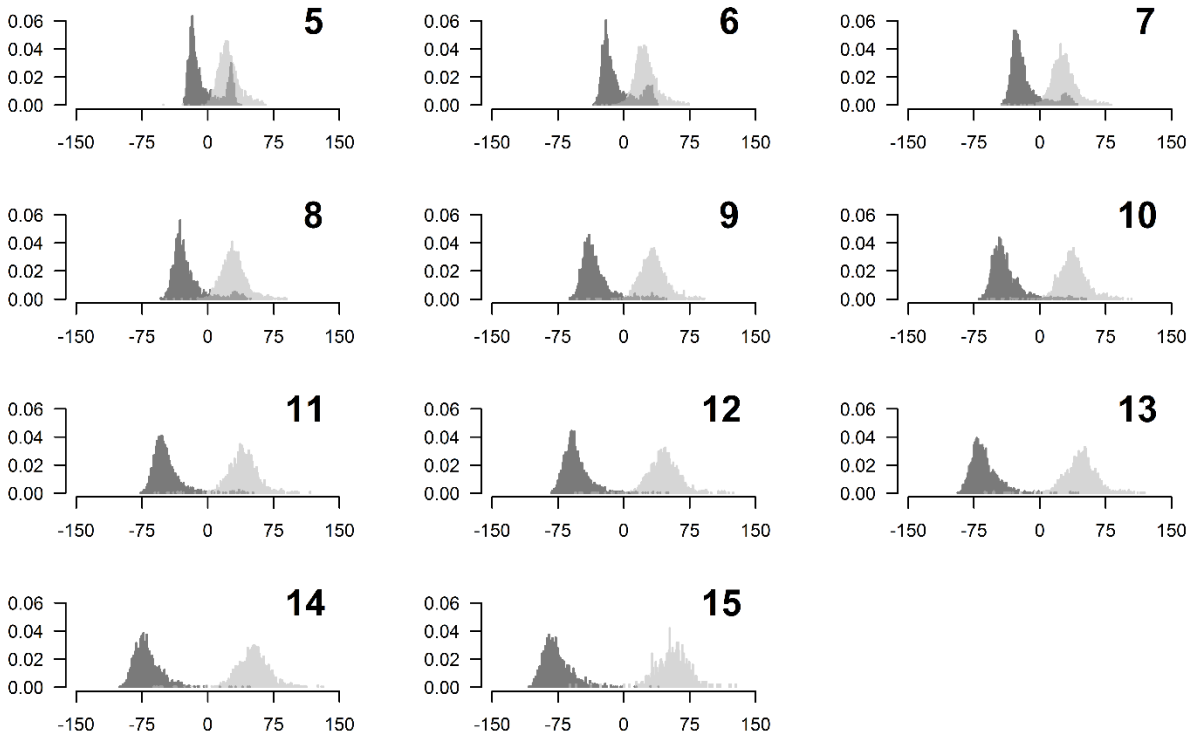


Figure 3-5. Empirical distributions of similarity statistic values for both non-mated (dark grey) and mated (light grey) samples for feature quantities 5 through 15. The X-axis represents the global similarity statistic values. The y-axis represents the density.

From figure 3-5, two important observations can be made. First, we see that the distributions appear to exhibit little overlap between the mated and non-mated datasets. Second, we see that the distributions appear to increase in separation as the feature quantities increase.

Sensitivity

The sensitivity was evaluated using the mated test dataset #1 (*known* to be mated). Mated test dataset #2 (*accepted* to be mated) and mated test dataset #3 (*believed* to be mated) were also utilized to evaluate the consistency between threshold ratio values and experts' interpretation of mated status ("consistency" is used here since it is not a true measure of sensitivity because mated status is not truly known). Each dataset was considered separately. Table 3-1 provides the sensitivity using mated test dataset #1. Table 3-2 provides the consistency between the method and experts' interpretation of mated status using mated test dataset #2. Table 3-3 provides the consistency between the method and experts' interpretation of mated status using mated test dataset #3.

Feature Quantity	**Number of configurations** *(Mated Dataset #1)*	**Sensitivity** *(Ratio >1)*	**Sensitivity** *(Ratio >10)*	**Sensitivity** *(Ratio >100)*
5	2,798	0.657	0.249	0.085
6	2,703	0.708	0.381	0.145
7	2,550	0.736	0.478	0.234
8	2,367	0.823	0.593	0.402
9	2,092	0.892	0.755	0.565
10	1,898	0.928	0.824	0.645
11	1,655	0.947	0.860	0.710
12	1,432	0.970	0.925	0.799
13	1,230	0.984	0.949	0.825
14	994	0.980	0.971	0.902
15	97	0.990	0.979	0.959

Table 3-1. Sensitivity of the method using mated test dataset #1 (known to be mated) for each quantity of features (ranging from 5 to 15). Sensitivity was evaluated using a ratio of 1, 10, and 100 as the thresholds.

Feature Quantity	**Number of configurations** *(Mated Dataset #2)*	**Consistency** *(Ratio >1)*	**Consistency** *(Ratio >10)*	**Consistency** *(Ratio >100)*
5	1,772	0.730	0.201	0.052
6	1,674	0.783	0.317	0.100
7	1,512	0.830	0.446	0.163
8	1,317	0.913	0.636	0.328
9	1,166	0.959	0.852	0.595
10	988	0.966	0.899	0.721
11	781	0.968	0.948	0.827
12	706	0.965	0.965	0.905
13	583	0.971	0.971	0.949
14	480	0.973	0.960	0.960
15	47	0.979	0.957	0.957

Table 3-2. Consistency between ratio values greater than 1, 10, and 100 and experts' interpretation of mated status using mated test dataset #2 (accepted to be mated) for each quantity of features (ranging from 5 to 15).

Feature Quantity	**Number of configurations** *(Mated Dataset #3)*	**Consistency** *(Ratio >1)*	**Consistency** *(Ratio >10)*	**Consistency** *(Ratio >100)*
5	6,050	0.794	0.287	0.088
6	6,038	0.840	0.436	0.150
7	5,982	0.870	0.530	0.239
8	5,830	0.927	0.716	0.437
9	5,526	0.955	0.889	0.690
10	5,040	0.961	0.927	0.805
11	4,441	0.965	0.934	0.868
12	3,876	0.971	0.953	0.910
13	3,226	0.970	0.958	0.920
14	2,638	0.978	0.974	0.961
15	258	0.981	0.977	0.970

Table 3-3. Consistency between ratio values greater than 1, 10, and 100 and experts' interpretation of mated status using mated test dataset #3 (believed to be mated) for each quantity of features (ranging from 5 to 15).

With respect to the sensitivity calculations listed above, it is important to note that the values were generated *without* the examiners having direct feedback regarding their annotation precision. Without such feedback, examiners have become acclimated to a relaxed environment in which they were accustomed to annotating the mere presence of a feature and in which measurements were not taken directly from the annotations. In practice, where a fingerprint expert recognizes the importance of precise annotations and adjusts accordingly, it is a reasonable assumption that the sensitivity will be *higher* (and thus the false negative rate will be *lower*) than what is represented in this section; however, a quantitative measure of *how much* higher the sensitivity would be in practice is unknown at this time. Nevertheless, the sensitivity of the method is expected to increase as examiners gain more experience and become more precise in their feature annotations – similar to when examiners gain a better understanding of how feature annotations impact the performance of AFIS search results and adjust their annotation habits accordingly.

Specificity

The specificity was evaluated using the non-mated test dataset #1 (*known* to be non-mated) as well as the non-mated test datasets #2a and #2b (*known* to be non-mated, "close non-match" from AFIS database search). The use of both datasets provides two different perspectives of the specificity as a result of prints being paired with non-mated impressions selected arbitrarily (non-mated dataset #1) as well prints being paired with the most-similar non-mated impression selected from a database of approximately 100 million others. Table 3-4 provides the specificity using non-mated test dataset #1. Table 3-5 provides the specificity using non-mated test datasets #2a and #2b (table 3-5a – "delta" region; table 3-5b – "core" region).

Feature Quantity	**Number of image pairs** *(Non-mated Dataset #1)*	**Specificity** *(Ratio <1)*	**Specificity** *(Ratio <10)*	**Specificity** *(Ratio <100)*
5	500	0.818	1.000	1.000
6	500	0.850	0.992	1.000
7	500	0.900	0.994	1.000
8	500	0.912	0.986	1.000
9	500	0.940	0.952	0.990
10	500	0.970	0.976	0.992
11	500	0.978	0.982	0.990
12	500	0.988	0.992	0.998
13	500	0.988	0.994	0.996
14	500	0.988	0.992	0.994
15	500	0.996	1.000	1.000

Table 3-4. Specificity of the method using non-mated test dataset #1 (known to be non-mated) for each quantity of features (ranging from 5 to 15). Specificity was evaluated using a ratio of 1, 10, and 100 as the thresholds.

Feature Quantity	**Number of image pairs** *(Non-mated Dataset #2a – "delta" region)*	**Specificity** *(Ratio <1)*	**Specificity** *(Ratio <10)*	**Specificity** *(Ratio <100)*
5	99	0.566	0.788	0.980
6	99	0.687	0.747	0.980
7	96	0.688	0.719	0.896
8	99	0.747	0.788	0.812
9	99	0.818	0.818	0.828
10	97	0.814	0.835	0.845
11	96	0.802	0.823	0.823
12	98	0.857	0.867	0.888
13	99	0.899	0.929	0.939
14	100	0.980	0.990	0.990
15	100	0.920	0.920	0.940

Table 3-5a. Specificity of the method using non-mated test dataset #2a (known to be non-mated; "close non-match" from AFIS database searches of the delta region) for each quantity of features (ranging from 5 to 15). Specificity was evaluated using a ratio of 1, 10, and 100 as the thresholds.

Feature Quantity	**Number of image pairs** *(Non-mated Dataset #2b – "core" region)*	**Specificity** *(Ratio <1)*	**Specificity** *(Ratio <10)*	**Specificity** *(Ratio <100)*
5	94	0.787	0.979	1.000
6	96	0.802	0.927	1.000
7	95	0.884	0.926	0.979
8	96	0.906	0.938	1.000
9	95	0.884	0.952	0.990
10	96	0.969	0.990	1.000
11	95	0.989	0.989	0.989
12	97	1.000	1.000	1.000
13	97	1.000	1.000	1.000
14	96	1.000	1.000	1.000
15	95	1.000	1.000	1.000

Table 3-5b. Specificity of the method using non-mated test dataset #2b (known to be non-mated; "close non-match" from AFIS database searches of the core region) for each quantity of features (ranging from 5 to 15). Specificity was evaluated using a ratio of 1, 10, and 100 as the thresholds.

With respect to the specificity calculations listed above, it is important to note that the values are limited to the output of the *FRStat* algorithm alone; thus, these values should not be confused with the overall specificity of the latent print examination method in general which is much improved by the input of the fingerprint expert. In practice, where a fingerprint expert's visual examination will precede the calculation of a similarity statistic value using *FRStat* and serve as an initial means of discrimination using details that *FRStat* is not designed to take into account, it is a reasonable assumption that the specificity will be much *higher* (and thus the false positive rate will be much *lower*) than what is represented in this section. However, because there are no publically available datasets to empirically measure how often non-mated impressions are falsely included by fingerprint experts *and* which result in sufficiently high similarity statistic values using this method, a quantitative measure of *how much* higher the specificity would be in practice cannot be determined at this time.

Receiver Operator Characteristic (ROC)

The Receiver Operator Characteristic (ROC) illustrates the performance of the method across the full range of potential threshold values. Figure 3-6 illustrates the ROC curves for mated test dataset #1 (*known* to be mated) and non-mated test dataset #1 (*known* to be non-mated) as well as the non-mated test datasets #2a and #2b (*known* to be non-mated, "close non-match" from AFIS database search). The use of both non-mated datasets provides two different perspectives of the performance of the method as a result of prints being paired with non-mated impressions selected arbitrarily (non-mated dataset #1) as well prints being paired with the most-similar non-mated impression selected from a database of approximately 100 million others.

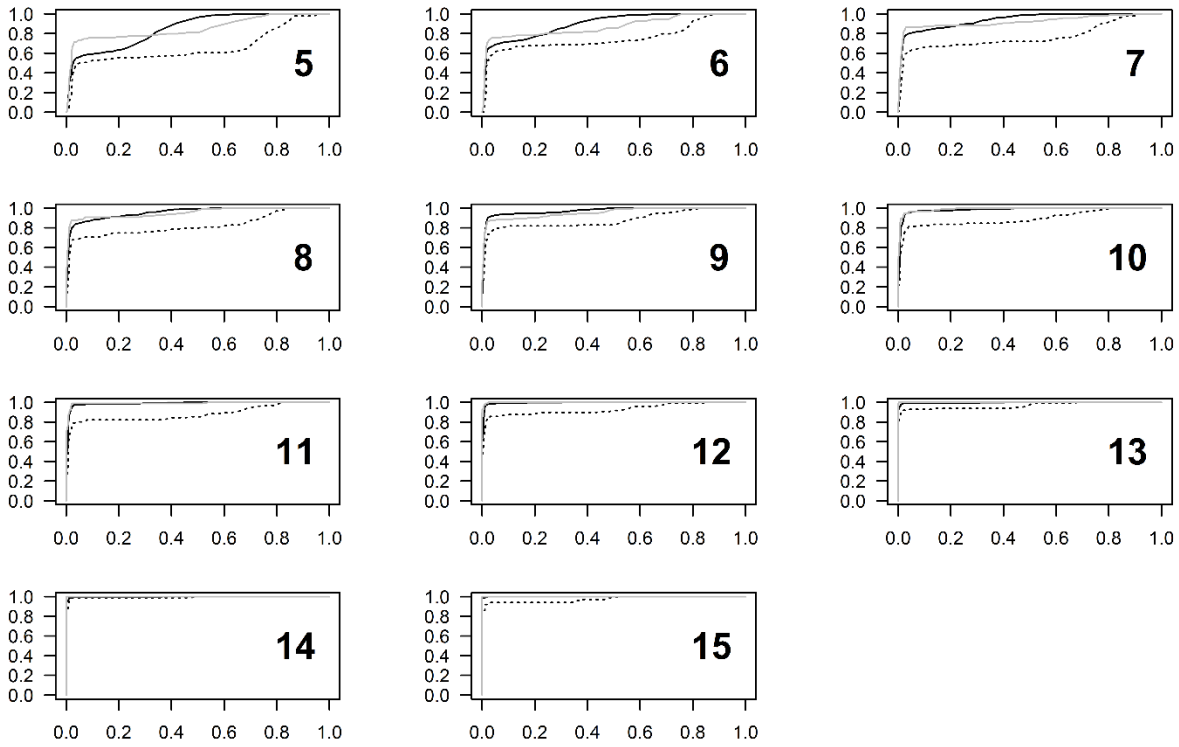


Figure 3-6. ROC curves illustrating the performance of the method using mated test dataset #1 (known to be mated) and non-mated test datasets #1 (known to be non-mated) as well as non-mated test datasets #2a (known to be non-mated; "close non-match" from AFIS database searches of the delta region) and #2b (known to be non-mated; "close non-match" from AFIS database searches of the core region) for each quantity of features (ranging from 5 to 15). The solid black line represents the ROC using non-mated test dataset #1(known to be non-mated). The dotted black line represents the ROC using non-mated test dataset #2a (known to be non-mated; "close non-match" from AFIS database searches of the delta region). The solid grey line represents the ROC using non-mated test dataset #2b (known to be non-mated; "close non-match" from AFIS database searches of the core region). The X-axis represents 1 - specificity. The y-axis represents the sensitivity.

From figure 3-6 as well as tables 3-4 and 3-5, we can make two important observations. First, the specificity rates from non-mated dataset #1 and non-mated dataset #2b are very similar to one another. Second, while the specificity rates from non-mated dataset #2a provides an indication of the "worst case-scenario" since it narrowly focuses on the #1 rank candidates out of approximately 100 million other non-mated prints as a result of AFIS searches *and* only considers the delta region of the fingerprint during the searches, the method still demonstrates the ability to accurately classify mated and non-mated impressions. Taking together, the performance characteristics discussed above may provide some general context to the results when non-mated samples are selected at random or whether they were selected on the basis of their similarity from large database searches. The samples comprising non-mated datasets #2a and #2b are limited in size due to operational constraints at the time of collection. A likely consequence of the small sample sizes is the subtle variability in the performance characteristics observed between the various feature quantities, particularly between 13, 14, and 15 features where the observed data suggests 14 features had better performance characteristics than 15 features. With a larger sample, the uncertainty associated with the performance characteristics will be reduced; therefore, further research into the impact of AFIS searches on the specificity rates is encouraged. Nevertheless,

because the intent of the method is to estimate the relative prevalence of similarity statistic values among the broader population of non-mated impressions rather than focus only on "close non-mates" from large database searches, the low sample size of these datasets (#2a and #2b) is not considered a critical limitation – their selection as the #1 rank candidate means they were already distinguished from all other impressions in the system using the high performance AFIS algorithms.

Within-Sample Variability

The within-sample variability captures the variation of the similarity statistic values as a result of multiple measurements of the *same* features without re-annotations (due to the random resampling scheme discussed in greater detail in Appendix C-1). Table 3-6 provides the within-sample variability of the method in terms of the combined standard deviation of similarity statistic values. These results demonstrate very low within sample variability and are insignificant compared to the between-sample variability.

Feature Quantity	**Combined σ GSS(t)**	**Mean GSS(t)**
5	0.593	20.742
6	0.648	20.202
7	0.651	24.736
8	0.692	25.104
9	0.831	25.869
10	0.903	32.910
11	0.916	33.371
12	0.969	37.555
13	1.067	39.275
14	1.196	42.979
15	1.244	47.464

Table 3-6. Within-sample variability (combined standard deviation from 25 repeat measurements each for 92 different images) of the similarity statistic value (GSS(t)) for each quantity of features (ranging from 5 to 15).

Between-Sample Variability

The between-sample variability captures the variation of the similarity statistic values as a result of multiple (different) measurements of *different* features. Table 3-7 provides the between-sample variability of the method in terms of the similarity test statistic. These results demonstrate between-sample variabilities consistent with those represented by the estimated parameters of the population distributions discussed in further detail in Appendix C-4 and are therefore consistent with expectations.

Feature Quantity	Number of configurations	Mean GSS(t)	σ GSS(t)
5	10,620	20.864	13.585
6	10,415	23.849	15.112
7	10,044	25.372	16.681
8	9,514	29.557	18.41
9	8,784	32.392	19.642
10	7,926	36.602	21.666
11	6,877	39.826	23.653
12	6,014	44.864	25.133
13	5,039	47.81	27.192
14	4,112	52.908	27.698
15	402	56.952	29.233

Table 3-7. Between-sample variability (standard deviation) of the similarity statistic value (GSS(t)) for each quantity of features (ranging from 5 to 15).

General Discussion

Ratio Values

The ratio values obtained with the method will vary depending on the measured similarity between the two impressions, reflected by the global similarity statistic, GSS(t), as well as the quantity of features. As the GSS(t) value and quantity of features increase, the ratio value will also increase indicating stronger significance of the association between the paired impressions. Theoretically, the ratio values can range from negative infinity to positive infinity; however, this provides little context to understanding the range of ratio values that one may plausibly observe in practice. Figure 3-7 illustrates the range of ratio values based on the GSS(t) values corresponding to 95% of the theoretical distribution modeling the mated source dataset (ranging from a left tail probability of 0.025 to 0.975) for each quantity of features.

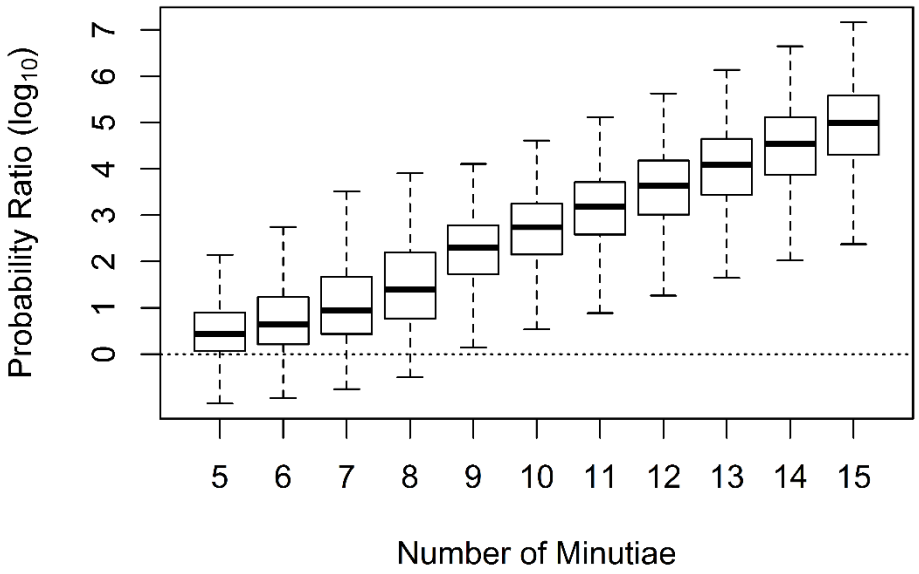


Figure 3-7. Box plots illustrating the plausible range of ratio values that may be reasonably expected for each quantity of features based on GSS(t) values corresponding to 95% of the theoretical distribution modeling the mated source dataset (ranging from a probability of 0.025 to 0.975). The X-axis represents the number of features (ranging from 5 to 15). The y-axis represents the log_{10} ratio value.

From figure 3-7, we observe a steady increase of ratio values as the quantity of features increases. This steady increase is a mathematical consequence of the algorithms for calculating the similarity statistic and consistent with the expected behavior of the method in terms of experience by forensic experts. Although the actual ratio values are much lower than what experts might expect, these ratio values are highly conservative since: (1) the method does not take into account all aspects of the impression, such as pattern type, feature type, ridge counts, and other types of features considered by an expert, (2) the similarity statistic value provides a single dimensional summary of the similarity between two impressions and does not consider the prevalence of the specific arrangement of features under consideration within the population, (3) the empirical distributions of similarity statistic values were conditioned such that the non-mated distribution was biased towards higher similarity statistic values (in terms of randomly paired impressions) and the mated distribution was biased towards lower similarity statistic values, and (4) logistic mixture distributions were chosen to model the empirical distributions of similarity statistic values on the basis of their heavier tails thus providing more conservative estimates of probabilities in the extreme ends of the distributions compared to Gaussian mixture distributions.

Although the ratio values provide a measure of the significance (i.e. strength of an association) between two impressions, common practice by forensic experts is to conduct an experience-based judgment and classify an impression as originating from a specific individual (i.e. individualization decision) based on personal confidence and subjective observation. The accuracy of expert determinations of individualization has been evaluated by Ulery et al. [66] finding approximately 0.1% false individualization rate. In a subsequent study, [67] Ulery et al. found that individualization determinations increase as the number of annotated features increase.

Further, among all individualization decisions (n = 1,653), only 1% were based on mated comparisons containing less than 7 features and among all mated comparisons with 12 or more features, 98.4% resulted in individualization decision. Table 3-8 provides the percentage of individualization decisions for each number of features (ranging from 5 to 15) from [67]. Although a loose comparison, given the accuracy of individualization determinations from [66] and the breakdown of individualization decisions as it relates to the number of annotated features from [67], these data may provide some general context for understanding how the results from this method compare to performance metrics and individualization decision behaviors by experts in traditional practice. Interestingly, if we compare the inter-quartile range of ratio values for each quantity of features from figure 3-7 above to the individualization determinations in table 3-8, we see that the inter-quartile ranges for 9 or more features exceeded a ratio of 10, which correspond to reasonably high specificity rates. Having discussed the comparisons between the ratio values of this method and experts' performance when making individualization decisions, caution should be exercised to ensure the probability estimates from this method are not incorrectly interpreted. The results provide the ratio of the estimated probabilities of a given similarity statistic value or more extreme among datasets of similarity statistic values from mated and non-mated comparisons. The results do not provide the probability of observing a *specific configuration* of features in the population or the probability that a *specific individual* is the source of an impression. Accordingly, although this method will provide an empirical foundation to the strength of an association between two impressions, determinations that specific individual is *the* source of an impression (i.e. individualization decisions) remain a subjective opinion by the expert.

Feature Quantity	**% Individualization Decisions**
5	2
6	17
7	47
8	64
9	81
10	90
11	92
12	95
13	97
14	99
15	96

Table 3-8. Percentage of individualization decisions by fingerprint experts on fingerprint images having different numbers of features (ranging from 5 to 15). Table values estimated from figure 3-3B in [67].

Method Limitations

The major limitations of the method include: (1) The similarity statistic values are dependent upon the subjective detection and annotation of friction ridge skin features by the human expert. (2) The method is only able to consider what the expert annotates and is not able to evaluate

the accuracy of feature annotations by the expert. (3) The method requires a minimum of five features and a maximum of fifteen features. The minimum of five features is due to the manner in which the similarity statistic is calculated. The maximum of fifteen features was a cutoff decision by the authors due to the computational impact of running the pairing algorithm on configurations containing higher numbers of features based on the current software implementation. For friction ridge skin impressions that contain more than fifteen features, only fifteen features can be encoded for statistical evaluation. This does not prevent the expert from making reference to the additional features available, but were not able to be encoded and evaluated by this version of the software application. (4) The weight functions are based on lateral distortions of friction ridge skin impressions on flat surfaces and may not capture all types of extreme distortions which may be encountered in practice, such as substrate, matrix, or photographic effects. (5) The method is not designed to evaluate all aspects of the impression, such as pattern type, feature type, ridge counts, and other types of features considered by an expert; thus, the quantitative results are artificially attenuated and conservative.

Considerations for Policy and Procedure

Taking into consideration the major limitations described above, general considerations for policy and procedure include: (1) The method should only be used *after* the expert has visually analyzed, detected, and annotated friction ridge skin features which are believed to correspond between two separate impressions of friction ridge skin. The method should not be used on impressions in which the analyst is able to visually exclude the two impressions as originating from the same source. (2) The method should be used in accordance with a set of strict policies and procedures to guard against potential cognitive biases in the analysis, detection, interpretation and annotation of friction ridge skin features as well as a quality assurance program to verify the accuracy of the annotated features. (3) The method should be used on digital images having a resolution of 500 pixels per inch or higher to ensure distance calculations are not impacted by lower resolution images.

Despite the limitations described above, this method provides several advantages which far outweigh the limitations. Most importantly, it provides fingerprint experts the capability to demonstrate the reliability of fingerprint evidence *for the case at hand* and ensure the evidence is reported with an empirically grounded basis. Further, having the ability to quantify the strength of fingerprint comparison, the evidence can be reported in a more transparent and standardized fashion with clearly defined criteria for conclusions and known error rate information. Appendix C-5 provides an example demonstrating the use of *FRStat*.

3.1.5 Conclusion

Over the years, the forensic science community has faced increasing amounts of criticism by scientific and legal commentators, challenging the validity and reliability of many forensic examination methods that rely on subjective interpretations by forensic practitioners. Among those concerns is the lack of an empirically demonstrable basis to evaluate and report the strength of the fingerprint evidence for a given case. In this paper, a method is presented which provides a

statistical assessment of the strength of fingerprint evidence. The method measures the similarity between friction ridge skin impressions using details annotated by human experts to calculate a similarity statistic (i.e. score), which is then evaluated against databases of similarity statistic values derived from pairs of impressions made by mated (same) and non-mated (different) sources of friction ridge skin impressions relevant for forensic casework. The distributions of similarity statistic values were developed such that the non-mated data are biased to *higher* similarity statistic values and mated data are biased to *lower* similarity statistic values. For non-mated data, this was accomplished by conditioning on (1) the delta region of friction ridge skin which was determined to maximize the opportunities of observing higher similarity statistic values, and (2) any set of *n* features determined to be "optimally paired" from a larger set of *m* possible features with respect to a combinatorial optimization algorithm under any condition of rotation and translation such that the similarity statistic values are maximized. For mated data, the bias to lower values was accomplished by conditioning on lateral pressures and other distortions such that the similarity statistic values are minimized and ensuring that the distributions represent the full range of plausible similarity statistic values that could reasonably be observed in casework when impressions are subject to various distortions during deposition. The empirical distributions were statistically modeled and plausible estimates of population parameters were evaluated using the Kolmogorov-Smirnov (K-S) "goodness of fit" test. The K-S test was selected for this purpose on the basis of its ubiquitous use as a non-parametric test of the equality of continuous probability distributions. The strength of the fingerprint evidence is calculated as a ratio of the tail probabilities from the distributions of similarity statistic values of mated and non-mated impressions. The numerator is the left tail probability of a given similarity statistic value or *lower* among the distribution of values from mated sources. The denominator is the right tail probability of a given similarity statistic value or *higher* among the distribution of values from non-mated sources. Although similar in appearance, the ratio is not a true likelihood ratio or Bayes' factor and therefore should not be used to estimate a posterior probability for a proposition.

The performance of the method was evaluated using a variety of different mated and non-mated datasets, including the most similar non-mated impressions from AFIS searches against a database of approximately 100 million other fingers. The results show strong performance characteristics. As the number of features increase, the magnitude of the ratio values increase as well as the ability to discriminate between mated and non-mated impressions, often with values supporting specificity rates greater than 99%. Despite the trend of increasing ratio values, there is still some overlap of the values between the different quantities of features. Consequently, similar to the findings in [37, 42], these data demonstrate the importance of evaluating the strength of the fingerprint evidence based on the measurable attributes of the given comparison rather than relying on generalizations based solely on the number of features.

As with any method, there are limitations to consider. For example, this method relies on the features annotated by the expert but does not take into account all aspects of fingerprint evidence. As a result, the quantitative results for reported associations using this method (*FRStat*) will be artificially low. Despite the limitations, *FRStat* provides fingerprint experts the capability to demonstrate the reliability of fingerprint evidence *for the case at hand* and ensure the evidence is evaluated with an empirically grounded basis. Further, having the ability to quantify the strength of the fingerprint comparison, the evidence can be reported in a more transparent and standardized fashion with clearly defined criteria for conclusions and known error rate information.

Although various aspects of the method may be further optimized, the performance characteristics described are proposed as a sufficient basis to demonstrate the foundational validity of the method to perform within the scope of its intended purpose – as a means of providing a statistical measure of the strength of a given fingerprint comparison. Further optimizations which may improve upon the method's performance are encouraged for future works.

3.2 Comparison with Other Methods

This section is supplemental to the published manuscript [50] and explores the utility of the FRStat compared to another method available at the University of Lausanne when applied as a quality control within a quality management system. Although both systems have their own benefits and limitations, the results show that each method has the capacity to distinguish between mated and non-mated impressions with reasonable accuracy and provide an additional layer of quality control to the overall examination scheme.

3.2.1 Background

The FRStat algorithm presented in this chapter provides a measure of the similarity between two sets of features identified by an analyst on two separate impressions which the analyst believes to correspond, and provides a statistical assessment of the significance of that correspondence. The FRStat algorithm, however, is not the only algorithm that has been proposed for purposes of assessing the strength of fingerprint evidence. Among several other methods that have been proposed, an updated version of the model originally described by Egli [33], which is based on the distribution of similarity scores from AFIS, has been made accessible for evaluation (referred to herein as the AFIS-SLR). Although both the FRStat and the AFIS-SLR provide measures of similarity and statistical assessments of the strength of fingerprint evidence, they do so in very distinct ways.

The FRStat, as described in [50], first calculates the similarity (referred to as the *Global Similarity Statistic*, or GSS) between the impressions, then it provides two estimates: one indicating how often impressions originating from common sources would result in a GSS that is equal to or less than the calculated GSS and another indicating how often impressions from different sources would result in a GSS that is equal to or greater than the calculated GSS. The two values are then combined as a ratio providing a single summary statistic indicating to what extent the GSS is consistent with impressions originating from a common source compared to different sources.

The AFIS-SLR model was originally developed by Egli [33] based on a commercial AFIS matcher developed by Sagem-Morpho (now Idemia). Initially, the prototype was relatively slow and required a case-by-case establishment of the within-source variability using pseudo-marks obtained from a range of prints provided by the person of interest. The model was later matured through the work of Marco De Donno[7] who added a distortion model based on the thin plate spline

[7] Marco De Donno is a current Ph.D. student at the School of Criminal Justice, University of Lausanne.

model of Bookstein [68] which improved computing times by taking advantage of multi-core and parallel processing. The model has also been adapted for the assessment of the expected weight of evidence to be assigned to a given mark described by Stoney et al. [61]. Generally, the AFIS-SLR first creates a dataset of pseudo-marks representing the mark in algorithmically distorted ways and calculates the similarity using the AFIS matcher between the mark and each of the pseudo-marks (resulting in a distribution of similarity scores representing the *within-source* variability). Next, it uses the AFIS algorithm to calculate the similarity between the mark and prints in the database (resulting in a distribution of similarity scores representing the *between-source* variability). Then, it calculates the similarity score between the mark and print for the case at hand. The distributions of scores representing the *within-source* variability and *between-source* variability are both fit to probability density distributions and the likelihood ratio for the similarity score resulting from the mark-print comparison for the case at hand is calculated (referred to as the score-based likelihood ratio, or SLR).

From the descriptions above, we note that the FRStat and AFIS-SLR apply very different mathematical approaches in their calculations of similarity and assessments of statistical strength (i.e., ratio values). Each method has its own benefits and limitations in terms of computational complexity and algorithmic transparency. The FRStat is a computationally light and transparent algorithm implemented into a stand-alone software application that is freely available; however, it lacks the ability to calculate the similarities of the feature configuration for the case at hand directly against a database of other impressions. The AFIS-SLR is a computationally complex system that benefits from intensive search engines powered by a commercial AFIS technology to calculate the similarities of the feature configuration for the case at hand directly against databases of other impressions; however, it requires greater computational resources to operate and lacks algorithmic transparency due to the proprietary nature of the commercial matching algorithms. Taking into account these attributes of the two methods, it is relevant to conduct an exploratory comparison of the performance of the FRStat and the AFIS-SLR to understand the impact of these tradeoffs to performance. Although the ratio value produced by the FRStat and the AFIS-SLR are fundamentally different mathematical constructs which prevents a straightforward comparison of the two systems on the basis of their magnitudes, they both share a common objective – to provide an empirical foundation to examiners' subjective assessments and help detect circumstances for which additional quality assurance review might be warranted. Thus, rather than comparing the magnitudes of the ratio values produced by the two systems, we can compare the performance of the two systems in terms of their ability to accurately distinguish between mated and non-mated impressions. This type of comparison is informative as it provides traditional performance characteristics of the systems as it relates to their use as an additional layer of quality control by flagging impressions as potentially problematic prior to issuing a conclusion in the case.

3.2.2 Materials & Methods

This comparison of the performance between the FRStat and the AFIS-SLR was conducted using the same datasets provided by Swofford et al. [50], which include the datasets used to model the empirical distributions as well as the mated and non-mated datasets used to test the performance of the FRStat (i.e., Mated Empirical Distribution [*known* to be mated], Non-mated Empirical Distribution [*known* to be non-mated], Mated Test Dataset #1 [*known* to be mated],

Mated Test Dataset #2 [*accepted* to be mated], Mated Test Dataset #3 [*believed* to be mated], Non-mated Test Dataset #1 [*known* to be non-mated], and Non-mated Test Dataset #2 [*known* to be non-mated; "close non-match" from AFIS database search – delta and core regions]). The makeup of each dataset is described in detail in [50].

The baseline performance of the FRStat to which the performance of the AFIS-SLR is compared against is derived from the raw performance data provided in [50]. The performance of the FRStat and the AFIS-SLR were evaluated in two distinct ways. First, the performance was evaluated in terms of the ability for each method to accurately distinguish between mated and non-mated impressions using values of their ratios alone—for both systems, ratio values greater than 1 are categorized as mated impressions and values less than 1 are categorized as non-mated impressions. Second, the performance was evaluated in terms of the ability for each method to accurately distinguish between mated and non-mated impressions using the values of their ratios combined with criteria to "flag" impressions as potentially misleading. For the FRStat, this criterion is based on a simple threshold of the ratio value selected on the basis of balancing sensitivity and specificity, as discussed in [50]. Under this approach, impressions resulting in a ratio value between 1 and 10 are flagged as potentially misleading. For the AFIS-SLR, this criterion is based on a more sophisticated approach using a machine learning model that accounts for the meta-data output by the AFIS-SLR. Under this approach, the output of the meta-data resulting from the AFIS-SLR measurements are used as inputs into a separate machine learning model trained to predict whether the resultant ratio value from the AFIS-SLR is potentially misleading—irrespective of the magnitude of the ratio value.

The machine learning model used as the basis for flagging impressions evaluated by the AFIS-SLR as potentially misleading was developed for purposes of this exploratory evaluation and is not currently a pre-existing component of the AFIS-SLR method. As such, to explore the utility of machine learning for this purpose and the extent to which performance of the AFIS-SLR could improve with such criterion, multiple machine learning classifiers were initially developed using the *caret* package in R [62] using a range of machine learning techniques (naïve based classifier, tree-based classifiers, discriminant analysis techniques, neural networks and support vector machines) called directly from the *caret* package, specifically: Linear Discriminant Analysis (LDA), Logistical Regression (LogReg), Classification and Regression Tree (CART), Random Forest (RF), Neural Network (NN), C5.0, Support Vector Machine (SVM), Gradient Boosting Machine (GBM), and XGBoost_Linear (XGB_Linear). The meta-data output by the AFIS-SLR that were used as inputs into the machine learning models included fifteen different predictor values (reduced from an initial set of twenty-seven different predictors to only include those which had a correlation less than 0.7), specifically: number of minutiae (nb_min), between-source skewness (inter-skewness), within-source skewness (intra-skewness), between-source kurtosis (inter-kurtosis), proportion of overlap between the within-source and between-source distributions (overlap), numerator of the LR (num), denominator of the LR (den), within-source distribution cumulative probability of the evidence score (pintra), calibrated LR value (lrCal), evidence score (ev), shearing value of the linear distortion from the thin-plate-spline (TPS) distortion algorithm (shearing), scale value of the linear distortion from the TPS distortion algorithm (scale), within-source distribution shape (intra_shape), bending energy of the thin-plate-spline distortion algorithm (be), and rank position of the mark when compared to the print (r). The machine learning techniques were developed and tested using a 10-fold cross validation against a

random 50/50 training-test split (cross validation only being performed on the training partition and predictions only being performed on the test partition) of the subset of impressions across all datasets resulting in a ratio value from the AFIS-SLR that contradicts the known (or assigned) mating status of the impression (i.e., AFIS-SLR ratio values greater than 1 when tested against impressions from non-mated datasets or ratio values less than 1 when tested against impressions from mated datasets). Among the various machine learning classifiers considered, the C5.0 resulted in the highest performance in terms of its ability to accurately predict misleading AFIS-SLR ratio values and was selected for purposes of comparing the performance of the FRStat and AFIS-SLR when each method is augmented by criterion to flag impressions as potentially misleading. Figures 3-8a and 3-8b illustrate the performance of the C5.0 compared to the other machine learning classifiers considered.

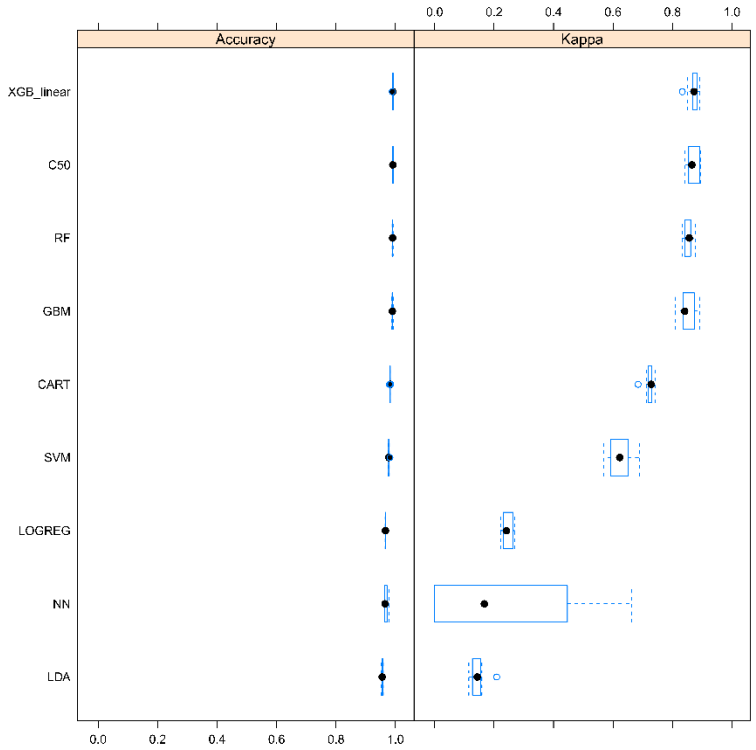


Figure 3-8a. Comparison of the performance of various machine learning classifiers.

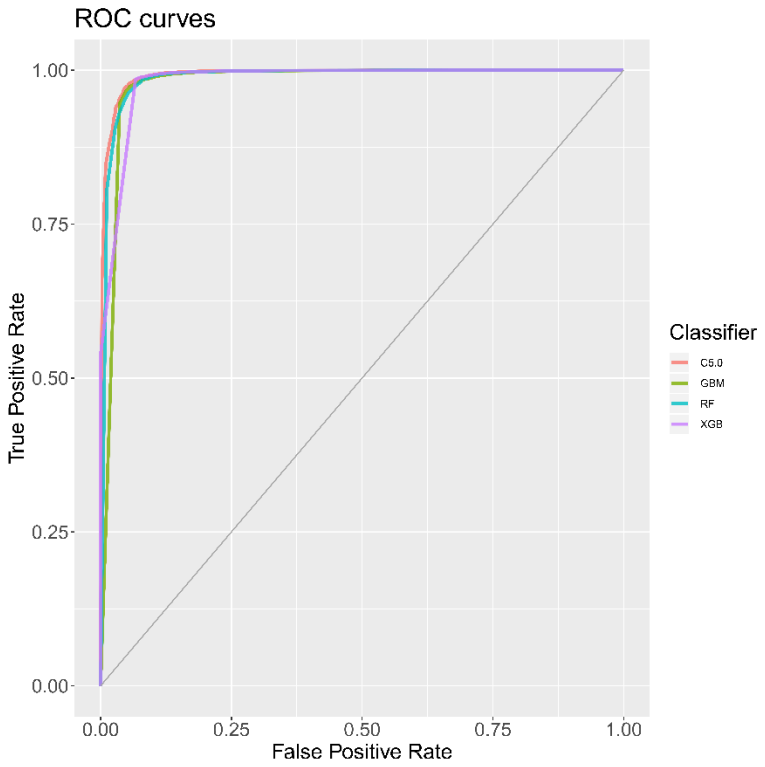


Figure 3-8b. Comparison of receiver operator characteristic (ROC) curves of various machine learning classifiers.

3.2.3 Results & Discussion

Performance with all data

Tables 3-9 through 3-14 provide the baseline performance of the FRStat and AFIS-SLR in terms of their ability to accurately distinguish between mated and non-mated impressions using values of their ratios alone. For both systems, ratio values greater than 1 are predicted as mated impressions and values less than 1 are predicted as non-mated impressions. Tables 3-9 and 3-10 provide the sensitivity of each method (i.e., the ability for each method to accurately predict impressions as mated) and Tables 3-11 through 3-14 provide the specificity of each method (i.e., the ability for each method to accurately predict impressions as non-mated) for the different mated and non-mated datasets.

Sensitivity – Mated Empirical Dataset				
	FRStat		**AFIS-SLR**	
Feature Quantity	*Number of configurations*	*Sensitivity (Ratio >1)*	*Number of configurations*	*Sensitivity (SLR >1)*
5	1,996	0.675	1,996	0.952
6	1,996	0.686	1,996	0.977
7	1,996	0.807	1,996	0.988
8	1,996	0.885	1,996	0.994
9	1,996	0.962	1,996	0.997
10	1,996	0.970	1,996	0.997
11	1,996	0.979	1,996	0.998
12	1,996	0.985	1,996	0.998
13	1,996	0.985	1,996	0.999
14	1,996	0.989	1,996	1.000
15	499	0.990	499	1.000

Table 3-9. Sensitivity of FRStat and AFIS-SLR using the Mated Empirical Dataset for each quantity of features (ranging from 5 to 15).

Sensitivity – Mated Test Datasets #1, #2, #3				
	FRStat		**AFIS-SLR**	
Feature Quantity	*Number of configurations*	*Sensitivity (Ratio >1)*	*Number of configurations*	*Sensitivity (SLR >1)*
5	10,593	0.687	10,593	0.924
6	10,385	0.779	10,385	0.953
7	10,015	0.840	10,015	0.972
8	9,491	0.915	9,491	0.981
9	8,762	0.952	8,762	0.989
10	7,923	0.961	7,923	0.994
11	6,877	0.967	6,877	0.995
12	6,012	0.967	6,012	0.995
13	5,036	0.970	5,036	0.996
14	4,106	0.976	4,106	0.999
15	402	0.978	402	1.000

Table 3-10. Sensitivity of FRStat and AFIS-SLR using the Mated Test Dataset #1, Mated Test Dataset #2, and Mated Test Dataset #3 combined for each quantity of features (ranging from 5 to 15).

Specificity – Non-Mated Empirical Dataset				
	FRStat		**AFIS-SLR**	
Feature Quantity	*Number of configurations*	*Specificity (Ratio <1)*	*Number of configurations*	*Specificity (SLR <1)*
5	1,850	0.793	1,850	0.927
6	1,850	0.833	1,850	0.911
7	1,850	0.903	1,850	0.924
8	1,999	0.934	1,999	0.936
9	2,000	0.955	2,000	0.933
10	1,999	0.972	1,999	0.946
11	2,000	0.980	2,000	0.933
12	1,998	0.988	1,998	0.959
13	1,999	0.991	1,999	0.949
14	2,000	0.993	2,000	0.951
15	1,849	0.994	1,849	0.959

Table 3-11. Specificity of FRStat and AFIS-SLR using the Non-Mated Empirical Dataset for each quantity of features (ranging from 5 to 15).

Specificity – Non-Mated Test Dataset #1				
	FRStat		**AFIS-SLR**	
Feature Quantity	*Number of configurations*	*Specificity (Ratio <1)*	*Number of configurations*	*Specificity (SLR <1)*
5	500	0.818	500	0.822
6	500	0.852	500	0.856
7	500	0.900	500	0.864
8	500	0.912	500	0.908
9	500	0.940	500	0.894
10	500	0.970	500	0.906
11	500	0.978	500	0.91
12	500	0.988	500	0.95
13	500	0.988	500	0.946
14	500	0.988	500	0.962
15	500	0.996	500	0.972

Table 3-12. Specificity of FRStat and AFIS-SLR using the Non-Mated Test Dataset #1 for each quantity of features (ranging from 5 to 15).

Specificity – Non-Mated Test Dataset #2a – "close non-match" delta region				
	FRStat		**AFIS-SLR**	
Feature Quantity	*Number of configurations*	*Specificity (Ratio <1)*	*Number of configurations*	*Specificity (SLR <1)*
5	99	0.566	99	0.636
6	99	0.677	99	0.545
7	96	0.688	96	0.531
8	99	0.737	99	0.515
9	99	0.818	99	0.545
10	97	0.804	97	0.536
11	96	0.802	96	0.417
12	98	0.857	98	0.418
13	99	0.899	99	0.444
14	100	0.980	100	0.570
15	100	0.920	100	0.460

Table 3-13. Specificity of FRStat and AFIS-SLR using the Non-Mated Test Dataset #2a – "close non-match" delta region for each quantity of features (ranging from 5 to 15).

Specificity – Non-Mated Test Dataset #2b – "close non-match" core region				
	FRStat		**AFIS-SLR**	
Feature Quantity	*Number of configurations*	*Specificity (Ratio <1)*	*Number of configurations*	*Specificity (SLR <1)*
5	94	0.787	94	0.809
6	96	0.792	96	0.833
7	95	0.884	95	0.737
8	96	0.896	96	0.750
9	95	0.874	95	0.653
10	96	0.969	96	0.760
11	95	0.989	95	0.684
12	97	1.000	97	0.753
13	97	1.000	97	0.763
14	96	1.000	96	0.854
15	95	1.000	95	0.853

Table 3-14. Specificity of FRStat and AFIS-SLR using the Non-Mated Test Dataset #2b – "close non-match" core region for each quantity of features (ranging from 5 to 15).

Performance after application of Supplemental QA criterion

Tables 3-15 through 3-20 provide the performance of the FRStat and AFIS-SLR in terms of their ability to accurately distinguish between mated and non-mated impressions using values of their ratios *after* applying the supplemental QA criterion to "flag" comparisons as potentially misleading. These data represent only those comparisons of feature configurations that were *not* flagged as potentially misleading (i.e., for FRStat, this includes those comparisons that resulted in a ratio value less than 1 *or* greater than or equal to 10, and for AFIS-SLR, this includes those

comparisons that were not flagged by the ML classifier). Tables 3-15 and 3-16 provide the sensitivity of each method and Tables 3-17 through 3-20 provide the specificity of each method for the different mated and non-mated datasets as well as proportion of the total number of configurations that were not flagged by the QA criterion.

Sensitivity – Mated Empirical Dataset						
	FRStat			**AFIS-SLR**		
Feature Quantity	*Number of configurations*	*Proportion of total number of configurations*	*Sensitivity After QA Criterion*	*Number of configurations*	*Proportion of total number of configurations*	*Sensitivity After QA Criterion*
5	917	0.459	0.293	1,777	0.890	1.000
6	996	0.499	0.371	1,838	0.921	1.000
7	1,076	0.539	0.645	1,912	0.958	1.000
8	1,226	0.614	0.812	1,947	0.975	1.000
9	1,764	0.884	0.957	1,967	0.985	1.000
10	1,895	0.949	0.969	1,972	0.988	1.000
11	1,948	0.976	0.979	1,974	0.989	1.000
12	1,984	0.994	0.985	1,978	0.991	1.000
13	1,985	0.994	0.985	1,986	0.995	1.000
14	1,989	0.996	0.989	1,981	0.992	1.000
15	499	1.000	0.990	491	0.984	1.000

Table 3-15. Sensitivity of FRStat and AFIS-SLR using the Mated Empirical Dataset for each quantity of features (ranging from 5 to 15) after applying criterion to "flag" comparisons as potentially misleading (data excludes those flagged as potentially misleading).

Sensitivity – Mated Test Datasets #1, #2, #3						
	FRStat			AFIS-SLR		
Feature Quantity	*Number of configurations*	*Proportion of total number of configurations*	*Sensitivity After QA Criterion*	*Number of configurations*	*Proportion of total number of configurations*	*Sensitivity After QA Criterion*
5	4,827	0.456	0.313	8,446	0.797	1.000
6	5,246	0.505	0.563	8,551	0.823	1.000
7	5,899	0.589	0.729	8,701	0.869	1.000
8	7,022	0.740	0.885	8,525	0.898	1.000
9	8,088	0.923	0.948	8,079	0.922	1.000
10	7,585	0.957	0.960	7,329	0.925	1.000
11	6,731	0.979	0.967	6,452	0.938	1.000
12	5,973	0.994	0.967	5,627	0.936	1.000
13	5,005	0.994	0.970	4,744	0.942	1.000
14	4,078	0.993	0.976	3,894	0.948	1.000
15	399	0.993	0.977	372	0.925	1.000

Table 3-16. Sensitivity of FRStat and AFIS-SLR using the Mated Test Dataset #1, Mated Test Dataset #2, and Mated Test Dataset #3 combined for each quantity of features (ranging from 5 to 15) after applying criterion to "flag" comparisons as potentially misleading (data excludes those flagged as potentially misleading).

Specificity – Non-Mated Empirical Dataset						
	FRStat			AFIS-SLR		
Feature Quantity	*Number of configurations*	*Proportion of total number of configurations*	*Specificity After QA Criterion*	*Number of configurations*	*Proportion of total number of configurations*	*Specificity After QA Criterion*
5	1,521	0.822	0.964	1,227	0.663	0.998
6	1,693	0.915	0.910	1,372	0.742	1.000
7	1,782	0.963	0.938	1,452	0.785	0.999
8	1,965	0.983	0.950	1,648	0.824	1.000
9	1,970	0.985	0.969	1,694	0.847	1.000
10	1,985	0.993	0.979	1,728	0.864	1.000
11	1,988	0.994	0.986	1,742	0.871	1.000
12	1,990	0.996	0.992	1,825	0.913	1.000
13	1,989	0.995	0.996	1,836	0.918	1.000
14	1,992	0.996	0.996	1,860	0.930	1.000
15	1,842	0.996	0.997	1,729	0.935	1.000

Table 3-17. Specificity of FRStat and AFIS-SLR using the Non-Mated Empirical Dataset for each quantity of features (ranging from 5 to 15) after applying criterion to "flag" comparisons as potentially misleading (data excludes those flagged as potentially misleading).

Specificity – Non-Mated Test Dataset #1						
	FRStat			AFIS-SLR		
Feature Quantity	*Number of configurations*	*Proportion of total number of configurations*	*Specificity After QA Criterion*	*Number of configurations*	*Proportion of total number of configurations*	*Specificity After QA Criterion*
5	409	0.818	1.000	175	0.350	0.863
6	430	0.860	0.991	235	0.470	0.923
7	453	0.906	0.993	236	0.472	0.975
8	463	0.926	0.985	322	0.644	1.000
9	494	0.988	0.951	348	0.696	1.000
10	497	0.994	0.976	386	0.772	1.000
11	498	0.996	0.982	416	0.832	1.000
12	498	0.996	0.992	448	0.896	1.000
13	497	0.994	0.994	460	0.920	1.000
14	498	0.996	0.992	470	0.940	1.000
15	498	0.996	1.000	471	0.942	1.000

Table 3-18. Specificity of FRStat and AFIS-SLR using the Non-Mated Test Dataset #1 for each quantity of features (ranging from 5 to 15) after applying criterion to "flag" comparisons as potentially misleading (data excludes those flagged as potentially misleading).

Specificity – Non-Mated Test Dataset #2a – "close non-match" delta region						
	FRStat			AFIS-SLR		
Feature Quantity	*Number of configurations*	*Proportion of total number of configurations*	*Specificity After QA Criterion*	*Number of configurations*	*Proportion of total number of configurations*	*Specificity After QA Criterion*
5	77	0.778	0.727	34	0.343	0.941
6	92	0.929	0.728	35	0.354	0.829
7	93	0.969	0.710	41	0.427	0.902
8	94	0.949	0.777	36	0.364	0.889
9	99	1.000	0.818	43	0.434	0.930
10	95	0.979	0.821	49	0.505	0.918
11	94	0.979	0.819	36	0.375	0.889
12	97	0.990	0.866	39	0.398	0.949
13	96	0.970	0.927	41	0.414	0.951
14	99	0.990	0.990	52	0.520	1.000
15	100	1.000	0.920	46	0.460	1.000

Table 3-19. Specificity of FRStat and AFIS-SLR using the Non-Mated Test Dataset #2a – "close non-match" delta region for each quantity of features (ranging from 5 to 15) after applying criterion to "flag" comparison as potentially misleading (data excludes those flagged as potentially misleading).

Specificity – Non-Mated Test Dataset #2b – "close non-match" core region						
	FRStat			AFIS-SLR		
Feature Quantity	*Number of configurations*	*Proportion of total number of configurations*	*Specificity After QA Criterion*	*Number of configurations*	*Proportion of total number of configurations*	*Specificity After QA Criterion*
5	76	0.809	0.974	31	0.330	0.871
6	83	0.865	0.916	37	0.385	0.919
7	91	0.958	0.923	36	0.379	0.917
8	92	0.958	0.935	37	0.385	1.000
9	92	0.968	0.902	35	0.368	0.943
10	94	0.979	0.989	61	0.635	0.967
11	95	1.000	0.989	56	0.589	1.000
12	97	1.000	1.000	66	0.680	1.000
13	97	1.000	1.000	64	0.660	1.000
14	96	1.000	1.000	79	0.823	1.000
15	95	1.000	1.000	78	0.821	1.000

Table 3-20. Specificity of FRStat and AFIS-SLR using the Non-Mated Test Dataset #2b – "close non-match" core region for each quantity of features (ranging from 5 to 15) after applying criterion to "flag" comparisons as potentially misleading (data excludes those flagged as potentially misleading).

Figures 3-9 through 3-14 illustrate the impacts to the performance of the FRStat and AFIS-SLR as a result of applying the supplemental QA criterion to "flag" comparisons as potentially misleading (i.e., graphically illustrating the sensitivity and specificity values listed in Tables 3-9 through 3-20). Figures 3-9 and 3-10 compare the sensitivity of each method before and after applying the QA criterion for the mated datasets and Figures 3-11 through 3-14 compare the specificity of each method before and after applying the QA criterion for the non-mated datasets.

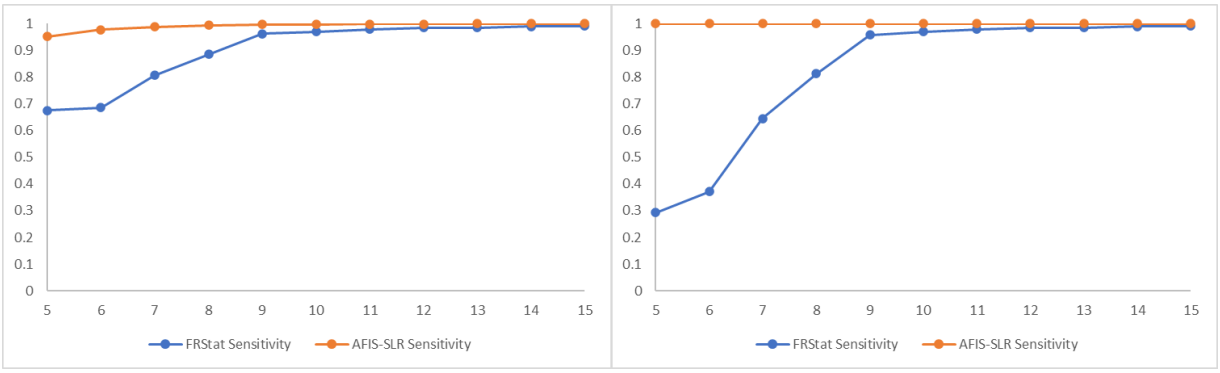


Figure 3-9. Comparison of the sensitivity of FRStat and AFIS-SLR using the Mated Empirical Dataset for each quantity of features (ranging from 5 to 15) before (left) and after (right) applying criterion to "flag" comparisons as potentially misleading (data excludes those flagged as potentially misleading).

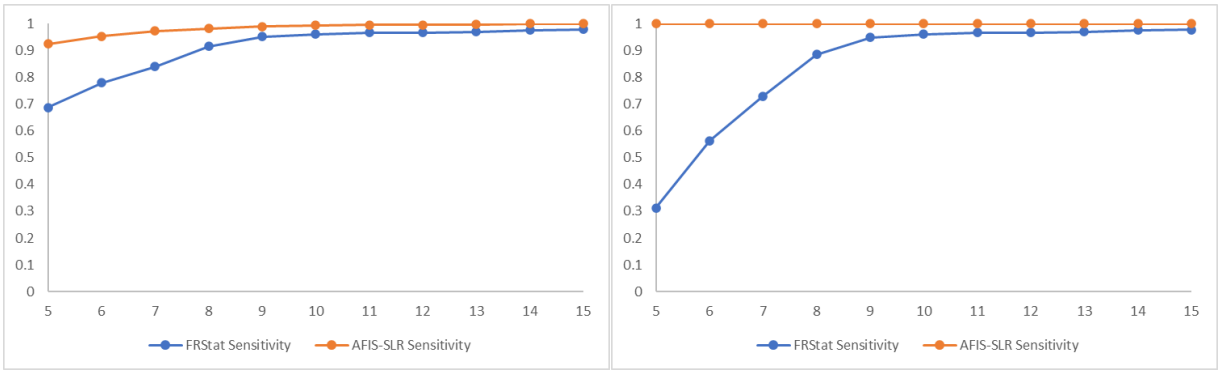


Figure 3-10. Comparison of the sensitivity of FRStat and AFIS-SLR using the using the Mated Test Dataset #1, Mated Test Dataset #2, and Mated Test Dataset #3 combined for each quantity of features (ranging from 5 to 15) before (left) and after (right) applying criterion to "flag" comparisons as potentially misleading (data excludes those flagged as potentially misleading).

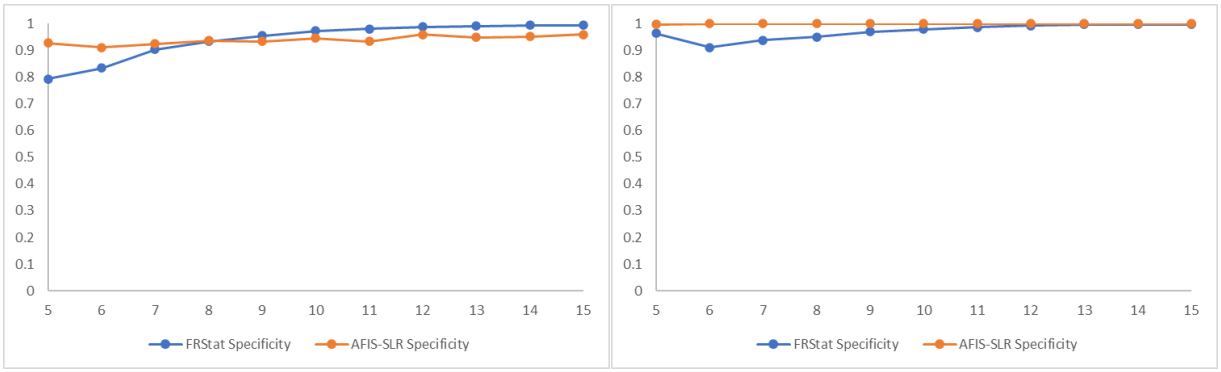


Figure 3-11. Comparison of the specificity of FRStat and AFIS-SLR using the using the Non-Mated Empirical Dataset for each quantity of features (ranging from 5 to 15) before (left) and after (right) applying criterion to "flag" comparisons as potentially misleading (data excludes those flagged as potentially misleading).

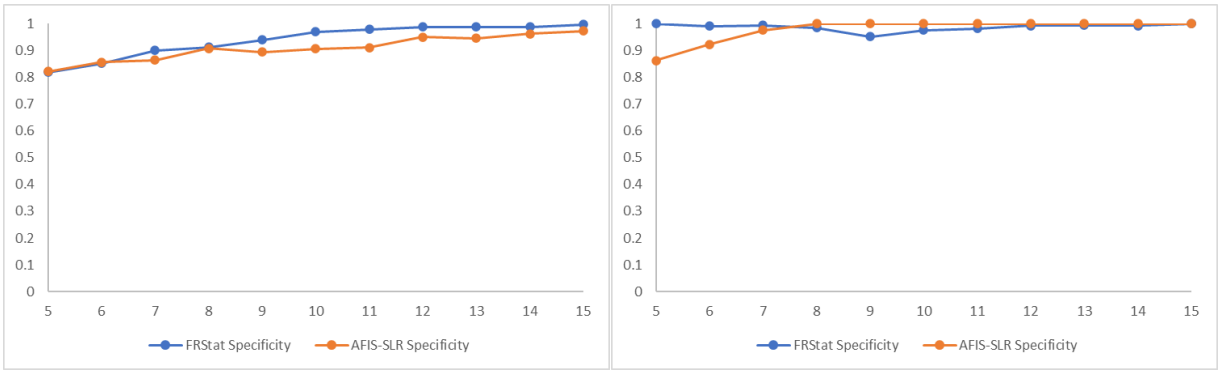


Figure 3-12. Comparison of the specificity of FRStat and AFIS-SLR using the using the Non-Mated Test Dataset #1 for each quantity of features (ranging from 5 to 15) before (left) and after (right) applying criterion to "flag" comparisons as potentially misleading (data excludes those flagged as potentially misleading).

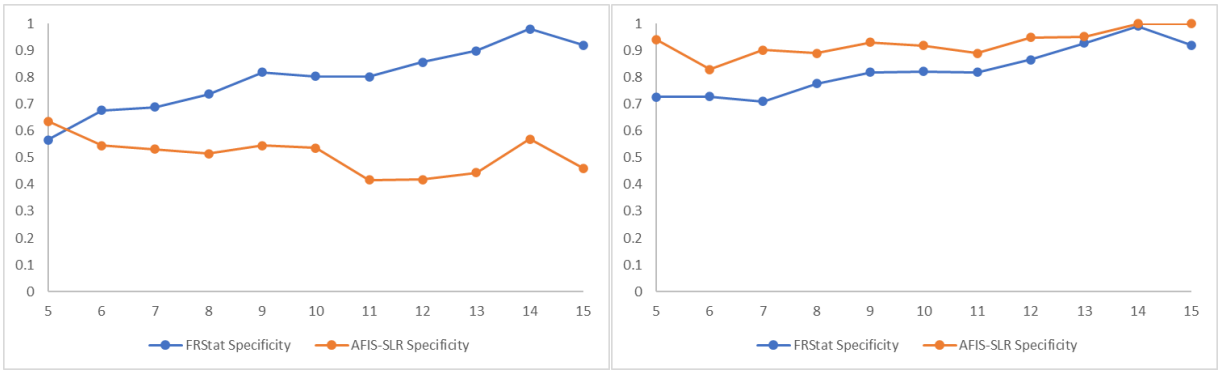


Figure 3-13. Comparison of the specificity of FRStat and AFIS-SLR using the Non-Mated Test Dataset #2a – "close non-match" delta region for each quantity of features (ranging from 5 to 15) before (left) and after (right) applying criterion to "flag" comparisons as potentially misleading (data excludes those flagged as potentially misleading).

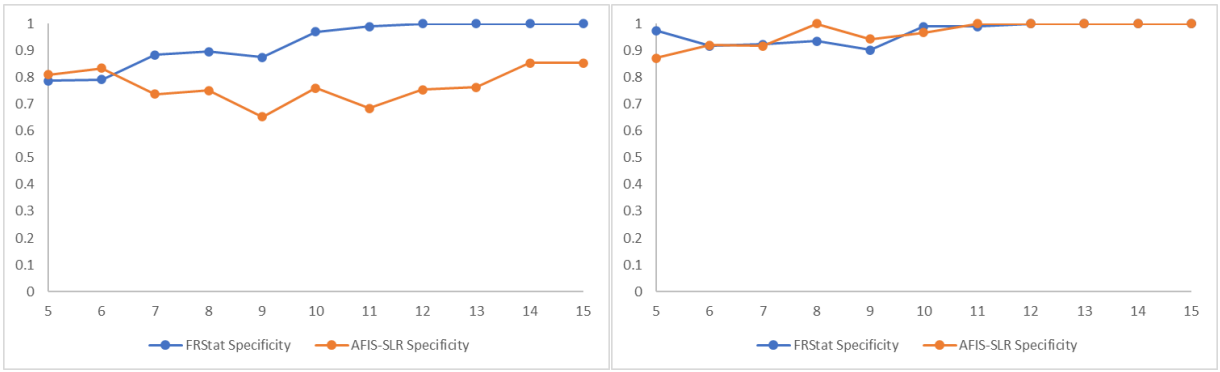


Figure 3-14. Comparison of the specificity of FRStat and AFIS-SLR using the Non-Mated Test Dataset #2b – "close non-match" core region for each quantity of features (ranging from 5 to 15) before (left) and after (right) applying criterion to "flag" comparisons as potentially misleading (data excludes those flagged as potentially misleading).

Tables 3-21 through 3-26 and Figures 3-15 through 3-20 illustrate the utility of the system (before and after the application of the supplemental QA criterion) as a function of both its performance (i.e., sensitivity, specificity) and the proportion of comparisons that were *not* flagged as potentially misleading. Tables 3-21 and 3-22 as well as Figures 3-15 and 3-16 demonstrate the utility of each system as it relates to their sensitivity (calculated by multiplying the values of sensitivity and the proportion of the total number of configurations not flagged by the supplemental QA criterion) and Tables 3-23 through 3-26 as well as Figures 3-17 through 3-20 demonstrate the utility of each system as it relates to their specificity (calculated by multiplying the values of specificity and the proportion of the total number of configurations not flagged by the supplemental QA criterion). The utility values listed in Tables 3-21 through 3-26 can be calculated directly from the data provided in Tables 3-9 through 3-20. Additionally, the utility values before the application of the supplemental QA criterion are equal to the performance of the system (i.e., sensitivity, specificity) since none of the comparisons were flagged by the supplemental QA criterion (i.e., the proportion of the total number of configurations *not* flagged by the supplemental QA criterion is 1.000).

Utility (Sensitivity) – Mated Empirical Dataset				
	FRStat		**AFIS-SLR**	
Feature Quantity	*Utility (before application of Supplemental QA criterion)*	*Utility (after application of Supplemental QA criterion)*	*Utility (before application of Supplemental QA criterion)*	*Utility (after application of Supplemental QA criterion)*
5	0.675	0.134	0.952	0.890
6	0.686	0.185	0.977	0.921
7	0.807	0.348	0.988	0.958
8	0.885	0.499	0.994	0.975
9	0.962	0.846	0.997	0.985
10	0.970	0.920	0.997	0.988
11	0.979	0.956	0.998	0.989
12	0.985	0.979	0.998	0.991
13	0.985	0.979	0.999	0.995
14	0.989	0.985	1.000	0.992
15	0.990	0.990	1.000	0.984

Table 3-21. Utility of FRStat and AFIS-SLR in terms of sensitivity and proportion of comparisons not flagged by supplemental QA criterion using the Mated Empirical Dataset for each quantity of features (ranging from 5 to 15).

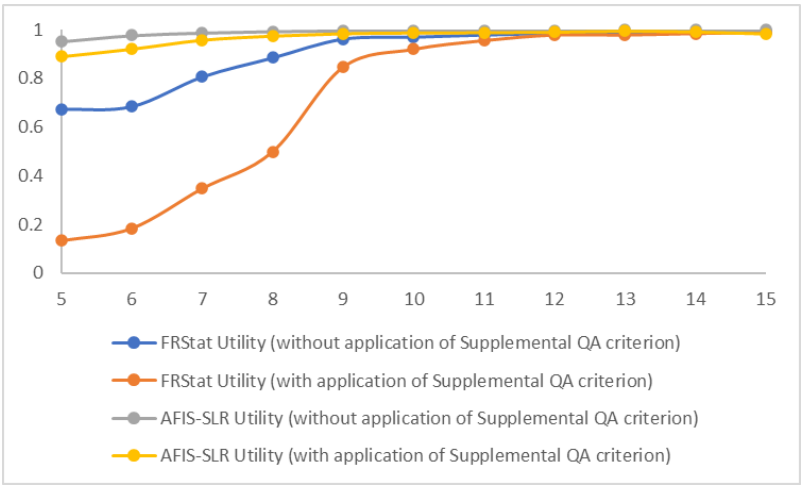


Figure 3-15. Utility of FRStat and AFIS-SLR in terms of sensitivity and proportion of comparisons not flagged by supplemental QA criterion using the Mated Empirical Dataset for each quantity of features (ranging from 5 to 15) before and after applying criterion to "flag" comparisons as potentially misleading (data excludes those flagged as potentially misleading).

Utility (Sensitivity) – Mated Test Datasets #1, #2, #3				
	FRStat		AFIS-SLR	
Feature Quantity	*Utility (before application of Supplemental QA criterion)*	*Utility (after application of Supplemental QA criterion)*	*Utility (before application of Supplemental QA criterion)*	*Utility (after application of Supplemental QA criterion)*
5	0.687	0.143	0.924	0.797
6	0.779	0.284	0.953	0.823
7	0.840	0.429	0.972	0.869
8	0.915	0.655	0.981	0.898
9	0.952	0.875	0.989	0.922
10	0.961	0.919	0.994	0.925
11	0.967	0.947	0.995	0.938
12	0.967	0.961	0.995	0.936
13	0.970	0.964	0.996	0.942
14	0.976	0.969	0.999	0.948
15	0.978	0.970	1.000	0.925

Table 3-22. Utility of FRStat and AFIS-SLR in terms of sensitivity and proportion of comparisons not flagged by supplemental QA criterion using the Mated Test Dataset #1, Mated Test Dataset #2, and Mated Test Dataset #3 combined for each quantity of features (ranging from 5 to 15).

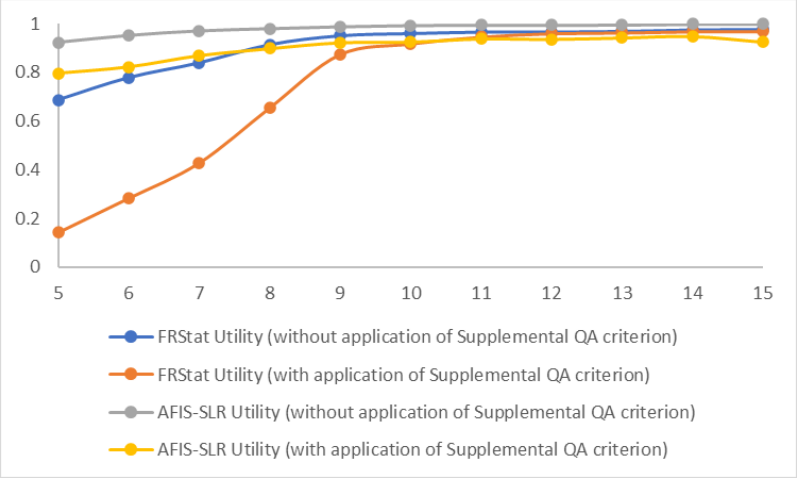


Figure 3-16. Utility of FRStat and AFIS-SLR in terms of sensitivity and proportion of comparisons not flagged by supplemental QA criterion using the using the Mated Test Dataset #1, Mated Test Dataset #2, and Mated Test Dataset #3 combined for each quantity of features (ranging from 5 to 15) before and after applying criterion to "flag" comparisons as potentially misleading (data excludes those flagged as potentially misleading).

Utility (Specificity) – Non-Mated Empirical Dataset				
	FRStat		**AFIS-SLR**	
Feature Quantity	*Utility (before application of Supplemental QA criterion)*	*Utility (after application of Supplemental QA criterion)*	*Utility (before application of Supplemental QA criterion)*	*Utility (after application of Supplemental QA criterion)*
5	0.793	0.792	0.927	0.662
6	0.833	0.833	0.911	0.742
7	0.903	0.903	0.924	0.784
8	0.934	0.934	0.936	0.824
9	0.955	0.954	0.933	0.847
10	0.972	0.972	0.946	0.864
11	0.980	0.980	0.933	0.871
12	0.988	0.988	0.959	0.913
13	0.991	0.991	0.949	0.918
14	0.993	0.992	0.951	0.930
15	0.994	0.993	0.959	0.935

Table 3-23. Utility of FRStat and AFIS-SLR in terms of specificity and proportion of comparisons not flagged by supplemental QA criterion using the Non-Mated Empirical Dataset for each quantity of features (ranging from 5 to 15).

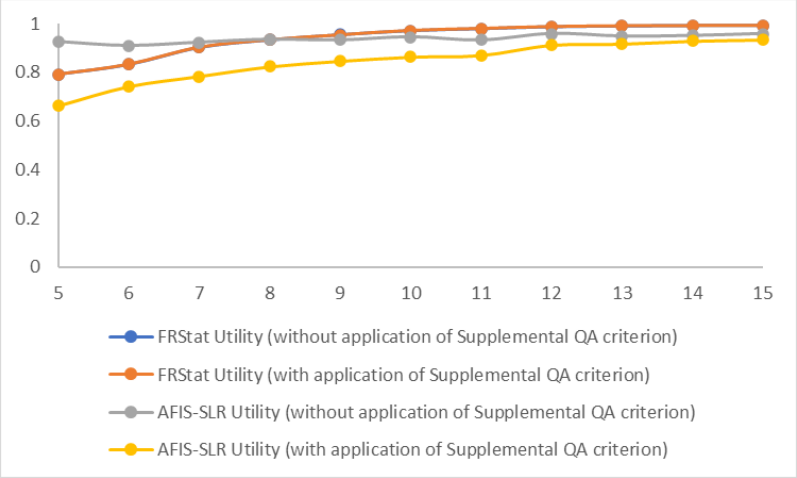


Figure 3-17. Utility of FRStat and AFIS-SLR in terms of specificity and proportion of comparisons not flagged by supplemental QA criterion using the using the Non-Mated Empirical Dataset for each quantity of features (ranging from 5 to 15) before and after applying criterion to "flag" comparisons as potentially misleading (data excludes those flagged as potentially misleading).

Utility (Specificity) – Non-Mated Test Dataset #1				
	FRStat		**AFIS-SLR**	
Feature Quantity	*Utility (before application of Supplemental QA criterion)*	*Utility (after application of Supplemental QA criterion)*	*Utility (before application of Supplemental QA criterion)*	*Utility (after application of Supplemental QA criterion)*
5	0.818	0.818	0.822	0.302
6	0.852	0.852	0.856	0.434
7	0.900	0.900	0.864	0.460
8	0.912	0.912	0.908	0.644
9	0.940	0.940	0.894	0.696
10	0.970	0.970	0.906	0.772
11	0.978	0.978	0.910	0.832
12	0.988	0.988	0.950	0.896
13	0.988	0.988	0.946	0.920
14	0.988	0.988	0.962	0.940
15	0.996	0.996	0.972	0.942

Table 3-24. Utility of FRStat and AFIS-SLR in terms of specificity and proportion of comparisons not flagged by supplemental QA criterion using the Non-Mated Test Dataset #1 for each quantity of features (ranging from 5 to 15).

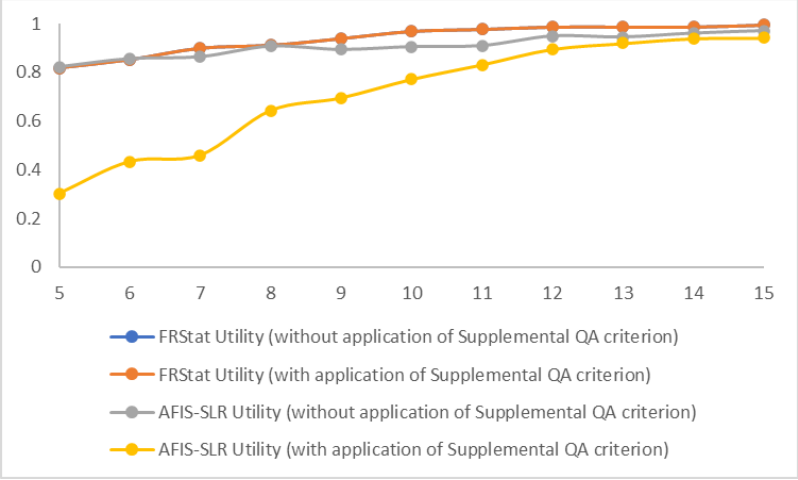


Figure 3-18. Utility of FRStat and AFIS-SLR in terms of specificity and proportion of comparisons not flagged by supplemental QA criterion using the using the Non-Mated Test Dataset #1 for each quantity of features (ranging from 5 to 15) before and after applying criterion to "flag" comparisons as potentially misleading (data excludes those flagged as potentially misleading).

Utility (Specificity) – Non-Mated Test Dataset #2a – "close non-match" delta region				
	FRStat		AFIS-SLR	
Feature Quantity	*Utility (before application of Supplemental QA criterion)*	*Utility (after application of Supplemental QA criterion)*	*Utility (before application of Supplemental QA criterion)*	*Utility (after application of Supplemental QA criterion)*
5	0.566	0.566	0.636	0.323
6	0.677	0.676	0.545	0.293
7	0.688	0.688	0.531	0.385
8	0.737	0.737	0.515	0.324
9	0.818	0.818	0.545	0.404
10	0.804	0.804	0.536	0.464
11	0.802	0.802	0.417	0.333
12	0.857	0.857	0.418	0.378
13	0.899	0.899	0.444	0.394
14	0.980	0.980	0.570	0.520
15	0.920	0.920	0.460	0.460

Table 3-25. Utility of FRStat and AFIS-SLR in terms of specificity and proportion of comparisons not flagged by supplemental QA criterion using the Non-Mated Test Dataset #2a – "close non-match" delta region for each quantity of features (ranging from 5 to 15).

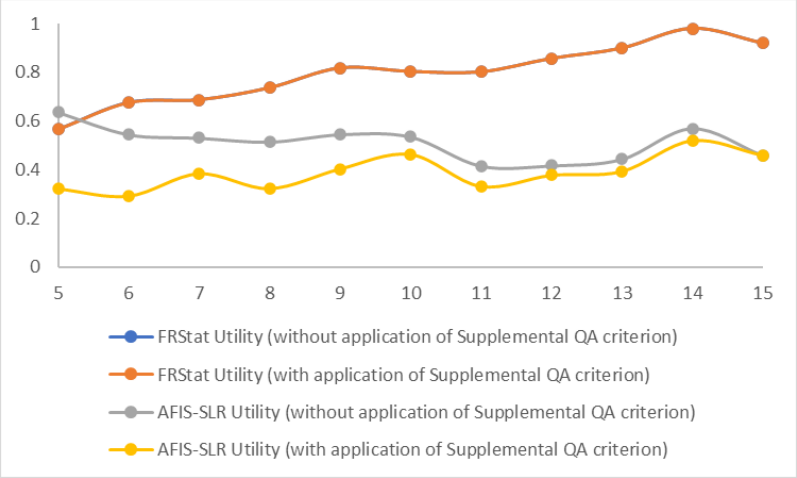


Figure 3-19. Utility of FRStat and AFIS-SLR in terms of specificity and proportion of comparisons not flagged by supplemental QA criterion using the Non-Mated Test Dataset #2a – "close non-match" delta region for each quantity of features (ranging from 5 to 15) before and after applying criterion to "flag" comparisons as potentially misleading (data excludes those flagged as potentially misleading).

Utility (Specificity) – Non-Mated Test Dataset #2b – "close non-match" core region				
	FRStat		AFIS-SLR	
Feature Quantity	*Utility (before application of Supplemental QA criterion)*	*Utility (after application of Supplemental QA criterion)*	*Utility (before application of Supplemental QA criterion)*	*Utility (after application of Supplemental QA criterion)*
5	0.787	0.788	0.809	0.287
6	0.792	0.792	0.833	0.354
7	0.884	0.884	0.737	0.348
8	0.896	0.896	0.75	0.385
9	0.874	0.873	0.653	0.347
10	0.969	0.968	0.76	0.614
11	0.989	0.989	0.684	0.589
12	1.000	1.000	0.753	0.680
13	1.000	1.000	0.763	0.660
14	1.000	1.000	0.854	0.823
15	1.000	1.000	0.853	0.821

Table 3-26. Utility of FRStat and AFIS-SLR in terms of specificity and proportion of comparisons not flagged by supplemental QA criterion using the Non-Mated Test Dataset #2b – "close non-match" core region for each quantity of features (ranging from 5 to 15).

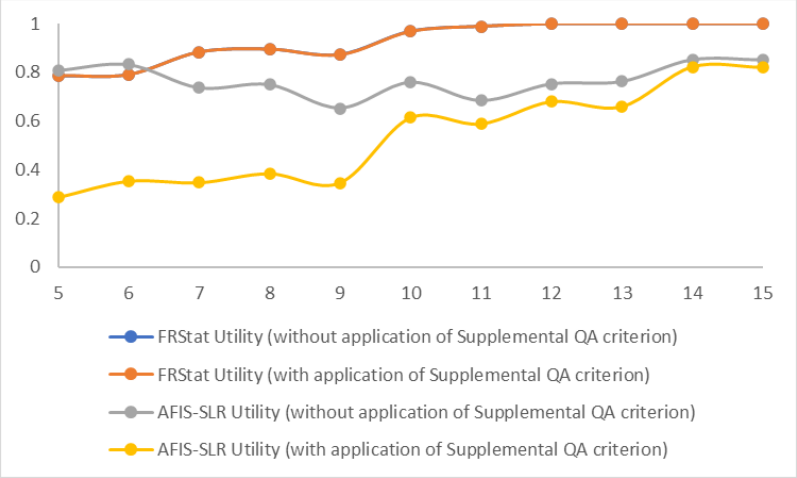


Figure 3-20. Utility of FRStat and AFIS-SLR in terms of specificity and proportion of comparisons not flagged by supplemental QA criterion using the Non-Mated Test Dataset #2b – "close non-match" core region for each quantity of features (ranging from 5 to 15) before and after applying criterion to "flag" comparisons as potentially misleading (data excludes those flagged as potentially misleading).

General Discussion

There are three key observations we can make based on these data presented in Tables 3-9 through 3-26 and Figures 3-9 through 3-20 as it relates to the performance of each system and the overall utility of each system (before and after the application of the supplemental QA criterion).

First, in terms of the performance of each system to accurately distinguish between mated and non-mated impressions on the basis of its ratio value alone (without any consideration of supplemental QA criterion), the AFIS-SLR demonstrated superior performance in terms of its *sensitivity* across all feature quantities (ranging from 5 to 15) for both sets of mated datasets (Mated Empirical Dataset and Mated Test Datasets 1, 2, and 3). When the quantity of features range between 5 and 8, the AFIS-SLR was far superior; however, once the quantity of features reached 9 or greater, the AFIS-SLR was only slightly better (e.g., approximately 3.5% difference or less). As the quantity of features increased, the sensitivity values also increased and the difference in the sensitivity values between the two systems decreased. These results of increasing sensitivity as the quantity of features increase align with our expectations of how the systems should behave. On the other hand, both the FRStat and AFIS-SLR demonstrated comparable performance in terms of its *specificity* for the Non-Mated Empirical Dataset and Non-Mated Test Dataset 1 (both datasets involving non-mated impressions arbitrarily paired together). Interestingly, the FRStat demonstrated superior performance in terms of its specificity for both Non-Mated Test Datasets 2a and 2b (close non-match "delta" and "core" regions, respectively). With minor exceptions between 5 and 6 features, these results were generally consistent across all other feature quantities, and the difference in specificity values between the two systems was quite large.

The second key observation is related to the impact on the performance of each system as a result of applying the supplemental QA criterion to "flag" comparisons for which the ratio value could be potentially misleading. In an operational context, those comparisons that are "flagged" would be subject to enhanced quality assurance review(s) prior to a result being finalized. Thus, those that are "flagged" are essentially considered "inconclusive" by the FRStat or AFIS-SLR, respectively, and therefore the output is not considered as part of this evaluation. To recall, for the FRStat, this evaluation includes those comparisons that resulted in a ratio value less than 1 *or* greater than or equal to 10, and for AFIS-SLR, this evaluation includes those comparisons that were not flagged by the ML classifier. Taking into account the performance of each system among those comparisons that were *not* flagged, we see the AFIS-SLR demonstrated perfect *sensitivity* across all quantities of features (sensitivity values of 1.00), illustrating that the supplemental QA criterion had a positive impact on the sensitivity of the AFIS-SLR; however, the sensitivity of the FRStat decreased substantially when the quantity of features ranged between 5 and 8 (compared to the sensitivity values without the supplemental QA criterion). Once the quantity of features reached 9 or greater, the supplemental QA criterion had minimal impacts to the sensitivity of the FRStat. On the other hand, the supplemental QA criterion had positive impacts on the *specificity* of both the FRStat and the AFIS-SLR across all non-mated datasets, and both the FRStat and AFIS-SLR demonstrated comparable performance in terms of its specificity for the Non-Mated Empirical Dataset and Non-Mated Test Dataset 1 (both datasets involving non-mated impressions arbitrarily paired together). For the Non-Mated Test Datasets 2a and 2b (close non-match "delta" and "core" regions, respectively), the AFIS-SLR demonstrated substantial improvement with the application of the supplemental QA criterion. The AFIS-SLR outperformed the FRStat across all quantities of features for Non-Mated Test Dataset 2a, and the results were generally comparable between the FRStat and AFIS-SLR for Non-Mated Test Dataset 2b.

The third key observation is related to the utility of each system (before or after applying the supplemental QA criterion). We can consider the utility of the system as a function of both its performance (i.e., sensitivity, specificity) and the proportion of comparisons that were *not* flagged

as potentially misleading (i.e., not flagged as "inconclusive" by the system). For example, we can propose a system that demonstrates high performance in terms of its sensitivity and specificity, but also flags a high proportion of comparisons as potentially misleading when the supplemental QA criterion is applied. In this example, although the performance of the system among those comparisons that were not flagged is high, the system is not very effective at providing a meaningful result for a high proportion of the comparisons; thus, the system would have moderate or low overall utility in an operational context. Compare that to a system that demonstrates high performance in terms of its sensitivity and specificity while flagging few, if any, of the comparisons as potentially misleading. In this example, not only is the performance of the system among those comparisons that were not flagged high, but the system is also very effective at providing a meaningful result for the vast majority of the comparisons; thus, the system would have high overall utility in an operational context. As illustrated by these examples, and for purposes of this evaluation, the utility of each system is calculated as the product of the performance (sensitivity, specificity) and the proportion of comparisons that were *not* flagged as potentially misleading. From the utility values represented in Tables 3-21 through 3-26 and illustrated by Figures 3-15 through 3-20, we see the AFIS-SLR (without the application of the supplemental QA criterion) provides superior utility in terms of *sensitivity* across all quantities of features. In terms of *specificity*, we see that the AFIS-SLR (without the application of the supplemental QA criterion) and the FRStat (with or without the application of the supplemental QA criterion) provide comparable performance for the Non-Mated Empirical Dataset and Non-Mated Test Dataset 1 (both datasets involving non-mated impressions arbitrarily paired together). For the Non-Mated Test Datasets 2a and 2b (close non-match "delta" and "core" regions, respectively), however, we see that the FRStat (with or without the application of the supplemental QA criterion) provides superior utility.

From the discussion above, we can draw three overall conclusions. First, the supplemental QA criterion had a negative impact to the overall performance of the FRStat (significant impacts to sensitivity and negligible impacts to specificity), and, although the application of the supplemental QA criterion improved performance of the AFIS-SLR, it did so at a significant cost of removing a large proportion of the dataset. Consequently, in terms of overall utility, neither system appears to benefit from the application of the supplemental QA criterion (i.e., the ratio value alone appears to be a sufficient basis for distinguishing between mated and non-mated comparisons). Second, the AFIS-SLR is superior compared to the FRStat in terms of maximizing both sensitivity and specificity across all quantities of features (except for "close non-match" comparisons); however, the greatest impact of this comes from the improved sensitivity of the AFIS-SLR for comparisons having less than 9 features. For comparisons with 9 features or more, both the AFIS-SLR and the FRStat demonstrated comparable performance for both sensitivity and specificity (except for "close non-match" comparisons). Third, the FRStat demonstrated significantly improved performance in terms of its specificity against "close non-match" comparisons. These results suggest that both the AFIS-SLR and the FRStat are viable systems to use as a means of distinguishing between mated and non-mated impressions (with the AFIS-SLR having a notable edge for impressions with less than 9 features and the FRStat having a notable edge for distinguishing between "close non-match" comparisons). Overall, these results suggest that there is some tradeoff of terms of performance and computational complexity, but not enough to warrant a wholesale rejection of one system over the other—both systems appear to have the

capacity to distinguish between mated and non-mated impressions with reasonable accuracy and provide an additional layer of quality control to the overall examination scheme.

4 Toward Objectivity: Integrating Algorithmic Outputs

This chapter explores the utility (i.e., usefulness), from a quality management standpoint, of integrating the DFIQI and FRStat algorithms into a single system for which the input to the FRStat is dependent upon the output from the DFIQI. An integrated system such as this could provide a more objective and semi-automated approach for ensuring not only that analysts' interpretations are empirically supported for all major decisions throughout the examination methodology but also that a means for monitoring and ensuring the quality of results meets minimum standards for quality assurance. This chapter describes how the two systems can be integrated and evaluates the impacts of such an application in practice.

4.1 Background

The DFIQI algorithm presented in chapter 2 (i.e., in [49]) provides measure of the overall quality of the impression for further examination purposes as well as an objective measure of the clarity of friction ridge features identified by the analyst. Although a primary utility of the DFIQI algorithm is the use of its global assessment of quality for purposes of flagging an impression for additional quality assurance review prior to further examination with an exemplar, the clarity of the friction ridges immediately surrounding each individual feature is measured as part of the algorithm and those results are included as an output to the user in the form of color-coded bins (green, yellow, or red) indicating whether the clarity of ridges in those areas is "high," "medium," or "low," respectively. The clarity of ridge detail immediately surrounding individual features can be used as a proxy for their reliability—features in high clarity areas are generally interpreted more consistently by analysts and are therefore considered more reliable whereas features in low clarity areas are generally interpreted less consistently by analysts and are therefore considered less reliable. As such, the DFIQI algorithm can be applied to a mark as a means of (1) measuring the quality of the overall impression for further examination, and (2) objectively assessing the reliability of the features identified by the analyst which are intended to be used in further examination with the exemplar. As a stand-alone algorithm, the DFIQI helps address an important need for the discipline; however, it is not a comprehensive solution. The DFIQI does provide a means of assessing the significance of an association between a mark-exemplar pair.

The FRStat algorithm presented in chapter 3 (i.e., in [50]) provides a means of quantitatively conveying the significance of an association observed by an analyst. As such, the FRStat provides an empirical foundation to analysts' opinions of association between a mark-exemplar pair as well as a tool for quality assurance managers to ensure the reported results meet minimum standards. Although the FRStat helps address a critical gap in the discipline, it has its limitations. The most notable limitation is that the FRStat is only able to consider the annotations by the analyst representing the locations and angles of the features and is not able to evaluate the reliability of the actual features. Consequently, without proper protocols in place to guard against potential cognitive biases related to the interpretation of features, it is foreseeable that the input to the FRStat could include annotations of low quality and unreliable features that are incorrectly represented. If the incorrect interpretation leads to a misalignment of similarity, then the FRStat is likely to produce a low similarity result, which would flag it for further quality assurance review. However, if the incorrect interpretation leads to a false alignment of similarity, then the FRStat is unlikely to flag it, which could lead to a more consequential outcome if not detected by other means available

in the quality system (e.g., verification, technical review, etc.). To address this, the FRStat should not be used in isolation—it should be used in accordance with a set of strict policies and procedures to verify the reliability of the annotated features. Different strategies can be employed to accomplish this, such as linear applications of ACE (i.e., only those features which were documented on the mark prior to exposure to the exemplar and which were not adjusted after exposure to the exemplar are included), consensus feature markups (i.e., only those features which were independently observed and documented on the mark by multiple examiners are included), or quality metrics (i.e., only those features measured by an algorithm and exceed a minimum standard for quality are included). The DFIQI algorithm presented in chapter 2 is an example of an algorithm to enable the latter approach. As a combined system, the integration of DFIQI and FRStat outputs could be a valuable tool for improving the objectivity, transparency, and standardization of the examination process.

Integrating the DFIQI and FRStat algorithmic outputs can be accomplished through governance of policy and procedure or through technology by consolidating the two programs into a single software application and creating a bridge between the two algorithms. Given the construct of the two software applications, both options are possible—each with their own pros and cons:

Governance by policy and procedure requires more effort in terms and workflow and would require discretizing the clarity of the features measured by DFIQI (reducing the clarity measures to one of three classes). Although this option is less automated, it offers the laboratory more flexibility and discretion if the analyst disagreed with the output of the algorithm midway through the examination process. In those situations, the laboratory could flag the impression for more in-depth quality assurance review and arbitrate the disagreement at that stage in the process prior to applying the FRStat algorithm. This option offers more precision when identifying and addressing the cause of the disagreement.

Governance by technology allows for either: (a) the clarity of features measured by DFIQI to be treated in a discretized manner and features input to the FRStat are automatically filtered to those which meet the minimum quality threshold (much like governance by policy and procedure but in an automated fashion), or (b) the clarity of features to be treated on a continuous scale such that all features are input into the FRStat, but the contribution of each feature to the FRStat algorithm is automatically weighted based on the clarity measured (as a continuous value) by the DFIQI algorithm. The former could be accomplished relatively easily without modification to the substantive elements of the FRStat algorithm. The latter would require substantive modifications to certain elements of the FRStat algorithm and is therefore presented here as a theoretical possibility but left for future work. This option offers the laboratory greater automation, but less flexibility and discretion if the analyst disagreed with the output of the DFIQI algorithm midway through the process versus waiting until the end after both algorithms have been applied.

In the sections that follow, the impact of integrating the two algorithms is evaluated by treating the clarity of features in a discretized manner and only permitting those features which are color coded green or yellow as eligible for entry into the FRStat. This approach simulates a form of integration that could be governed by either policy and procedure or technology and therefore represents the most likely scenario in which the two algorithms would be integrated operationally.

4.2 Materials & Methods

Although it is possible to integrate the two algorithms by consolidating the two programs into a single software application and creating a bridge between the two algorithms, for purposes of this evaluation both programs remain as separate software applications. The impact of integrating the two algorithms as a combined system is evaluated by running a dataset of impressions through the DFIQI using the global quality measure ($Value_{GQS}$) as an initial gating function followed by FRStat *without* pre-assessment by DFIQI LQS (i.e., all features observed by the analysts are input into FRStat irrespective of their clarity measures) and *with* pre-assessment by DFIQI LQS (i.e., only those features color-coded as green or yellow by DFIQI clarity measures are input into FRStat).

The dataset used for this evaluation consists of 605 marks and corresponding exemplars arbitrarily collected from casework during the course of routine operations by fingerprint experts in a federal crime laboratory in the United States. All impressions in this dataset were initially determined subjectively by analysts to be "suitable" or "of value" for identification purposes and subsequently identified to the exemplars. All impressions collected in this dataset were examined as part of routine casework operations prior to the use of algorithms to augment traditional subjective examination practices. The impressions were collected from a wide variety of cases, substrates, and assigned fingerprint experts. The corresponding features (ranging between 7 and 15) were manually annotated by the assigned fingerprint expert during the initial case examination. The selected features were then annotated later in a format suitable for DFIQI and FRStat analysis by the same fingerprint expert that did the initial examination. For clarity, this dataset is the same as that referenced as "GQS-Dataset-2" in chapter 2 (i.e., [49]) related to the evaluation of DFIQI and "Mated Test Dataset #3 (*believed* to be mated)" in chapter 3 (i.e., [50]) related to the evaluation of FRStat. The utility of this particular dataset for purposes of this evaluation is to consider the impact of the algorithms in terms of the proportion of cases for which the analysts' opinions would be supported by the algorithms versus those which would be flagged for further review when applied to impressions derived from actual casework and assessed under normal operational conditions. It should be noted that this dataset does not include those impressions deemed to be "no value." Operational procedures at the time these impressions were examined did not require retention of annotated images for "no value" outcomes; thus, the impact of algorithms is only considered in terms of the proportion of mark-exemplar pairs that the algorithms supported the analysts' opinion of "value" and subsequent association or flagged for further review. For purposes of this evaluation, results are analyzed using the following thresholds for DFIQI and FRStat discussed in chapters 2 and 3, respectively: DFIQI $Value_{GQS}$ result of 0.50 or greater for analysts' subjective suitability opinions to be supported without further review, DFIQI LQS result of green or yellow (LQS values of 0.20 or greater) for the features to be eligible for entry into FRStat, and a FRStat ratio result of 10 or greater for analysts' opinions of a positive association between the mark and exemplar to be reported without further review.

4.3 Results & Discussion

Among the 605 mark-exemplar pairs, 591 (98%) produced a DFIQI $Value_{GQS}$ result of 0.50 or greater thereby supporting the analysts' initial opinion of "value" and determination to proceed to further examination. Thus, at the outset, the DFIQI $Value_{GQS}$ would have flagged 14 (2%) of the marks as warranting further review from a quality assurance perspective prior to further examination. Of the 591 that were supported by the DFIQI $Value_{GQS}$ pre-assessment, when run through the FRStat *without* consideration of the measured clarity of features, 559 produced a FRStat ratio of 10 or greater. Among the 14 that were flagged by the DFIQI $Value_{GQS}$ pre-assessment, 7 resulted in an FRStat ratio supporting an association. These results are summarized in Table 4-1.

Total mark-exemplar pairs	**DFIQI $Value_{GQS}$**	**FRStat (*without* DFIQI LQS filter)**
605	591 Supported	559 Supported
		32 Flagged
	14 Flagged	7 Supported
		7 Flagged

Table 4-1: Proportion of mark-exemplar pairs evaluated by the DFIQI and FRStat algorithms (without consideration of the measured clarity of features) which resulted in algorithmic outputs that supported analysts' subjective opinions of "association" or flagged as warrating further quality assurance review. Note: This dataset of impressions was taken from a single federal laboratory in the United States which were considered "value for identification" and subsequently identified to exemplar impressions. Given the lack of quantifiable standards for "value for identification" at the time these impressions were examined, the extent to which these results are generalizable is unclear.

Among those same 591 that were supported by the DFIQI $Value_{GQS}$ pre-assessment, when run through the FRStat *with* consideration of the measured clarity of features (i.e., limiting the features eligible for FRStat to only those which were categorized by DFIQI LQS as green or yellow indicating "high" or "medium" clarity / reliability, respectively), 520 produced a FRStat ratio of 10 or greater. Of the 14 that were flagged by the DFIQI $Value_{GQS}$ pre-assessment, 4 resulted in an FRStat ratio supporting an association. These results are summarized in Table 4-2.

Total mark-exemplar pairs	DFIQI $Value_{GQS}$	FRStat (*with* DFIQI LQS filter)
605	591 Supported	520 Supported
		71 Flagged
	14 Flagged	4 Supported
		10 Flagged

Table 4-2: Proportion of mark-exemplar pairs evaluated by the DFIQI and FRStat algorithms (with consideration of the measured clarity of features limiting those features eligigble for entry into FRStat to those occuring in "high" or "medium" clarity regions) which resulted in algorithmic outputs that supported analysts' subjective opinions of "association" or flagged as warrating further quality assurance review. Note: This dataset of impressions was taken from a single federal laboratory in the United States which were considered "value for identification" and subsequently identified to exemplar impressions. Given the lack of quantifiable standards for "value for identification" at the time these impressions were examined, the extent to which these results are generalizable is unclear.

Taken together, under the most stringent circumstances when DFIQI $Value_{GQS}$ is applied as an initial gating function *and* DFIQI LQS is applied as a means of controlling the reliability of features eligible for entry into FRStat, 520 (86%) of the mark-exemplar pairs resulted in algorithmic outputs which supported the analysts' subjective judgments, warranting reports to be issued without additional quality assurance reviews. The remaining 85 (14%) would have been flagged by either DFIQI, FRStat, or both, thus warranting further examination prior to a report being issued. It is important to note that impressions (or mark-exemplar pairs) flagged for additional quality assurance review by the algorithms do not imply the analysts' original opinion was incorrect or inappropriate. Although this very well could be the case, it could also mean that those impressions might be lower quality and/or higher complexity compared to others (e.g., "borderline" cases) or a potential mistake by the analysts in their initial interpretation or annotation of features (e.g., additional features overlooked, mistakes during feature annotation, issues preventing the algorithms from accurately detecting image details, etc.). In either situation, an additional review seems prudent from a quality assurance perspective, and the algorithms can be an effective tool to triage cases and suggest when and where to strategically focus resources compared to arbitrary sampling schemes which are unlikely to detect potential issues as efficiently as the algorithms.

5 Evaluation of Practitioners' Perspectives

This chapter presents a manuscript entitled "'Mt. Everest—We are Going to Lose Many': A Survey of Fingerprint Examiners' Attitudes toward Probabilistic Reporting" (Swofford et al., 2021) [51] published in *Law, Probability & Risk* that explores practitioners' perspectives related to probabilistic reporting practices (with or without algorithmic tools) in terms of their reactions, attitudes, and sources of resistance toward probabilistic methods. Practitioners' perspectives are evaluated quantitatively and qualitatively using a structured survey instrument with Likert-scale response and free-text responses choices.

Mt. Everest—We Are Going to Lose Many:
A Survey of Fingerprint Examiners' Attitudes toward Probabilistic Reporting

[1]Swofford, H.; [2]Cole, S.; [2]King, V.

[1]School of Criminal Justice, Forensic Science Institute, University of Lausanne, Switzerland
[2]Department of Criminology, Law & Society, University of California, Irvine, CA, U.S.A.

5.1 Abstract

Over the last decade, with increasing scientific scrutiny on forensic reporting practices, there have been several efforts to introduce statistical thinking and probabilistic reasoning into forensic practice. These efforts have been met with mixed reactions—a common one being skepticism, or downright hostility, toward this objective. For probabilistic reasoning to be adopted in forensic practice, more than statistical knowledge will be necessary. Social scientific knowledge will be critical to effectively understand the sources of concern and barriers to implementation. This study reports the findings of a survey of forensic fingerprint examiners about reporting practices across the discipline and practitioners' attitudes and characterizations of probabilistic reporting. Overall, despite its adoption by a small number of practitioners, community-wide adoption of probabilistic reporting in the friction ridge discipline faces challenges. We found that almost no respondents currently report probabilistically. Perhaps more surprisingly, most respondents who claimed to report probabilistically, in fact, do not. Further, we found that two-thirds of respondents perceive probabilistic reporting as "inappropriate"—their most common concern being that defense attorneys would take advantage of uncertainty or that probabilistic reports would mislead, or be misunderstood by, other criminal justice system actors. If probabilistic reporting is to be adopted, much work is still needed to better educate practitioners on the importance and utility of probabilistic reasoning in order to facilitate a path toward improved reporting practices.

Keywords: Reporting, Testimony; Fingerprint; Categorical; Probability; Attitudes.

5.2 Introduction

Recent years have witnessed increasing efforts to introduce probabilistic reasoning into forensic practice—particularly in the pattern evidence disciplines. We define probabilistic reasoning in forensic practice as formally recognizing and articulating the uncertainties inherent in forensic interpretation using probabilistic logic. Forensic statisticians' efforts in these areas have primarily, and understandably, been concentrated in their area of technical expertise: statistics. Therefore, these efforts have been focused on such activities as developing statistical models [29, 31, 32, 37, 41-43, 50], building useful data sets [69, 70], and developing quality metrics [14, 18, 23, 26-28, 44, 49].

However, for probabilistic reasoning to be adopted in forensic practice, more than statistical knowledge will be necessary. Probabilistic reasoning will have to be adopted by the current workforce of forensic practitioners. It is not clear that this workforce is either knowledgeable about or committed to a probabilistic approach. Indeed, some practitioners have expressed skepticism, or downright hostility, toward probabilities and statistics (e.g., [46, 71, 72]).

In addition to statistical knowledge, therefore, social scientific knowledge will be necessary to actually enact the introduction of probabilistic reasoning into forensic practice. Such knowledge can help us understand issues such as: whether and to what extent forensic practitioners understand probabilistic reasoning; how better to educate practitioners in probabilistic reasoning; and whether practitioners welcome the introduction of probabilistic reasoning or are actively resistant to it and the reasons for, and sources of, such reactions.

The present study was intended to be a contribution to that effort. It used a survey of practitioners in a single forensic discipline—friction ridge examination—and it focused on a single deployment of probabilistic reasoning, which we call "probabilistic reporting"— that is, the reporting of forensic findings in probabilistic (as opposed to "categorical") terms. Friction ridge analysis was chosen because the researchers are familiar with the discipline and had connections with the large practitioner community, it is a widely used and influential pattern evidence discipline, the debate over probabilistic reporting is familiar to many in the discipline, and statistical tools have been developed and are familiar to the community. This study aims to capture baseline data on reporting practices across the discipline in order to: (i) ascertain what kind of reporting language friction ridge examiners and Forensic Service Providers (FSPs) currently use, and to what extent examiners and FSPs use probabilistic reporting, (ii) gauge friction ridge examiners' attitudes toward probabilistic reporting, and the reasons for, or sources of, those reactions; (iii) understand examiners' characterization of probabilistic reporting and what it means to report probabilistically.[8]

It is hoped that this study will be useful for scientists interested in fostering the use of probabilistic reasoning in forensic science. It may also be of interest to forensic practitioners, laboratory administrators, legal scholars, social scientists, and others interested in the introduction of statistical thinking into forensic practice. The findings may help these groups better understand

[8] The study originally had a fourth goal: to capture and record the experiences of latent print examiners who have adopted probabilistic reporting. However, we received only 6 survey responses (2%) from such examiners. We deemed this an insufficient sample, and we do not address this goal further here.

the degree of penetration of probabilistic reasoning that has already been achieved, the reasons practitioners may welcome or resist the introduction of probabilistic reasoning, and how to improve education and implementation efforts.

5.3 Background

Friction ridge impression evidence (colloquially referred to as "fingerprint evidence") has long been considered one of the most important kinds of forensic evidence used in criminal and civil litigation and is often regarded by jurors and other criminal justice system actors as incontrovertible proof that an individual touched an item in question [73-77]. This is based upon decades of testimony that fingerprint evidence is unique to an individual and that no two individuals, including identical twins, share the same arrangement of friction ridge skin [78]. Friction ridge examination consists of visual observation and comparison of friction ridge details between two impressions. Traditionally, the process for conducting friction ridge examinations is described by the acronym ACE-V, which stands for "Analysis," "Comparison," "Evaluation," and "Verification" [79]. ACE-V has been described in the forensic literature as a means of comparative analysis of evidence since 1959 [3].

For nearly a century, latent print examiners have expressed their findings in categorical terms with statements or implications of absolute certainty, something also true of many other forensic disciplines [80]. When we characterize reporting as "categorical," we mean that reporting follows a system in which reports are assigned to "categories" which are treated as homogeneous within and mutually exclusive. For example, in friction ridge analysis it is common to report results in three categories often named "exclusion," "inconclusive," and "identification." Although categorical reporting does not require statements of certainty (and often allows for statements of uncertainty), historically it has been common to treat one or both of the endpoints of categorical frameworks (i.e., "exclusion" and "identification" in the framework above) as statements either of certainty or of some state of quasi-certainty that can be treated as tantamount to certainty [6]. So, for example, lay fact-finders were often told that fingerprints "matched" with "100% certainty" and the two impressions were made by the same source [81]. Over time, terms such as "match," "identification," and "individualization" became synonymous expressions, all of which meant that a specific individual was determined to be *the* source of an impression [6]. Such claims have been criticized as unsupportable by individual scientists and scholars [37, 45, 82-89] and a number of governmental and scientific reports [3, 7-9, 90]. While this paper is not intended to review these debates, we summarize the criticisms as follows: First, statements of certainty, to the extent that they are being made, are inherently misleading and unscientific—they systematically overstate the value of the evidence. Forensic results, particularly those with an inclusionary outcome, cannot preclude the possibility of any considered hypothesis. Proper reporting of forensic results should therefore account for the probability of the evidence under the considered hypotheses and some probability, even if small, must necessarily be assigned to each hypothesis [91]. Second, statements of certainty aside, categorical frameworks are too simplistic. They treat all forensic results as equivalent, no matter how different, assigned to the same category. And, they may overstate the difference between two forensic results that are quite similar but fall on opposite sides of the arbitrarily defined boundary between two categories. Ideally, then, forensic results should be reported along a continuum rather than in categories [45]. How this should be done is not

something we will discuss in this paper, but, to summarize, proposals range from expanded "verbal scales" to expressing probabilistic statements along a continuum. Methods for expressing probabilistic findings range from the use of likelihood ratios to the use of accuracy data, and from the use of statistical models and associated software to the use of subjective probabilities based on human judgment.

The friction ridge discipline has responded with statements that limit strength of the claim that the words "identification" and "individualization" are supposed to convey [10, 92]. These changes, however, insisted on retaining the terms themselves and the claim that *two impressions were made by the same source* while dispensing with the phrase "*to the exclusion of all others*," resulting in ensuing criticisms that the change had no practical impact [93, 94]. At least one crime laboratory, the United States Army Criminal Investigation Laboratory (USACIL), the primary forensic laboratory supporting the criminal investigative mission of the Department of Defense, announced a policy change to abandon the term "identification" and report their findings in a probabilistic framework [95]. In 2017, the USACIL went a step further and announced the implementation of a statistical software application, *FRStat*, to provide probabilistic support to fingerprint associations [96].[9] In 2018, the Organization for Scientific Area Committees (OSAC) for Forensic Science, Friction Ridge Subcommittee (OSAC FRS), which is responsible for the promulgation of standards and best practices related to the forensic examination of friction ridge skin impression evidence nationwide, released the proposed Standard for Friction Ridge Examination Conclusions [97], which took an additional step toward ensuring a probabilistic expression. While the proposed standard maintains the term "identification," it was redefined in a probabilistic likelihood ratio format [97]. In addition to the revised definition, the OSAC FRS states that "an examiner shall not assert that a source identification is the conclusion that two impressions were made by the same source or imply an individualization to the exclusion of all other sources" [97].

This debate over reporting practices provides the context for the present study. However, the purpose of this study was not to advance the debate for or against probabilistic reporting. Rather, it was to try to elicit the perspectives of a practitioner community on the prospect of probabilistic reporting.

[9] The *FRStat* software is method developed by the United States Army Criminal Investigation Laboratory (USACIL) designed to serve as a quality assurance tool and a means of quantitatively conveying the significance of an association observed by an examiner. The *FRStat* development and validation is described by Swofford et al., 2018. Briefly described, the *FRStat* first calculates a similarity value (called GSS) between two sets of features identified by an examiner on two separate impressions which the analyst believes to correspond. The software then provides two estimates, one indicating how often prints originating from common sources would result in a GSS that is equal to or greater less than the calculated GSS and another indicating how often prints from different sources would result in a GSS that is equal to or greater than the calculated GSS. The two values are then combined as a ratio providing a single summary statistic indicating to what extent the GSS is consistent with originating from a common source compared to different sources. Generally speaking, higher values of this ratio indicate greater evidence in favor of the analyst's opinion of association; lower values indicate less evidence in favor of the analyst's opinion of association and may serve as a quality assurance tool to flag a comparison as potentially problematic due to insufficient similarity to support an association, based on the thresholds and standards set by an organization's policies.

5.4 Methods

Participant Recruitment and Survey Administration

Participants were recruited to participate in the study by invitation through their membership in the International Association for Identification (IAI), the largest professional organization of forensic fingerprint practitioners in the world, and through word of mouth by members of the friction ridge (fingerprint) community. The survey was emailed to approximately 1,700 IAI members listed as having background in friction ridge examination (see Appendix D-1 for the recruitment email). On the Study Information Sheet (but not in the Recruitment Email) participants were informed that they would receive a Center for Statistical Applications in Forensic Evidence (CSAFE)-branded coffee mug for completing the survey. Eligible participants were forensic practitioners 18 years of age or older. Participants were provided a link to an online survey using a commercial survey platform, Qualtrics®. All responses to the survey were anonymous. As will be discussed below, the survey received a total of 301 survey responses.

The survey was open for a two-month period during August and September 2018. After giving informed consent, participants were presented with a series of questions pertaining to their demographics (gender, age, and education), employment and testimony experience. Participants were then provided a closed-response question in which they were asked to choose which of three sample statements most closely resembled the wording they currently used in reports of an association between two friction ridge impressions (see Appendix D-2). The first option was meant to encompass a variety of different "categorical" ways in which friction ridge examiners tend to report and testify. The second was intended to encompass the variety of ways in which friction ridge examiners currently try to testify "probabilistically." The third, which we call "demonstrability," was intended to capture a kind of reporting currently advocated by some practitioners which emphasizes the "demonstrability" of the conclusions more than their probabilistic nature [98]. We refer to this question as the "trigger question" because it was used to initially divide the subject pool into two groups, labelled "probabilistic" and "categorical," which were administered questions slightly differently in parts of the remainder of the survey.[10] For purposes of this binary assignment, "demonstrability" respondents were aggregated with the "probabilistic" group.[11]

Next, all participants were given an open-response question which asked them to provide the actual language used in their written examination reports when reporting an association

[10] Participants responding to the trigger question indicating that their reporting language was "probabilistic" were asked to report their attitudes toward probabilistic reporting: (1) *before* making a change to probabilistic reporting and (2) *after* making a change to probabilistic reporting (currently) in order to understand the degree to which their views have changed over time, if at all. This group was also asked an additional set of questions concerning what was most and least effective with helping the participant understand the importance of probabilistic reporting and gain comfort with reporting and testifying using probabilistic conclusions. However, we received only 6 survey responses (2%) from examiners who have adopted probabilistic reporting. We deemed this an insufficient sample, and we do not address this further.

[11] We doubt that either forensic statisticians or the proponents of the approach themselves would consider the statements associated with "demonstrability" probabilistic. We combined these responses with the "probabilistic" group because, though not probabilistic, they do represent a desire to move beyond the *conventional* categorical approach, even if the demonstrability approach still does consider itself categorical [99].

between two friction ridge impressions in their practice. In a follow-up binary, closed-response question, participants were then asked whether they believed their actual reporting language was "probabilistic" or "non-probabilistic (categorical)."

After being divided into two groups based on their responses to the trigger question, participants were administered a series of Likert-scale and free-response questions regarding their positions toward probabilistic reporting. Likert-scale questions included five response choices indicating the extent participants agree or disagree with the statements provided. The Likert-scale response choices included: "strongly agree," "somewhat agree," "neither agree or disagree," "somewhat disagree," "strongly disagree." Likert-scale responses were evaluated quantitatively and free text responses were evaluated qualitatively through researcher coding and analysis using *Atlas.ti*® software. The raw surveys and our coding are publicly available through the CSAFE data portal [70].

Current Reporting Practices

Categorical versus Probabilistic

The first aim of the survey endeavored to capture current reporting practices for associations between friction ridge impressions. This was accomplished in two different ways. First, examiners were asked to choose from three fixed options (referred to earlier as the trigger question). Second, we offered participants the opportunity to articulate their reporting language in their own terms. Following this second probe, participants were asked to self-report whether they believed the language they used was "probabilistic" or "non-probabilistic (categorical)." This second probe allowed us the opportunity to evaluate whether the submitted language was or was not probabilistic and compare participants' self-reports to our own evaluations. The free text responses for which participants provided samples of the actual language used in their written examination reports when reporting an association between two impressions were evaluated and coded as "probabilistic" or "non-probabilistic (categorical)" independently by two of the researchers using the criteria outlined below:

> *Statements are coded "Probabilistic" if they openly and transparently assign in any way (verbal or numerical) a probability to the alternative hypothesis.*
>
> *Statements are coded "Categorical" if they do not assign a probability to the alternative hypothesis or if they do assign a probability to the alternative hypothesis but, in the same statement, minimize, belittle, or otherwise encourage the disregarding of that probability.*

Coding discrepancies between the two researchers were reviewed by the third researcher and discussed until a consensus was reached. This design allowed us to compare participants' self-reports to our own evaluations of whether or not statements were probabilistic.

Types of Categorical and Probabilistic Reporting

In order to achieve greater specificity about the nature of the statements being used, we subdivided "categorical" and "probabilistic" statements into two subcategories each. We subdivided categorical statements into "Traditional" or "Elaborated," following the nomenclature proposed by Bali *et al.* [80]. Traditional statements are generally those kinds of statements that have pervaded the friction ridge discipline for the past century. Elaborated statements are those that appear to recognize that the manner of reporting needs to change, but appear to do so subtly. An example would be the redefinition of the term "individualization" by the Scientific Working Group for Friction Ridge Analysis Study and Technology (SWGFAST) to mean "the decision that the likelihood the impression was made by another (different) source is so remote that it is considered as a practical impossibility" [10]. More specifically, "Traditional" statements (e.g., "the two prints are from the same source"; "this finger made this print"; "the print was identified to the defendant"; "I made an identification," etc.) assign no probability to the alternative hypothesis. "Elaborated" statements assign a probability to the alternative hypothesis but also minimize it with a statement that encourages disregarding it (e.g., "practical impossibility," "negligible," "discounted," etc.).

We subdivided probabilistic statements into two categories according to the degree of rigor with which the statements follow the logical and formal rules of probabilistic reporting (e.g., clearly articulating hypotheses) [100]. Statements in the first category, which we call "Probability of Findings," tend to articulate two hypotheses and characterize the probability of the evidence. Statements in the second category, which we call "Probability of Hypothesis," tend to articulate only one hypothesis and characterize the probability of the hypothesis (i.e. posterior probabilities) as opposed to the probability of the evidence. Based on our reading of the statement, we believe the "Probability of Hypothesis" statements represent *efforts* to testify in a logical probabilistic manner which have, like many efforts to speak probabilistically, inadvertently transposed the conditional [101].

Support for Probabilistic Reporting (Agency Policy versus Personal Belief)

In addition to gaining a general understanding of the extent to which examiners report categorically vs. probabilistically, we were also interested in understanding the extent to which examiners support reporting probabilistically. By asking examiners to report the statements that they use in their actual reports, we may have captured agency policy rather than examiners' personal beliefs. In order to better understand examiners' personal positions toward probabilistic reporting, thus capturing whether such examiners may be at least open to the idea of a transition, we asked several questions designed to elicit their personal beliefs about categorical and probabilistic reporting. These were explored using the following Likert-scale response questions for all respondents:

> *I feel that the proposed shift away from "identification" and the use of probabilistic language is an appropriate direction for the fingerprint community.*

I do not understand why there is concern with expressing positive conclusions in absolute terms, such as "identification."

I support probabilistic reporting because it is a scientifically more appropriate means of expressing positive fingerprint conclusions.

I do not understand why probabilistic conclusions are more appropriate means of expressing positive fingerprint conclusions.

I am willing to take an active role in helping other practitioners become more understanding and accepting of probabilistic reporting.

Attitudes Toward Probabilistic Reporting

Receptivity to Probabilistic Reporting

The second aim of the survey endeavored to capture examiners' attitudes toward probabilistic reporting and the reasons for, or sources of, those reactions. This was accomplished in several ways. First, our survey question "I feel that the proposed shift away from 'identification' and the use of probabilistic language is an appropriate direction for the fingerprint community" probed the current state of examiners' receptivity to probabilistic reporting. Second, in order to gain greater insight into examiners' views, we solicited a free text response which invited participants to elaborate on why they agree or disagree "that the proposed shift away from 'identification' and the use of probabilistic language is an appropriate direction for the community." The free text responses were analyzed in three groups according to aggregated responses from the Likert-scale: (1) those who perceive probabilistic reporting as *appropriate* (i.e., those who responded "somewhat" or "strongly agree"); (2) those who perceive probabilistic reporting as *inappropriate* (i.e., those who responded "somewhat" or "strongly disagree"); and (3) those who were *neutral* as to the appropriateness of probabilistic reporting (i.e., those who responded "neither agree nor disagree"). Free text responses were single-coded by the second author in order to derive themes that emerged from the data according to a grounded theory approach. The second author has been studying friction ridge analysis from historical, sociological, epistemological, and rhetorical perspectives for more than 20 years and, therefore, is, we believe, sufficiently familiar with the jargon of the discipline to interpret the responses. After a provisional list of themes was generated, the themes were re-evaluated, and some themes were aggregated, disaggregated, or deleted. The researcher then made a second pass through the data using this final list of themes. There was no maximum placed on the number of themes which could be applied to any single response, but the minimum was 1 to ensure the assignment of at least one theme to each response (i.e., if a response did not fit any existing theme, a new theme was added).

General Opposition to Probabilistic Reporting

In addition to the free text responses allowing participants to elaborate on why they agree or disagree "that the proposed shift away from 'identification' and the use of probabilistic language

is an appropriate direction for the community," we were also interested in understanding key reasons for the opposition in a more structured way. We accomplished this by asking all participants several Likert-scale response questions designed to elicit general reasons why they may be opposed to probabilistic reporting. The questions were selected based on our anecdotal observations of examiners' claims or implications at conferences, online chat boards, and informal discussions over the last several years:

I feel that law enforcement, special agents, and/or other investigators would not understand how to interpret probabilistic conclusion language.

I feel that defense attorneys would take advantage of probabilistic conclusion language to create reasonable doubt.

I feel that prosecutors would be less willing to use fingerprint evidence in court because of the probabilistic conclusion language.

I feel that judges and/or jurors would not understand probabilistic conclusion language.

I feel that I do not sufficiently understand probabilities and would not to be able to properly testify to my conclusion in court.

I feel that a probabilistic conclusion is too weak of a conclusion.

I feel that a probabilistic conclusion would negatively impact the outcome of a trial.

I feel that if I were to report and/or testify to probabilistic language that my certification with the International Association for Identification (IAI) would be in jeopardy.

I feel that probabilistic reporting will cause the number of erroneous associations to significantly increase.

Characterizations of Probabilistic Reporting

The third aim of the survey endeavored to understand examiners' characterization of probabilistic reporting and what it means to report probabilistically. Although the concept of "probabilistic reporting" has been advocated by proponents, it is unclear what examiners understand those words to mean and whether they differ from one another. In a free-text response question, we sought to allow the respondents to tell *us* what they understood the term "probabilistic reporting" to mean with the following question: "How would you describe probabilistic reporting, compared to non-probabilistic (categorical) reporting?" For analysis purposes, we divided respondents into two groups (categorical reporters and probabilistic reporters) based on the trigger question discussed above. Using the same analysis procedures described above, the free text responses were reviewed by one of the researchers in order to derive themes that emerged from the data according to a grounded theory approach.

5.5 Results

Participant Responses

A total of 435 raw survey responses were received in the two-month period the survey was open (August and September 2018); however, 134 survey responses omitted the trigger question and were largely incomplete or blank altogether, and we discarded them. Given that the survey was made freely available on the internet, many "responses" may have reflected individuals who preferred to view the survey, rather than complete it. After removing the incomplete surveys, a total of 301 completed surveys were available for evaluation—yielding a response rate of 17.7% out of approximately 1,700 IAI members invited to participate. A completed survey, however, does not mean that the respondent answered every question.

Of the 301 respondents who completed surveys, the demographics are as follows: 44% were male and 54% were female (2% unreported), 88% reported being employed in the United States and 10% reported being employed outside of the United States, 83% reported having a Bachelor's Degree or higher, 84% have testified in court, 54% testified in court in the past year, the average years of experience is 16.5 years (standard deviation of 11.2 years), and the distribution of participants' ages was: 8% reported as 20-29 years, 36% reported as 30-39 years, 28% reported as 40-49 years, 19% reported as 50-59 years, 8% reported as 60-69 years, 2% reported as 70-79 years, and 2% unreported.

Current Reporting Practices

Categorical versus Probabilistic

Among the 301 completed surveys, all participants responded to the trigger question with three fixed reporting options. The responses to this question are shown in Table 5-1.

Response	Brief description	n
Categorical	Same source; identified to, matched	88% (264/301)
Probabilistic	Likelihood same or different source	10% (31/301)
Demonstrability	The conclusion is easily demonstrable to others	2% (6/301)

Table 5-1. Breakdown of fixed options of current reporting practices for associations. Examiners selected the samples of fixed reporting options that most closely resembled their own. The fixed reporting options acted as a "trigger question" to initially categorize respondents as reporting categorically or probabilistically. The "demonstrability" option was treated as probabilistic for purposes of this initial categorization.

Table 5-1 shows that the vast majority of respondents (88%) use categorical reporting language. Probabilistic reporting is rare (10%) and only 2% of respondents use "demonstrability"

language. In short, 9 out of 10 friction ridge examiners surveyed responded that they report categorically.

Among the 301 completed surveys, only 247 provided free text responses with a sample of their actual reporting language. For those 247 respondents, we were able to code whether the reporting language was categorical or probabilistic (coding discrepancies between the two researchers occurred in 3 of the 247 responses—all three related to whether a categorical statement should be sub-coded as "traditional" versus "elaborated"). Table 5-2 compares respondents' self-reports with researcher coding. The self-reports for the smaller group of 247 respondents are consistent with those found in the larger group of 301 (Table 5-1): 88% of examiners described themselves as reporting categorically (Table 5-2, column 2). Surprisingly, even among the small number of examiners who purported to report probabilistically, the majority of them, in our view, report categorically. Conversely, two respondents described themselves as reporting categorically, but provided a sample statement that we interpret to be probabilistic (columns 4 and 6).

1	2	3	4	5	6
Self-report		**Researcher coded**		**Total (Researcher coded)**	
Categorical	88% (217/247)	**Categorical**	87% (215/247)	**Categorical**	98% (241/247)
		Probabilistic	<1% (2/247)		
Probabilistic	12% (30/247)	**Categorical**	10% (26/247)	**Probabilistic**	2% (6/247)
		Probabilistic	2% (4/247)		

Table 5-2. Breakdown of categorical vs. probabilistic reported based on respondents' self-report compared to researcher coding of the actual reporting language. The arrows show the consequences of reassigning respondents whose self-reports the researchers deemed incorrect.

Readers may wonder whether and why they should treat our assessment of whether a statement is probabilistic as dispositive. We believe that readers who examine the statements for which we disagreed with the participant's self-report will find our assessments uncontroversial (taking note of the definitions of these two categories we provide above). For example, one respondent self-reported that the following statement was "probabilistic"; however, we interpreted it to be categorical: "Identification is the opinion of an examiner that there is sufficient quality and quantity of detail in agreement to conclude that two impressions originated from the same source." Conversely, another respondent self-reported that the following statement was "categorical"—we coded it as probabilistic: "In the opinion of this examiner the likelihood that the impressions were made by a different source other than the one listed is very small." A full listing of all the statements for which researcher coding conflicted with self-report is presented in Appendix D-3.

Table 5-2 suggests that categorical reporting is even more prevalent than the 90% figure found in the self-reports. According to our coding of actual provided sample reporting language,[12]

[12] Readers may wonder whether multiple participants from the same laboratory participated in the survey. The anonymous nature of the survey precludes any insights into this.

98% of respondents are using categorical statements to describe associations between friction ridge impressions. Only 6 of the 247 respondents provided statements that we interpret as probabilistic.

Types of Categorical and Probabilistic Reporting

Table 5-3 provides the breakdown of respondents' categorical statements subdivided as "traditional" versus "elaborated" and respondents' probabilistic statements subdivided as "probability of findings" versus "probability of hypothesis."

	Researcher coded		**Statement subtype**
Categorical	98% (241/247)	**Traditional**	89% (220/247)
		Elaborated	9% (21/247)
Probabilistic	2% (6/247)	**Probability of Findings**	2% (4/247)
		Probability of Hypothesis	<1% (2/247)

Table 5-3. Breakdown of categorical and probabilistic reporting into subtype based on determinations of categorical vs. probabilistic from researcher coding (traditional vs. elaborated for categorical and probability of findings vs. probability of hypothesis for probabilistic).

Support for Probabilistic Reporting (Agency Policy versus Personal Belief)

Table 5-4 shows the responses from participants who actually report categorically related to questions designed to elicit their personal support for probabilistic reporting. As Table 5-4 shows, between 32% and 43% of respondents who report categorically responded in ways that suggest they personally support probabilistic reporting.

Question	**Likert-Measure**	**Total**	**Degrees of dis/agreement aggregated**
I feel that the proposed shift away from "identification" and the use of probabilistic language is an appropriate direction for the fingerprint community.	Strongly agree	9% (24/265)	32% (85/265)
	Somewhat agree	23% (61/265)	
	Neither agree or disagree	10% (26/265)	10% (26/265)
	Somewhat disagree	28% (73/265)	58% (154/265)
	Strongly disagree	31% (81/265)	

I do not understand why there is concern with expressing positive conclusions in absolute terms, such as "identification."	Strongly agree	22% (59/265)	49% (129/265)
	Somewhat agree	26% (70/265)	
	Neither agree or disagree	9% (23/265)	9% (23/265)
	Somewhat disagree	25% (65/265)	43% (113/265)
	Strongly disagree	18% (48/265)	
I support probabilistic reporting because it is a scientifically more appropriate means of expressing positive fingerprint conclusions.	Strongly agree	12% (34/280)	37% (104/280)
	Somewhat agree	25% (70/280)	
	Neither agree or disagree	12% (33/280)	12% (33/280)
	Somewhat disagree	21% (58/280)	51% (143/280)
	Strongly disagree	30% (85/280)	
I do not understand why probabilistic conclusions are more appropriate means of expressing positive fingerprint conclusions.	Strongly agree	21% (59/280)	46% (128/280)
	Somewhat agree	25% (69/280)	
	Neither agree or disagree	14% (38/280)	14% (38/280)
	Somewhat disagree	25% (69/280)	41% (114/280)
	Strongly disagree	16% (45/280)	
I am willing to take an active role in helping other practitioners become more understanding and accepting of probabilistic reporting.	Strongly agree	13% (37/280)	34% (94/280)
	Somewhat agree	20% (57/280)	
	Neither agree or disagree	34% (95/280)	34% (95/280)
	Somewhat disagree	13% (35/280)	33% (91/280)
	Strongly disagree	20% (56/280)	

Table 5-4. Participants' (categorical respondents based on researcher coding) responses to Likert-scale questions related to examiners' personal beliefs and support for probabilistic reporting.

Attitudes Toward Probabilistic Reporting

Receptivity to Probabilistic Reporting

The first question on Table 5-4 shows the range of responses on the Likert-scale to the question "I feel that the proposed shift away from 'identification' and the use of probabilistic language is an appropriate direction for the fingerprint community" for respondents and then shows aggregations of the Likert-scale responses into broader categories. From these data, we see most friction ridge examiners feel that probabilistic reporting is not an appropriate direction for the community (58%). Few examiners are neutral on the issue (10%). Among the 239 respondents with opinions (responding "somewhat" or "strongly dis/agree" in either direction), approximately two-thirds view it as inappropriate and one-third view it as appropriate. Although the proportion who view it as appropriate is far greater than those who actually apply probabilistic reporting, the majority of the examiner community at large still remains generally opposed to, or at least skeptical of, probabilistic reporting.

When invited to elaborate on why participants agree or disagree "that the proposed shift away from 'identification' and the use of probabilistic language is an appropriate direction for the community" in a free text response, respondents expressed a wide diversity of opinions to this question. Some respondents wrote long disquisitions, and at least one complained about the lack of space in which to enter a response. The mean and median number of themes for each response was 3. The minimum was 1 (our coding rules required the assignment of at least one theme to each response), and the maximum was 8.

Respondents supporting probabilistic reporting

Among the 85 participants who consider probabilistic reporting *appropriate* (i.e., those who responded "somewhat" or "strongly agree" on the first question in Table 5-4), a total of 36 different themes were identified suggesting *why* they viewed it as appropriate, and a total of 113 themes were coded across the 85 respondents. Table 5-5 shows all themes that were mentioned by more than one respondent. The full list of themes is available through the CSAFE data portal (see [70]).

Number	Theme	Frequency
1.	more accurate	12
2.	more scientific	12
3.	uncertainty	10
4.	jury clarity	9
5.	transparency	7
6.	weight of evidence	7
7.	finer	6
8.	uniqueness unproven	6
9.	objectivity	5
10.	it's happening	3
11.	statistic ok with verbal	3
12.	appropriate	2
13.	consistent with other disciplines	2
14.	external scrutiny	2
15.	law	2
16.	overselling	2
17.	reliance on stock phrases/expertise	2
18.	sound reasoning	2
19.	step in right direction	2

Table 5-5. "Shift to probabilistic reporting is appropriate": all coded themes (n=85).

The most common responses were that probabilistic reporting was an improvement over past or current practice. This was most commonly described as either "more accurate" or "more scientific." An example of a "more accurate" statement is:

> *It is a more accurate description of my observations and the limits of my observations (38).*[13]

An example of a "more scientific" statement is:

> *I think it's time for LP [latent print] examination to apply more scientific rigor to the practice of latent print examination. This would include articulating probabilities when reporting results of examinations. I would hope that this ultimately leads to more credibility for friction ridge examination as a forensic discipline (416).*

These respondents also viewed probabilistic reporting's ability to convey uncertainty as an advantage. For example, one respondent wrote:

[13] All quotations or references to specific responses from the survey are followed by a parenthetical reference to the respondent number. Respondent numbers were assigned to all respondents who opened the survey, including those who did not complete it. Therefore, some respondent numbers are greater than the total number of respondents, 301. Spelling errors in the responses are corrected, but grammatical errors are not.

> *We shouldn't speak in absolute terms. Identification over estimates the evidence and our conclusions should convey the level of certain [sic] we know. Even if we know what we mean by identification it does imply to a jury absolute certainty. It's not scientific. I would like the field to be better (31).*

Several respondents specifically cited the "weight of evidence," an important concept in forensic statistics, as an advantage of probabilistic reporting. Many of these respondents perceived probabilistic reporting as offering greater "jury clarity" and "transparency" for the jury. This contrasts with many respondents who do not support probabilistic reporting and cited "jury confusion" as a reason for their opposition (see below).

Overall, respondents offered a rich and diverse set of reasons for the appropriateness of probabilistic reporting. There were reference to epistemological considerations (e.g., "uniqueness unproven"; "objectivity"), external forces (e.g., "it's happening," "external scrutiny," "law"), and perceived problems with current practice (e.g., "overselling," "reliance on stock phrases/expertise," "conclusions are not overstated"). However, in contrast to the responses that probabilistic reporting is inappropriate, those who consider it appropriate were characterized by a clearly perceptible degree of ambivalence. Many responses coded as "appropriate" listed some advantages of probabilistic reporting, but then turned to some perceived disadvantage.[14] An example is this:

> *It gives a more accurate representation of the validity of the conclusion and results reported. However, I do worry it may confuse the issue in the event of a distracted/inattentive jury (419).*

Recognizing the importance of this ambivalence, we coded such responses as containing "reservations." Fully 36 of the 85 responses (42%) that considered probabilistic reporting appropriate contained such reservations. Such ambivalence was not nearly as common among the responses that considered it inappropriate, and so "reservations" in those responses were not counted.

[14] It is important to recall that in this analysis respondents were categorized according to their responses to the Likert-scale question, *not* according to researcher coding of their free-text responses. Thus, a respondent who reported that they "somewhat agree" that probabilistic reporting is appropriate and submitted a free-text response criticizing probabilistic reporting was still analyzed in the "appropriate" group. For example, the following free-text responses were made by respondents who self-reported that they agreed or somewhat agreed that probabilistic reporting was appropriate: *"I believe much more study into the usage of probabilistic statements is required. Specifically, in district and county courts within the US, not just military court as is currently being done"* (86); and *"The shift towards probabilistic language seems contradictory to the principles of fingerprint identification that I learned in my training years. I was taught to give definitive conclusions - not maybes. Additionally, I worry that probabilistic conclusions may have the negative effect of increasing the number of people being erroneously associated with a given latent print"* (374).

Respondents not supporting probabilistic reporting

Among the 154 participants who considered probabilistic reporting *inappropriate* (i.e., those who responded "somewhat" or "strongly disagree" on the first question in Table 5-4), a total of 49 different reasons were identified suggesting *why* they viewed it as inappropriate, and a total of 265 themes were coded across the 154 respondents. Table 5-6 shows all themes that were mentioned by more than one respondent. The full list of themes is available through the CSAFE data portal (see [70]).

Number	Theme	Frequency
1.	jury confusion	36
2.	probability not ready	26
3.	underselling	23
4.	quantification impossible	18
5.	opinionization	15
6.	uncertainty can be eliminated	14
7.	misleading	12
8.	uniqueness	12
9.	wealth of empirical data	9
10.	DNA paradigm inappropriate	7
11.	unnecessary	7
12.	appropriate in some cases	6
13.	combine probability with id	6
14.	models don't capture all information	6
15.	vague	6
16.	only problem uncertainty	5
17.	verbal/subjective probabilities unacceptable	5
18.	politics	4
19.	skeptical of statistics as discipline/all models wrong	4
20.	customer is police/attorneys	3
21.	examiner confusion	3
22.	risk contradicting ground truth more often	3
23.	transparency sufficient	3
24.	whole population problem	3
25.	accuracy more important	2
26.	defense exploitation	2
27.	doesn't prefer	2
28.	not science	2

Table 5-6. "Shift to probabilistic reporting is inappropriate": all coded themes (n=154).

The most common response was that probabilistic reporting would confuse the jury:

> *Our job in court is to make the jury understand the evidence. I feel that the language being referred to will just confuse a lay person (36).*

For many examiners, the "confusion" they feared lay in the move away from communicating results in the simple form of certainty:

> *This shift seems to add confusion. It's giving some degree of uncertainty to our conclusions (185).*

This contrasts with the respondents favoring probabilistic reporting who viewed "uncertainty" as an advantage.

A similar concern was invoked when examiners expressed concern about "underselling"—the concern that probabilistic reporting is too "weak" and would understate the probative value of the evidence:

> *No accurate or "full" way to express LPE's opinions. The numbers are weak and meaningless (413).*

Compared to respondents who consider probabilistic reporting appropriate, some examiners cited "jury confusion" to express the opposite concern: that probabilistic reporting would overstate, rather than understate, the value of the evidence:

> *Until probabilistic language is valid and reliable, as well as easily understood by all practitioners and easily explained by those practitioners to a judge and jury, they are useless numbers. They provide to confuse the trier of fact, add little to the data, and very likely will serve to bolster the testimony and evidence (24).*

The second most common reason given was "probability not ready." These respondents communicated that they were not opposed to probabilistic reporting in principle. Their opposition, rather, was a practical consideration based on their perception of the current state of affairs with regard to the development of a statistical model useable for assigning a probability to a friction ridge association:

> *There is no current scientific basis for a probabilistic model. Reporting latent print conclusions in probabilistic language is misleading and unscientific. Statistics are not in themselves scientific or objective (254).*

Some other common themes cited fundamental epistemological concerns about the use probability to communicate the findings of friction ridge analysis. One such concern is "quantification is impossible":

> *I believe there cannot be probabilistic language involved with latent print examination. There are too many variables involved. Latent prints examination is too subjective to have probability (169).*

These responses suggest the impossibility of quantification as deriving from the limits of friction ridge analysis: it is inherently uncertain, subjective, and reliant on continuous rather than discrete measures and, therefore, impervious to quantification. In contrast, other respondents resisted the premise that friction ridge analysis is inherently uncertain, claiming "uncertainty can be eliminated":

> *When I form the opinion that I have made an identification - I am certain that this is the person. There is no probability of it being from someone else. If I am not certain then I will not say it's an identification but that I can not exclude them as being the contributor (258).*

> *Ident is ident I don't see the purpose of assigning a probability number, it's 100% or I wouldn't call it a ident (172).*

A third such concern is reference to "uniqueness":

> *Strongly disagree because it is impossible to apply a probability to something, anything, that is unique. The scientific basis in biology and other natural sciences is well established. Therefore, probabilistic language is inappropriate and unscientific (270).*

Some respondents expressed outright hostility toward the discipline of statistics, which was called, for example, a "bandwagon" and a "fad" (239). Another respondent commented:

> *The probabilistic method (not language) is put upon us because it is supposed to be more scientific. It comes from the DNA science. I see that nowadays DNA evidence in practice is accepted as Empirical fact in spite of statistical humdrum (296).*

There was also some hostility expressed toward DNA, primarily for imposing an inappropriate paradigm on friction ridge analysis. "Politics," defense attorneys, defense experts, and "critics" were also perceived as imposing the shift toward probabilistic reasoning on the discipline.

Respondents neutral to probabilistic reporting

Among the 26 participants who were *neutral* as to the appropriateness of probabilistic reporting (i.e., those who responded "neither agree nor disagree" on the first question in Table 5-4), a total of 27 different reasons were identified, and a total of 36 themes were coded across the 26 respondents. Table 5-7 shows all themes that were mentioned by more than one respondent. The full list of themes is available through the CSAFE data portal (see [70]).

Number	Theme	Frequency
1.	jury confusion	5
2.	truly undecided	4
3.	doesn't understand	2
4.	unnecessary	2

Table 5-7. "Neutral on shift to probabilistic reporting is appropriate": all coded themes (n=26).

General Opposition to Probabilistic Reporting

Table 5-8 shows the responses to the Likert-scale questions designed to elicit key reasons why examiners may be opposed to probabilistic reporting and provides them in rank order based on the proportion of examiners who responded "agree" or "strongly agree" to each question.

Key Reasons for Opposition to Probabilistic Reporting (in rank order)	Agree or Strongly Agree
Defense attorneys would take advantage of probabilistic conclusion language to create reasonable doubt	79% (211/266)
Judges and/or jurors would not understand probabilistic conclusion language	79% (209/264)
Law enforcement, special agents, and other investigators would not know how to interpret probabilistic conclusion language	69% (184/265)
A probabilistic conclusion is too weak of a conclusion	48% (127/265)
I do not sufficiently understand probabilities and would not be able to properly testify	44% (116/266)
A probabilistic conclusion would negatively impact the outcome of a trial	41% (110/266)
Prosecutors would be less willing to use fingerprint evidence in court	38% (100/266)
Probabilistic reporting will cause the number of erroneous associations to significantly increase	28% (80/281)
My certification with the International Association for Identification would be in jeopardy	15% (40/266)

Table 5-8. Participants' (categorical respondents based on researcher coding) responses to Likert-scale questions related to possible reasons for opposition to probabilistic reporting (from tables 5-6 and 5-7) in rank order. Note: n's vary by question.

Characterizations of Probabilistic Reporting

When asked to describe probabilistic reporting compared to non-probabilistic (categorical) reporting in a free text response, 192 participants responded. Among those, 177 were respondents who had self-reported that they report categorically and 15 were respondents who had self-reported that they report probabilistically. Due to the small number of respondents indicating they report probabilistically, we only discuss the results of the 177 participants who indicated they report categorically. Among the 177 categorical respondents, a total of 94 different themes were identified, and a total of 326 themes were coded across the 177 respondents. Table 5-9 shows all themes that were mentioned by more than one respondent. The full list of themes is available through the CSAFE data portal (see [70]).

Number	Theme	Frequency
1.	uncertainty	25
2.	quantification	19
3.	jury confusion	17
4.	probability not ready	17
5.	random match probability	16
6.	weight of evidence	15
7.	likelihood ratio inverted conditional	13
8.	scientific	11
9.	likelihood ratio	9
10.	confusing for customer	7
11.	undermining of fingerprinting	7
12.	continuum	6
13.	incomplete	6
14.	vague	6
15.	confusing	5
16.	confusing to practitioners	5
17.	don't know	5
18.	sliding scale	5
19.	too weak	5
20.	underselling	5
21.	unnecessary	5
22.	DNA model	4
23.	driven by critics/self-serving agenda	4
24.	error prone	4
25.	statistical model	4
26.	the way forward	4
27.	disaster	3
28.	misleading	3
29.	more data to support conclusion	3
30.	possible associations	3
31.	problematic	3

32.	anti-ground truth	2
33.	consider with other evidence	2
34.	dropping exclusion of all others	2
35.	expectation of seeing similarity	2
36.	greater exploitation of evidence	2
37.	imprecise	2
38.	inappropriate	2
39.	insufficient weight	2
40.	more accurate	2
41.	not opposite of categorical	2
42.	objective	2
43.	probability impossible	2
44.	protects incompetence	2
45.	reliance on technology	2
46.	score-based	2
47.	transparent	2
48.	wiggle room for witnesses	2
49.	wordy	2
50.	world population paradigm	2

Table 5-9. Describe probabilistic reporting – Categorical respondents: all coded themes (n=177).

The most common response characterized probabilistic reporting as the quantification, or at least the communication, of "uncertainty":

> *Probabilistic reporting assigns uncertainty to each examination while non-probabilistic reporting offers a conclusion (16).*

The second most common response characterized probabilistic reporting in terms of "quantification":

> *Using numbers, statistics, and/or frequencies to explain a conclusion rather than words or descriptions (181).*

As shown in Table 5-9, two of the more common themes that appeared in response to questions about the appropriateness of probabilistic reporting also appeared in response to questions about its definition: "jury confusion" and "probability not ready."

A number of specific technical statistical concepts appeared in the responses. The notion of "random match probability" appeared more frequently than "weight of evidence." This is notable because the latter of the two terms better captures the current thinking among forensic statisticians, especially with regard to pattern evidence, such as friction ridge analysis. For example, many respondents described probabilistic reporting in terms of a random match probability:

> *I would describe it as using a statistic [sic] model to convey the likelihood of finding someone else with the same characteristics in those prints (25).*

Slightly fewer described the weight of evidence:

> *Probabilistic reporting would involve some sort of calculation to give a weight to the conclusion based on the information present in the known and unknown (180).*

The likelihood ratio, which is currently a very common topic of discussion in forensic statistics, was mentioned more frequently than either random match probability or weight of evidence. However, it was more common to describe the likelihood ratio incorrectly, with the conditional inverted [101], than it was to describe it correctly. For example, one respondent correctly characterized the likelihood ratio:

> *Probabilistic reporting puts a number on the result. In some models, that number is a similarity score, sort of like the scores we get when we search a print in AFIS. In other models, the number is a likelihood ratio, telling you how likely it is that the prints having so many features in common come from the same source versus how likely it is that prints having those features in common come from a different source (20).*

But others described it with the conditional inverted:

> *Instead of just saying "identification" which would be categorical reporting, probabilistic reporting would include the likelihood of the latent print being made by the subject (189).*

5.6 Discussion

Participant Responses

The usable response rate of 17.7% (301 out of approximately 1,700 IAI members invited to participate) poses a limitation on our survey. Viewed in combination with the recruitment scheme and voluntary participation, it is difficult to know whether these responses were representative of the views of the latent print examiner population at large. However, the demographics of our participants hint that our participants were not an unusual subset of the IAI membership. Most of our participants were mid-career latent print examiners between the ages of 30 and 50 years with testimony experience.

Current Reporting Practices

Probabilistic reporting appears to remain rare in the friction ridge discipline, despite its adoption by a small number of FSPs and practitioners. Approximately 98% of friction ridge examiners surveyed report categorically using terms and phrases that are reminiscent of over a century of practice. This is consistent with the findings of other studies in a variety of disciplines. In their analysis of mock forensic reports from proficiency tests, Bali *et al.* [80] found categorical

statements in 100% of toolmark reports, 100% of fiber analysis reports, 98% of firearms examination reports, 97% of glass analysis reports, 93% of questioned documents reports, 87% of handwriting examination reports, 85% of paint analysis reports, and 79% of shoeprint impression reports. Across all disciplines, 94% of reports were categorical.[15] Morrison *et al.* [102] found that categorical reporting was by far the most common for speaker identification results. One interesting observation from our survey is that approximately 80% of the examiners surveyed who claimed to be reporting probabilistically gave as examples statements that were actually categorical. This suggests some examiners may have a false belief that they are reporting probabilistically and therefore may not recognize many of the concerns over categorical reporting are applicable to them.

Even among those who report categorically, fewer than 10% appear to have adopted the "elaborated" approach espoused by SWGFAST as early as 2011. Nearly a decade of calls by the scientific community to move toward probabilistic reporting seem to have had limited impact. However, as noted above, approximately one-third of respondents report categorically but responded in ways that suggest they personally support probabilistic reporting. This indicates that some examiners may be "captured" by agency policy—reporting in a manner dictated by agency policy rather than by personal belief. This raises several open questions. Why is there such a large discrepancy between examiners who support the idea of probabilistic reporting and those who actually practice it? Are there some examiners whose reporting is inconsistent with what they personally believe is appropriate? Or did these responses merely reflect the respondents' perception that probabilistic reporting was the "socially desirable" answer?

Attitudes Toward Probabilistic Reporting

Most respondents perceived probabilistic reporting as inappropriate for friction ridge analysis. Only around one third of our respondents described a shift toward probabilistic reporting as appropriate. At the same time, it can reasonably be argued that finding approximately one-third of friction ridge practitioners described probabilistic reporting as "appropriate" would have been unthinkable as recently as one decade ago, let alone two or three decades.

Longstanding "myths" about friction ridge evidence—for example, the claim that the "uniqueness" of friction ridge skin eliminates uncertainty in associations between friction ridge impressions—did appear in our data, and so they cannot be considered "dead" and still lurk behind the scenes [6]. It should be emphasized, however, that they were uncommon among our survey respondent population, who—precisely because they took the time to complete the survey—may reasonably be presumed to more aware of, and interested in, current debates and developments within the discipline.

In assessing the reasons respondents offered for their attitudes, it may be helpful to distinguish between what we might call "consumption" issues—relating to how evidence is used by other criminal justice system actors—and what we might call "technical" issues concerning the

[15] It should be noted, however, that Bali *et al.* [80] coded categorical statements "absent" or "present"; thus, a single forensic report could contain both a categorical and a probabilistic statement. We, in contrast, coded "categorical" and "probabilistic" mutually exclusively: a single statement could not be both probabilistic and categorical.

merits of probabilistic reporting itself. Among the majority of respondents who viewed the shift toward probabilistic reporting as inappropriate, the degree of concern about "consumption" issues—fact-finder comprehension, prosecutor interests, and defense exploitation—is conspicuous. For example, from the data in Table 5-8, we see that the primary reason respondents opposed probabilistic reporting was the concern that defense attorneys will take advantage of the probabilistic conclusion to sow reasonable doubt (79%). Respondents further supported this through a number of free text responses, such as:

> *I believe it will confuse the jury and give the defense a chance to place reasonable doubt* (425).
>
> *I don't know why we would be testifying to 'probable' outcomes—it would make cross examination significantly more difficult for the expert witness* (310).
>
> *The conclusion that would essentially replace 'identification', in my opinion, could easily be misunderstood by jurors as meaning that there is reasonable doubt in the 'same source' conclusion* (409).

This was closely followed by the concerns that judges, jurors, law enforcement, and other investigators would not understand or know how to interpret the probabilistic conclusion language (79% and 69%, respectively). Similarly, the leading free-text reason for probabilistic reporting being inappropriate was "jury confusion," a related "consumption" issue that continues to be the subject of on-going research [74, 103-108]. When considering various options for reporting, it seems reasonable for examiners to express concern whether consumers of those reports are able to take appropriate action based on the information. However, whether the concerns are warranted depend on how "confusion" is being defined. Several of the free-text responses suggest those respondents who view probabilistic reporting as inappropriate on the basis of "juror confusion" do so over claims that jurors want a binary answer of "yes" or "no," probabilistic language seems "wishy-washy," and expressing a probabilistic view does not align with the examiner's belief that the two impressions were made by the same source. For example, some respondents stated:

> *Why would we be needed if we used the term probabilistic. That means probably him or her. SO if we say that then there is no need for fingerprint examiners. Then I guess you could say he probably wasn't there* (199).
>
> *I think the matter is being overcomplicated and we are watering down our testimony, which is a disservice to the victims* (56).
>
> *Meaningless to jury, not accurate, wishy-washy* (222).

This leads us to question whether "juror confusion" is being used as another way of expressing concern that jurors might not place as much weight on the conclusion as the examiner believes

they should (compared to a traditional categorical identification conclusion).[16] For example, one respondent stated:

> *Probabilistic Identification is Deceitful. The purpose of having Expert witness is so they can state their belief; not force the blame on to the Jury* (334).

The respondent's invocation of the notion of "blame" is interesting. It recalls Wells's [109] remark that fingerprint examiners' testimony that their conclusion the person of interest is the source of an unknown impression may be particularly persuasive to fact-finders because it satisfies a "bidirectional test": if the ultimate fact is wrong, the evidence must be wrong as well. As Wells notes, this mode of reporting evidence shifts the moral hazard of legal decision-making from the fact-finder to the expert. If the conclusion is wrong, the fact-finder can reason that they were misled by the expert. Wells found that fact-finders prefer such evidence even to statistically equivalent evidence in which the moral hazard is not assumed by the expert—the expert simply states the evidence, rather than the probability of the ultimate fact (i.e., the posterior probability). The respondent above was not merely willing to assume the moral hazard of the evidence; they appeared to perceive it as "deceitful" *not* to. This position is in marked contrast to a common argument in forensic science which insists that experts should report only the evidence [110]. However, it is consistent with our anecdotal impressions that many forensic experts genuinely believe that "the legal system" prefers, or even requires, them to state their beliefs about posterior probabilities, rather than confining their reports to just the evidence. This may reflect a belief that the expert's posteriors are superior to the fact-finder's, or it may simply reflect a sense of obligation.

Put in simple terms, respondents were less concerned that probabilistic reporting was "wrong"—although there was certainly a significant number of respondents who espoused that view—than they were that defense attorneys would take advantage of uncertainty or that it would mislead, or be misunderstood by, other criminal justice system actors, such as jurors, judges, attorneys, and police investigators. Arguably, such concerns are external to the discipline of friction ridge analysis and belong the realm of policy, rather than science. Admittedly, this is a complicated issue and raises questions as to the extent institutional factors could be at odds with scientific advancements for the forensic sciences. It could be argued that this is a second-order problem and one that the forensic scientist need not necessarily face alone. Instead, the "consumption problem" can reasonably be construed as falling within the purview not solely of forensic scientists, but perhaps also of lawyers and legal scholars, social scientists, educators, and policy makers. Although it is clearly a barrier to probabilistic reporting, whether it *should* be a barrier to probabilistic reporting and how to mediate practitioners' concerns over these issues remain open questions.

Also of interest is the concern that "probabilistic reporting will cause the number of erroneous associations to significantly increase" shared by just over a quarter of the practitioner community. This refers to an issue that has long lurked behind proposals for probabilistic reporting: that it would enable the use of *more* friction ridge evidence, not less, albeit evidence of

[16] Although, it is unclear how much weight *should* be placed on the conclusion given traditional categorical statements of single-source attribution have been criticized as unsupportable by individual scientists [37, 45, 82-89] and a number of governmental and scientific reports [3, 7-9, 90].

more marginal value [45]. In this sense, respondents who agreed with this prompt may have been expressing anxiety about the transition from the comforts of the categorical regime to the discomforts of the probabilistic. Under the categorical regime, experts could feel reassured by the belief that the categorical regime was ostensibly conservative: evidence whose strength was less than overwhelming was ruthlessly discarded, and thus it was believed that it was unlikely that comparisons that fell into the "identification" category would result in error. In a probabilistic regime, however, all evidence, no matter how marginal, can in theory be reported. While these reports will be "truer" to the weight of the evidence and potentially provide more information for investigators and courts to consider, if fact-finders interpret them as equivalent to the old categorical terms (e.g., "identification,") they could interpret the evidence as having *more* weight than was intended by the examiner, thus creating an opportunity for fact-finders to, for example, improperly infer that a person of interest is *the* source of an impression. Indeed, some free-text responses also touched on what we interpret to be this concern:

> *I think it creates room for errors* (326)
>
> *I am concerned with using probabilities because I think that it will wrongfully convict someone based on a % of something that I wouldn't consider reliable as a print examiner* (179).
>
> *I believe probabilistic language will confuse and mislead the jury. I think there will be a spike in wrongful convictions* (393).

In contrast to the "consumption" issues discussed above, only 15% of examiners reported that they were motivated by concerns that their certification would be placed in jeopardy if they were to report probabilistically. Feedback on this source of opposition was solicited to evaluate the influence of the historical policy of the IAI that codified longstanding culture at the time and formally remained in effect for over 30 years. Between 1979 and 2010, the IAI officially opposed any testimony or reporting of "possible, probable, or likely friction ridge identification" with the threat of formal charges of "conduct unbecoming" and revocation of professional certification [111-113]. This formal opposition to probabilistic reporting has undoubtedly shaped the perspectives of many experienced practitioners. However, it appears to be less important to our respondents than the "consumption" issues.

Characterizations of Probabilistic Reporting

Our findings suggest that examiners' characterizations of probabilistic reporting are quite diverse. A review of Table 5-9 suggests that survey participants fell into two broad categories in responding to this question. A significant number responded in what we might call "technical" terms. They responded by describing back to us what, technically, probabilistic reporting is: quantification, uncertainty, "weight of evidence" and so on. Some used specific types of measures, such as random match probability and likelihood ratio. Given the current discourse around forensic interpretation in, for example, this journal, it is interesting that "Bayes' Theorem" was only mentioned once.

A second group responded to the question by treating "probabilistic reporting" as an institutional phenomenon. They described back to us what the move toward probabilistic represented within the context of institutional debates over the role of forensic science in the criminal justice system. This elicited a number of what we might call "skeptical," and at times even cynical responses, e.g., "undermining of fingerprinting," "unnecessary," "driven by critics/self-serving agenda," "disaster," "misleading," "problematic," "imprecise," "protects incompetence," "wiggle room for witnesses," "wordy," "baseless," "guessing," "hysteria," "massive undertaking," "meaningless," "not a panacea for error," "quantification for its own sake," "scary to most examiners," "scientific veneer," "something we don't do," "threatening to examiners careers," "uncharted territory," and "unhelpful." One respondent described probabilistic reporting simply by: "Mt. Everest--we are going to lose many." While we admit that we are not entirely certain how to interpret that remark, it captures the general tone of anxiety that pervaded many responses well enough that we feel it makes an appropriate title for this paper.

However, the same prompt also elicited many positive responses, e.g., "the way forward," "transparent," "appropriate," "natural evolution of science," "safer," and "tool." In retrospect, we can see that the open-endedness of this prompt allowed respondents to choose whether to respond in technical or non-technical terms and, to some extent, invited them to editorialize. As with the responses to our other open-ended questions, one important observation is the heterogeneity of perspectives within the friction ridge discipline—not only as it relates to what it means technically to report probabilistically, but also how examiners have characterized it as a concept. These findings are important as they suggest that even among those who might be welcoming of probabilistic reporting, there are many different perspectives as to what it means and how it might be accomplished. We note, however, that it is possible respondents could have been primed by previous survey questions, so we cannot be sure that these responses truly reflect the respondents' unadulterated opinions.

5.7 Conclusion

The purpose of this survey was to provide proponents of probabilistic reporting with a sense of the state of progress in one important forensic discipline: friction ridge examination. We found that probabilistic reporting has not been widely adopted and remains extremely rare. Among those who responded to our survey, 98% of respondents continue to report categorically with explicit or implicit statements of certainty. Although we found that approximately one-third of respondents evinced receptivity to probabilistic reporting, which may well represent a more receptive audience than some might have expected, we also found that significant resistance to probabilistic reporting remains across the discipline.

The most common reasons for opposition to probabilistic reporting, shared by approximately 80% of respondents, were that defense attorneys would take advantage of the uncertainties as a litigation strategy and that probabilistic language would mislead, or be misunderstood by, other criminal justice system actors, such as jurors, judges, attorneys, and police investigators, respectively. Free-text responses related to their opposition were diverse and not limited to issues of whether probabilistic reporting is scientifically more appropriate. In fact, some respondents acknowledged probabilistic reporting may be more scientifically appropriate yet

continued to defend traditional categorical reporting practices. Rather, attitudes toward probabilistic reporting appear to be influenced by educational, philosophical, psychological, and complex judicial implications and longstanding cultural and institutional norms.

For forensic statisticians looking for guidance, we believe our findings offer three useful lessons. First, we would emphasize the sheer heterogeneity of the responses found in, e.g., Tables 5-5 through 5-7. Practitioners' perspectives, even on a narrowly framed issue such as probabilistic reporting for a single forensic discipline, are quite varied and complex. This will present a challenge to educators and trainers. They will not face a handful of widely held "myths" that they need to debunk or perspectives that they need to realign. Instead, they will face a diverse array of strongly held opinions about what, if anything, ails the friction ridge discipline, how it can and should be improved, and to what extent statistics offers a solution to those problems or would be the cause for other problems. Probabilistic reporting in latent print examination is not a "bi-partisan" issue; it is more complicated than that.

Second, we would direct readers' attention to our finding that, what we have called "consumption" issues, seem to dominate respondents' attitudes toward probabilistic reporting. This suggests something important about where our respondents perceive the boundaries of their role as experts to lie. Our respondents appear to believe that as experts, they are responsible not only for the evaluation and articulation of the evidence, but also for how that evidence will be consumed by litigators and the decisions about the evidence that *they believe* fact-finders *should* be making.

We believe the appropriate role of the expert is narrower. We believe that litigation strategies and juror concerns are not within the remit of the forensic scientist to do forensic science properly. Rather, we believe the role of the expert is to educate the fact-finder about the evidence and report their findings within the limits of what the science can support, but leave it to litigators to fit that information into their arguments and to fact-finders to weigh that information when making their ultimate decisions. This requires experts to neutrally represent the evidence and clearly articulate the strengths and limitations related to those findings so that fact-finders can make an informed decision, but resist the temptation to "simplify" the evidence for fact-finder consumption (especially when such simplifying would entail rounding the probative value of the evidence up, as when a strong belief that two impressions derive from the same source is expressed as "the impressions originate from the same source"). To be sure, we are not denying the importance of fact-finder comprehension of statistical evidence, which is well understood to be an important problem and is the subject of a wealth of research. However, we are surprised at the degree to which bench practitioners seem to understand it be *their* problem as opposed to a problem for legal actors, psychologists, policy makers, etc. This puts statisticians in a bind because many practitioners seem to view probabilistic reporting inappropriate not because it is an incorrect way to report evidence, but rather because fact-finders have difficulty understanding statistics. The latter point is undoubtedly true, but it also may well be an intractable problem. In this sense, concerns about consumption can begin to seem like stalling tactics.

From our findings, few of the respondents appear to share this view. This is most evident in how they approached several of the open-ended questions. When asked about the scientific merits of probabilistic reporting over categorical reporting or how they characterize probabilistic,

compared to categorical, reporting they responded by discussing how such reporting language could be (mis)used in court. This presents another challenge. If experts are expected to report their findings within appropriate limits, the role of the expert will need to be clarified by policy makers and enforced by the judiciary. As long as experts are allowed to express their opinions categorically, they will continue to do so. Proponents of probabilistic reporting, therefore, will need to not only include forensic statisticians, educators, and trainers devising statistical tools and recommending reporting frameworks for experts, but also policy makers and members of the judiciary to require it of experts and enforce it during litigation.

Third, our survey suggests that many respondents do not share a common understanding of what is meant by the term "probabilistic reporting." We would point to the high number of respondents who claimed to be reporting probabilistically but, in fact, were not. The survey results offer forensic statisticians ample further reasons for pessimism about their educational efforts thus far: the insistence that uncertainty can be eliminated in friction ridge analysis, the claim that quantification of friction ridges is impossible, the claim that a statistical model is "impossible" for the same reason claims of certainty are impossible, and the skepticism and mistrust directed toward statistics as a discipline. We might even go a step further and suggest that many respondents did not understand what it means to report probabilistically or why categorical reporting, as we define it, has come under so much criticism. However, this was not unexpected. When prompted, almost half of our respondents (44%) acknowledged that they did not feel they sufficiently understand probabilities and would not be able to properly testify. This was further evident in some of the free-text responses to other questions. For example:

> *Probabilistic language is currently very confusing to me and I would need a lot more training and understanding of it, before I would be comfortable putting it in a report and testifying to it in a courtroom setting* (117).
>
> *I agree that the shift toward probabilistic language is appropriate but I still don't fully understand the impact at this time and have had no training on the subject of probabilistic language yet* (108).
>
> *For the latent examiner not as comfortable with explaining statistics and probability, it could open the door for the attorney to discredit the examiner. That is not only problematic for the case it could be detrimental for the examiner's career* (103).

We are sympathetic to this concern because it has never been a formal requirement for practitioners to have any background knowledge in principles of probabilities and statistics. This presents yet another challenge: if practitioners are expected to testify using probabilistic language, it will require a coordinated investment by forensic science administrators, educators, and trainers to ensure practitioners have the fundamental education and training on probabilistic and statistical principles so that they understand what they are reporting and feel comfortable and confident in their own knowledge on the subject. For example, as one respondent suggested:

> *As a scientist, I understand why the community cannot testify to absolutes in terms of "identification." The reasoning behind that argument is sound. However, I think it will take a change in the dogma of the science to get practitioners to 1) understand probabilities*

> *and what they're actually reporting (and that their ultimate conclusion is not actually changing), and 2) want to change how they testify/report their findings (because they are accustomed to "this is how I've always done it")* (30).

It will also require outreach to attorneys and judges so that they understand the transition and what it means. This will require more than a mere policy-change; it will require a commitment by forensic science administrators, educators, trainers, practitioners, and policy makers to address the foundational gaps in education and training curricula as well as establish operational environments that are conducive to allowing practitioners to explore what it means to report probabilistically and how to do so in a way that they are comfortable with. This is important because the transition to probabilistic reporting should not be done in haste. Fortunately, though, probabilistic reporting, as we define it, does not necessarily require the use of numerical quantities, algorithms and other statistical tools, or the reporting of evidence along a full continuum, although those measures are preferred by some. Instead, it can be achieved by simply avoiding claims that an individual is *the* source of an impression or using terms that imply certainty for single-source attribution. As practitioners and stakeholders gain comfort with reporting using probabilistic language for comparisons which would normally be categorized as "identification" under traditional categorical reporting schemes, it can be expanded along the continuum to include more marginal comparisons that still provide useful information, but do not warrant stronger conclusions. The use of numerical quantities, algorithms, and other statistical tools will provide more precise information related to the strength of the evidence, but this transition need not be done in a single act nor contingent upon the availability of such technologies.

6 Evaluation of Stakeholders' Perspectives

This chapter presents a manuscript entitled "Probabilistic Reporting and Algorithms in Forensic Science: Stakeholder Perspectives within the American Criminal Justice System" (Swofford & Champod, 2022) [52] published in *Forensic Science International: Synergy* that explores perspectives from key criminal justice stakeholders (laboratory managers, prosecuting attorneys, defense attorneys, judges, and academic scientists and scholars) related to interpretation and reporting practices (with or without algorithmic tools) and the use of computational algorithms in legal settings. Stakeholders' perspectives are evaluated qualitatively from semi-structured interviews.

Probabilistic Reporting and Algorithms in Forensic Science:
Stakeholder Perspectives within the American Criminal Justice System

Swofford, H. and Champod, C.

School of Criminal Justice, Forensic Science Institute, University of Lausanne, Switzerland

6.1 Abstract

In recent years, there have been increased efforts to promote probabilistic reporting and the use of computational algorithms across several forensic science disciplines. Reactions to these efforts have been mixed—some stakeholders argue they promote greater scientific rigor whereas others argue that the opacity of algorithmic tools makes it challenging to meaningfully scrutinize the evidence presented against a defendant resulting from these systems. Consequently, the forensic community has been left with no clear direction on how to navigate these mounting concerns as each proposed solution seemingly has countervailing benefits and risks. In order to explore these issues in greater depth and provide a foundation for a path forward, this study draws on one-on-one semi-structured interviews with fifteen participants to elicit the perspectives of key criminal justice stakeholders, including forensic laboratory managers, prosecuting attorneys, defense attorneys, judges, and other academic scientists and scholars, on issues related to interpretation and reporting practices and the use of computational algorithms in forensic science within the American legal system.

Keywords: Forensic science, Pattern evidence, Probabilities, Statistics, Algorithms

6.2 Introduction

Forensic science has long been considered a cornerstone for advancing investigations and establishing facts in question to support criminal and civil litigation. Under the powerful aura of science, interpretations and conclusions made by forensic experts are often presented as tantamount to fact—the silent witness—that courts can rely on in their pursuit of justice. For decades on end, forensic evidence was broadly considered infallible and rarely questioned. In February 2009, however, that all changed with the release of the National Research Council's

(NRC) report on the needs of the forensic science community, highlighting that "[t]he law's greatest dilemma in its heavy reliance on forensic evidence, however, concerns the question of whether—and to what extent—there is *science* in any given forensic science discipline" [3]. Following their analysis of several forensic science disciplines, the NRC noted: "The simple reality is that the interpretation of forensic evidence is not always based on scientific studies to determine its validity. This is a serious problem. Although research has been done in some disciplines, there is a notable dearth of peer-reviewed, published studies establishing the scientific bases and validity of many forensic methods." The NRC goes on to assert "no forensic method other than nuclear DNA analysis has been rigorously shown to have the capacity to consistently and with a high degree of certainty support conclusions about 'individualization' (more commonly known as 'matching' of an unknown item of evidence to a specific known source)" [3]. The NRC report, although positive in the sense that it raised awareness of the need for greater resources, offered damning critiques to a body of evidence that was often presented, and perceived, as essentially infallible.

In the years that followed, these types of critiques have become commonplace—particularly as it relates to concerns over the high reliance on subjectivity and lack of statistical foundations supporting the interpretation of results, as well as concerns over the expression of conclusions asserting a level of certainty that implies infallibility. For example, in 2012 a committee supported by the National Institute of Standards and Technology (NIST) and the National Institute of Justice (NIJ) issued several recommendations specific to improving friction ridge examinations, claiming: "Because empirical evidence and statistical reasoning do not support a source attribution to the exclusion of all other individuals in the world, latent print examiners should not report or testify, directly or by implication, to a source attribution to the exclusion of all others in the world" [9]. This was followed by another landmark report offered by the President's Council of Advisors on Science and Technology (PCAST) in 2016, asserting: "Statements claiming or implying greater certainty than can be demonstrated by empirical evidence are scientifically invalid. Forensic examiners should therefore report their findings with clarity and restraint, explaining in each case that the fact that two samples satisfy a method's criteria for a proposed match does not necessarily imply that the samples come from a common source. … [C]ourts should never permit scientifically indefensible claims" [7]. Finally, in 2017, the friction ridge community was faced with, yet again, another critique, but this time coming from the American Association for the Advancement of Science (AAAS)—the world's largest scientific society. Following a scientific gap assessment of the research supporting the existing methods, the AAAS committee stated: "Examiners should be careful not to make statements in reports or testimony that exaggerate the certainty of their conclusions. ... [T]hey should avoid statements that claim or imply that the pool of possible sources is limited to a single person. Terms like 'match,' 'identification,' 'individualization,' and their synonyms, imply more than the science can sustain" [8].

In light of these concerns, increasing calls have been made for the introduction of probabilistic reasoning and the use of validated statistical methods into forensic practice—particularly in the pattern evidence disciplines—to formally recognize and articulate the uncertainties inherent in forensic interpretation and to reduce the heavy reliance on subjective judgment [3, 7-9]. Over the years, a number of reputable efforts have been made by researchers to explore the optimal approach for expressing forensic conclusions to maximize lay fact-finders'

interpretation (e.g., see [108]) and, in the friction ridge discipline in particular, to introduce probabilistic models—often through computational algorithms[17]—to provide statistical foundations to the analysis and evaluation of evidence [17-43, 50]. Although probabilistic reporting is often presented as a scientifically superior approach to expressing forensic results compared to traditional categorical assertions, it is often more difficult for lay fact-finders to interpret [108]. Likewise, although algorithmic tools generally possess remarkable potential to provide advanced scientific capabilities and promote more objective foundations to the evaluation of forensic evidence, they often do so at the cost of transparency and explainability [114-120], which have been argued to stifle meaningful scrutiny and accountability of the evidence resulting from these tools thereby infringing on criminal defendants' Constitutional rights (e.g., see [114, 115, 117, 118]). Consequently, the forensic community has been left with no clear path forward on how to navigate these mounting concerns as each proposed solution seemingly has countervailing benefits and risks. In recent work, we began to explore some of these issues in greater detail based on perspectives that have been raised in the literature thus far and provided some initial recommendations relating to the operational implementation of computational algorithms [53]. This current study further explores those issues with greater breadth and depth, but it is only a start to what we consider to be a much needed, and much more extensive, discussion on these issues so that the forensic and legal communities can begin addressing these challenges that are no longer over the horizon.

As the forensic community continues to grapple with these issues, widespread reform efforts have been understandably slow. However, a few notable steps have been taken in an effort to heed the recommendations from various scientific committees. In 2015, the United States Army Criminal Investigation Laboratory (USACIL), the primary forensic laboratory supporting the criminal investigative mission of the Department of Defense, announced a policy change to abandon the term "identification" and report their findings in a probabilistic framework (albeit in the absence of a computational algorithm) [95]. In 2017, the USACIL went a step further and announced the implementation of a statistical software application, *FRStat*, to provide statistical support to fingerprint associations [50, 96]. This has been considered by some as a step in the right direction to reduce variability and improve overall consistency between analysts (e.g., [121, 122]). Then, in 2018, the Organization of Scientific Area Committees (OSAC) for Forensic Science, Friction Ridge Subcommittee (OSAC FRS), which is responsible for the promulgation of standards and best practices related to the forensic examination of friction ridge skin impression evidence throughout the United States, released the proposed standard for Friction Ridge Examination Conclusions [97], taking an additional step toward promoting probabilistic expressions on a national level. While the proposed standard maintains the term "identification," which has traditionally been used to express categorical conclusions, it was redefined in a probabilistic framework as a qualitative (non-numeric) expression of a likelihood ratio. In addition to the revised definition, the OSAC FRS stated that "an examiner shall not assert that a source identification is the conclusion that two impressions were made by the same source or imply an individualization to the exclusion of all other sources" [97], a claim which has been a common hallmark of categorical statements.

[17] The term "computational algorithms" refers to automated or semi-automated computer implementable processes designed to compute mathematical outputs for purposes such as forecasting, predictions, statistical evaluations and decision making. For purposes of this paper, the term "algorithm" and "computational algorithm" are synonymous. The term "algorithmic tools" refer to devices enabling the applications of computational algorithms.

Despite these efforts, probabilistic reporting and statistical interventions continue to be a contentious topic within the forensic science community, with some forensic friction ridge practitioners welcoming it with open arms as a more "scientifically defensible" approach while others express passive skepticism or outright opposition [51]. Although significant resistance remains across the friction ridge discipline and probabilistic reporting remains rare, approximately one-third of survey participants who currently report categorically seem to be receptive to the idea of reporting probabilistically, but remain hesitant to adopt for one reason or another [51]. Practitioners' perspectives have been instrumental in highlighting a number of social scientific issues that are believed to have contributed to this hesitancy (i.e., educational, philosophical, psychological and complex judicial implications and longstanding cultural and institutional norms) thereby allowing us to consider strategies to address their concerns [51]. While forensic practitioners will ultimately be responsible for implementing the proposed solutions, it would be incomplete to focus *solely* on perspectives of forensic practitioners.

To fully understand the issues and more effectively facilitate improvements to traditional practices, we must also account for the perspectives of *all* stakeholders within the criminal justice system—not just forensic practitioners. Recognizing that prior work has captured the broad perspectives of friction ridge practitioners (i.e., [51]), this study aims to explore the individual perspectives of other key criminal justice stakeholders based on their different roles in the criminal justice system—including forensic laboratory managers, prosecuting attorneys, defense attorneys, judges, and other academic scientists and scholars—to provide a better understanding of their distinct values and interests on issues related to: (i) interpretation and reporting practices (with or without algorithmic tools) and (ii) the implications of the use of algorithms in legal settings as a means of calculating the probabilistic values assigned to the evidence.

6.3 Materials & Methods

This study was conducted as one-on-one semi-structured interviews between the first author and each individual stakeholder using the video-based virtual meeting platform Zoom®. Although the qualitative nature of this approach prohibits broad generalizations and quantitative representations, it does allow us to explore these various perspectives in greater depth and with more clarity than if it were presented as a structured survey. Participants were solicited by invitation (see Appendix E-1) based on having been actively engaged in issues concerning forensic science policies, procedures, and practices. These participants have occupied prominent roles in their disciplines (e.g., senior and executive level positions in their organizations and professional societies), have been selected to serve on boards and committees steering policy and practice recommendations (e.g., National Commission on Forensic Science, Organization of Scientific Area Committees for Forensic Science), have made academic contributions to forensic science practices through professional publications and presentation, or have influenced the practices of others across the broader community, either directly through supervision or indirectly through training and continuing education activities. Overall, a total of twenty-two individuals were invited to participate in the study and seven individuals declined to participate (four individuals did not respond to the invitation [one forensic laboratory manager, one prosecuting attorney, and two judges], two individuals cited competing priorities and commitments to participate within the

intended timeframe [one forensic laboratory manager and one judge], and one individual expressed support for the study but felt unable to answer the questions related to the use of algorithms [academic scholar]). Invitations were extended to potential participants until three individuals agreed to participate for each stakeholder group (forensic laboratory managers, prosecuting attorneys, defense attorneys, judges, and other academic scientists and scholars) resulting in a total of fifteen participants. Specific details related to the backgrounds and experiences for those individuals who agreed to participate are provided in the Results section for each stakeholder group.

Interviews were conducted between September and November 2021 and were scheduled based on participants' availability, thereby enabling an arbitrary sequence of participants (i.e., stakeholder participants were arbitrarily spread throughout and not interviewed in any particular sequence). Participants' personal identities are not disclosed or publicly attributed to any specific statements. Each participant was assigned a unique identifier within their stakeholder group to distinguish among responses from individual participants. Prior to the study commencing and as part of the initial invitation, participants were provided an Information and Informed Consent sheet that summarized the structure of the study (see Appendix E-2), a summary of the purpose and background of the study that included specific terms and definitions related to the interview questionnaire (see Appendix E-3), and a general outline along with a set of structured questions to guide the interview (see Appendix E-4).

Participants were first presented with a series of questions pertaining to their demographics (occupation, experience, education, and exposure to algorithms). Participants were then asked a series of structured questions addressing various topics (described below) pertaining to their perspectives related to interpretation and reporting and the use of computational algorithms for court purposes. Although most participants offered responses to all of the structured questions, in a few instances some questions were omitted during the interviews due to time constraints; thus, not every participant provided a separate response to each individual question. Throughout the interview, unstructured questions were raised *ad hoc* to explore participants' responses in further detail and to elicit their perspectives related to responses provided by other participants interviewed thus far.

Questions related to the broader issue of interpretation and reporting sought to elicit participants' perspectives around four broad topics:

The first topic focuses on the validity, appropriateness, benefits, and limitations/risks of categorical reporting compared to probabilistic reporting methods. These concepts have become central to the broader discourse concerning how forensic science testimony should be delivered and have been at the forefront of the friction ridge discipline for over a decade (e.g., see [3, 7-9, 51])—often resulting in heated debates within the forensic practitioner community [51].

The second topic points to salient concerns raised by friction ridge practitioners as it relates to the use of probabilistic reporting. In a recent study surveying various reasons for practitioners' opposition to probabilistic reporting, the most common concerns cited by friction ridge practitioners related to how defense attorneys might (mis)use probabilistic reporting to "create reasonable doubt" and whether jurors would understand the conclusion being conveyed [51]. The

findings from this survey raise other questions concerning the role/duties of experts as it relates to the limits of their testimony and whether, and to what extent, such factors ought to be taken into account by forensic practitioners when considering the most appropriate means of expressing forensic conclusions. In other words, should forensic practitioners focus on not only the validity and appropriateness of such claims, but also how those conclusions might factor into litigation strategies for one or both sides or be perceived by fact-finders? All these concerns are relevant, but how they should be addressed and by whom remains an open question.

The third topic focuses on whether it is necessary for forensic practitioners to disclose underpinnings or statistical data to support their testimony. This topic was motivated primarily by the PCAST argument that "[s]tatements claiming or implying greater certainty than can be demonstrated by empirical evidence are scientifically invalid" and "[n]othing—not personal experience nor professional practices—can substitute for adequate empirical demonstration of accuracy" [7]. Such claims by the PCAST suggest all forensic testimony must be accompanied by empirical foundations underpinning such claims. It also raises the question whether statistical data is meant to be the means for providing the empirical foundations. This is impactful to friction ridge practitioners, as traditional practices encourage experts to base their conclusions on "training and experience" and to couch their conclusions as an expression of their opinion rather than basing them on statistical measurements. It raises the question as to whether other stakeholder groups share the perspective suggested by the PCAST and how this might be more explicitly required in the longer term. Indeed, proposed amendments to Federal Rule 702 have been made to address "the problem of overstating results" and "emphasize that the court must focus on the expert's opinion, and must find that the opinion actually proceeds from a reliable application of the methodology" when considering the admissibility of expert testimony [123]. The full implications of such a proposal, however, remains unclear.

The fourth topic focuses on what participants view as the most significant challenges facing the pattern evidence disciplines relating to examination and reporting. This topic is intended to highlight how the pattern evidence disciplines might need to consider adapting in light of the various perspectives raised by the different stakeholders on this broader issue of interpretation and reporting.

Questions related to the broader issue of the use of computational algorithms for court purposes sought to elicit participants' perspectives around five broad topics:

The first topic focuses on the role computational algorithms should play in forensic science for court purposes along with the benefits and limitations/risks of such applications. These issues have become central to the broader discussion of responsible applications AI in society. As computational algorithms have advanced and automated decision systems have become more accessible, researchers, advocates, and policy makers are debating when and where these systems are appropriate—including particularly sensitive domains such as criminal justice [124]. Questions have been raised on how to fully assess the short and long-term impacts of these systems and the appropriateness of their applications given many operate as "black-boxes" [124]. These are broad questions for which stakeholders often disagree. In the context of forensic science, perspectives on these issues have yet to be fully explored.

The second topic focuses on the concept of "trust" with computational algorithms and what artifacts are needed for stakeholders to be comfortable with the use of an algorithmic tool. For example, is source code a necessary requisite for an algorithm to be trusted? In recent years, particularly in the context of probabilistic genotyping algorithms, courts have grappled with legal issues surrounding whether they can or should compel disclosure of source code due to countervailing positions related to trade secret violations. These issues have become a growing source of controversy affecting whether algorithms should be used in forensic science more broadly [114].

The third topic expands on the concept of "trust" and points specifically to computational algorithms based on AI/ML methods. Recognizing that source code has often been the focus of legal debates as it relates to the admissibility of algorithms based on human interpretable rules or processes, what about algorithms that are based on non-human interpretable processes, such as those developed through AI/ML methods? Computational algorithms based on AI/ML are often "black boxes" even to their developers, irrespective of the availability of source code. Given this additional layer of opacity, is it appropriate to use computational algorithms based on AI/ML methods in forensic science for court purposes? If so, under what circumstances should they be used?

The fourth topic addresses the issue of regulating computational algorithms. This issue was motivated by recently proposed legislation, the Justice in Forensic Algorithms Act of 2019, to "prohibit the use of trade secrets privileges to prevent defense access to evidence in criminal proceedings, provide for the establishment of Computational Forensic Algorithm Testing Standards and a Computational Forensic Algorithm Testing Program, and for other purposes" [125]. Among other implications of this proposed legislation, it would prohibit the use of computational forensic algorithms unless they have been tested by the Computational Forensic Algorithm Testing Program and the developers of the algorithmic tools agree to waive any and all legal claims related to the defense analyzing or testing the computational forensic software [125]. Although this proposed legislation remains early stage, it raises the question of whether computational algorithms should be regulated, and, if so, by whom and how. Is the adversarial system sufficiently positioned to regulate computational algorithms as they currently do with the admissibility of expert testimony? Should specific algorithmic tools be "approved" by an external authority prior to authorizing their use? If so, should it be administered by a government entity (federal, state or local) or other non-government institution? The issue of regulation raises several other complex questions and takes on several different dimensions that have yet to be fully explored.

The fifth topic focuses on what participants view as the most significant challenges facing the pattern evidence disciplines relating to the operational use of computational algorithms in forensic science for court purposes. This topic is intended to highlight how the pattern evidence disciplines might need to consider adapting in light of the various perspectives raised by the different stakeholders on this broader issue of the use of computational algorithms.

Interviews were recorded (audio and video) using the Zoom® virtual meeting platform. The full recording was transcribed using the Descript® transcription platform [126] using a two-stage approach. First, transcriptions were initially performed using the Descript® commercial

machine transcription software to automatically detect speakers, transcribe the audio, and align transcribed text to the audio and video [126]. Second, using the manual transcription editing features with the text, audio, and video, aligned within the Descript® platform [126], the machine transcription was reviewed by the first author to confirm accurate transcription and manually correct any errors. The transcribed interview was then exported to a Microsoft Word® document. Overall, this resulted in over twenty hours of recorded interviews and over three hundred pages of written transcripts. The transcribed text from the interviews were then qualitatively analyzed by categorizing participants' responses based on the specific topics being explored (e.g., within the broader issue of "interpretation and reporting," participants' responses that were related to the validity, appropriateness, benefits, and limitations/risks of categorical reporting were categorized separately from the other topics described earlier). Then, within the categorized responses for each participant, specific excerpts were identified that succinctly represented each participant's viewpoint. This approach allows us to capture specific comments made by individual participants *in their own words*, summarize participants' perspectives for each topic explored, and compare those perspectives both within and between the different stakeholder groups.

The perspectives of each stakeholder group are presented separately. This enables us to understand the source(s) of the different perspectives and compare those perspectives across the different stakeholder groups, which is a key objective of this study. Although all stakeholders share a common goal for an effective administration of justice, they each serve very different roles and responsibilities, and therefore may view various issues differently based on those roles. For example, forensic laboratory managers are responsible for ensuring they have the personnel, resources, and equipment to examine cases effectively and efficiently to keep pace with the growing demands and are therefore often focused on ways of increasing capacity while maintaining acceptable quality standards. Prosecuting attorneys, as legal representatives of the government, are responsible for convincing a court that a particular individual is guilty of committing the crimes that they have been charged with and are therefore often focused on presenting their arguments in a manner that is comprehensible to lay fact-finders. Defense attorneys, as legal representatives of the defendant, are responsible for defending their client's interests and rights and are therefore often focused on confronting and challenging the evidence presented against them to ensure it meets the appropriate legal standards. Judges are responsible for overseeing the legal process and are therefore often focused on ensuring that applicable rules, regulations, and laws are followed by all parties and that the integrity of the process is upheld. Finally, other scientific and academic scholars are responsible for researching complex issues and making recommendations for improving policy, procedure, or practice, and therefore are often focused on considering issues in terms of scientific or legal ideals. Understanding the different perspectives from each stakeholder group and how their interests may differ as they relate to fulfilling their specific roles and responsibilities within the criminal justice system is important for us to lay the foundation and begin to navigate a path forward on these issues that is responsive to the needs of all stakeholder groups.

In order to provide such an analysis and synthesis of these various stakeholder perspectives, we have organized the information into two distinct sections. In the Results section, we present a summary of each participant's background and experiences and responses to questions addressing key topics related to the broader issues of "interpretation and reporting practices" and "use of algorithms" within each stakeholder group. Organizing the Results of the interviews in this

manner allows us to compare the extent to which perspectives from individual participants are consistent with others *within* the same stakeholder group. In the Discussion section, we characterize the collective perspective representing each stakeholder group by topic and compare those perspectives across the different groups. Organizing the Discussion in this manner allows us to consider the extent to which perspectives may vary *between* different stakeholder groups and begin to understand the sources of those differences and lay a foundation for us to explore why those differences might exist. Throughout the Results section, we provide short specific quotes from individual participants to illustrate certain views or discussion points. While these quotes are intended to be illustrative, we recognize that some readers might desire to consider participants' statements in greater context of their responses from the interviews. Although full transcripts cannot be released to protect the anonymity of participants, in Appendices E-5 and E-6 we provide more elaborate quotes from participants related to each topic discussed in the interview. In the Discussion section, we provide a fewer set of more elaborate quotes from participants, primarily from responses to *ad hoc* questions presented to participants throughout the interviews to illustrate other interesting points.

6.4 Results

6.4.1 Laboratory Managers

Background & Experience:

Three laboratory managers participated in the study—all male. All three laboratory managers are actively working in large metropolitan jurisdictions in the United States and have between 20 and 38 years of experience in forensic science. One participant's experience is dominated by trace evidence, including physical match comparisons, shoe print, tire track, textile, hair comparisons, and fiber comparisons as well as forensic serology and DNA (LM#1). The other two participants experiences were dominated by toxicology (LM#2) and analytical chemistry (LM#3). All three participants, however, currently serve as the director for their respective laboratory system, overseeing a wide range of forensic disciplines, including DNA, drug chemistry, toxicology, fingerprints, firearms, and crime scene, among others. Participants' experiences working with algorithms are varied, and include analytical instrumentation (e.g., GCMS, LCMS, etc.), breathalyzers for breath alcohol quantitation, database searching (e.g., AFIS), imaging technologies (e.g., 3D imaging for firearms), and DNA mixture interpretation (e.g., probabilistic genotyping software). One participant (LM#3) has experience developing computer software and teaches computer science (among other courses, such as physics and chemistry) at the local college. All three participants are actively engaged in national and international professional bodies and have been vocal representatives of the needs of forensic laboratories throughout the United States.

Interpretation & Reporting Practices:

All three laboratory managers expressed the perspective that categorical reporting in pattern evidence disciplines using terms such as "Identification" or "Individualization" have the

potential to mask the uncertainty and limitations associated with the conclusion. All of the participants acknowledged that the forensic science community has historically made claims in various disciplines that were overly generalized and implied greater certainty than can be supported by the empirical evidence. However, as long as the examiners caveated the claims as being their opinion, the participants were less concerned. For example:

> *Absolutes and conclusions, I think, are probably inappropriate. I, however, do not have a problem with experts giving their opinion. I think we have very good experts. I think expertise matters. I think exposure to casework matters. I do agree with a lot of the defense experts and the academics that we need a reasonably good way to express uncertainty (LM#3).*

Participants suggested that probabilistic reporting, in theory, is superior to categorical reporting because it explicitly acknowledges the uncertainty in the conclusion; however, all three participants suggested probabilistic reporting in practice had its own pitfalls. Participants were concerned that probabilistic statements would be confusing or incorrectly interpreted by lay fact-finders or would be relied upon too heavily by fact-finders assuming the numerical references were based on empirical measurements. One participant made it clear that probabilistic statements with numbers should not be used unless it was clearly based on some empirical data source (LM#3). For example:

> *I like [numbers] because it provides [context]. On the other hand, even numbers have their limitations. … How do you throw somebody just a number and expect them to understand it? ... It's still not standalone (LM#1).*

> *From a philosophical standpoint, I think it is more appropriate. What I see though, is a hell of a lot of confusion on the part of the lay person and lawyers and juries (LM#2).*

> *I have no problem with subjective interpretations [such as] "in my experiences," [or] "is very likely," just as a subjective conclusion, but if you're going to put a number on it, I think you need to have some basis [of] where you're pulling the number from (LM#3).*

Overall, participants generally considered the benefits of categorical reporting as its simplicity and ease for fact-finders to base their decision and it provides a more holistic assessment of the examination. However, categorical reporting is "fuzzier" and can mask the uncertainty associated with the conclusion. Participants generally considered probabilistic reporting as favorable in principle. However, noting the confusion that often accompanies probabilistic references, participants were hesitant to suggest probabilistic reporting was superior in practice. Ultimately, all participants suggested applying both approaches as part of examiners' explanation of the evidence.

When responding to concerns raised by practitioners as it relates to probabilistic reporting, participants agreed with practitioners, expressing the view that probabilistic reporting would be confusing to lay fact-finders. However, participants did not consider this as a reason not to adopt probabilistic reporting. Two participants suggested the challenges would not be insurmountable (LM#1 and LM#2). The other participant was more cautious, suggesting the optimal approach

moving forward is to adopt probabilistic reporting as supplemental to traditional categorical reporting. For example:

> *Watching what I've seen happened with biology, yes, it will be confusing. Is it irrevocably confusing? No. I think everybody in the system can learn how to deal with it and how to explain it. … The practitioners are confused by it right now. But that is (1) not a reason to not go there, and (2) not an indelible absolute. The confusion will subside. The confusion will abate and people will get better about explaining it (LM#2).*

> *I think the type of testimony that we're currently giving plus this is the best model for the future (LM#3).*

Participants were also sympathetic to practitioners' expressing concerns that defense attorneys would use probabilistic reporting to create "reasonable doubt;" however, none of the participants expressed the view that it should be a reason not to consider probabilistic reporting. Rather, it represents an additional barrier that will need to be addressed by proponents of probabilistic reporting. Two of the participants considered this reaction from practitioners as reinforcement for their perspective that probabilistic reporting should not be use alone—it should always be combined with an expert opinion providing an overall conclusion (LM#1 and LM#3). The other participant expressed the view that it should not be a concern from the standpoint of being rational and neutral to the issues, but at the same time recognized the human side of practitioners and suggests that it is impractical for people to be completely divorced from the emotional aspects that motivate them to be forensic scientists to start with (LM#2).

> *The last thing I want is to put something out there that can be misused. … That's why you should have the opinion that we believe that this has a likelihood of association, then you throw in the number but you give the whole package as opposed to just reporting a number that potentially could be misinterpreted (LM#1).*

> *I think there is a huge grade of the concerns that all come back to the fear of the uncertainty … their fear is if we change this, I don't know what's going to happen on the other side of it (LM#2).*

When responding to questions raised about the role and duties of experts and the limits of their testimony, participants expressed the view that it is incumbent upon experts to convey those limitations to ensure the results are properly interpreted, and the conclusions are not overstated or understated. One participant pointed to consensus-based guidelines to drive how the results should be framed in order to ensure greater standardization across the field (LM#1). The other two participants recognized the challenges associated with conveying the limitations, suggesting there is not a straightforward solution (LM#2 and LM#3). One participant claimed the limitations should be explicit on the report so that stakeholders did not have to pull it out during testimony, although acknowledged this is a practice they have not yet implemented and are still working through how to accomplish it (LM#2). The other participant expressed frustration that courts have made it challenging to convey limitations unless they are directly asked, but even then, the participant recognized the difficulty of conveying them (LM#3). For example:

I think it is an inherent obligation on the part of the expert to convey those limitations and do the best they can trying to explain the inherent uncertainty there. … [However,] this is not saying that we have effectively managed to accomplish this, we haven't (LM#2).

I think all of us have an ethical obligation to understand the limitations of what we're saying. ... [However,] most of the time the court hearings won't allow us [to express those limitations] unless they directly ask us. ... So, articulating that uncertainty is something we're not perfect [doing] yet. But, it's also one of the reasons why we don't say to the exclusion of all others [for example] (LM#3).

When asked about whether participants find it acceptable for experts to express their opinion in court without disclosing the underpinnings or statistical data to support those opinions, all three participants strongly advised to do so; however, they also recognized it does not always come out in practice and, in some situations, suggested it may not be absolutely necessary. One participant expressed frustration that despite the laboratory's best efforts to convey those details, the legal system makes it challenging for the experts to do so during testimony (LM#2). Another participant echoed similar challenges but seemed to be more resigned to the realities of the court room environment (LM#3). For example:

I would strongly encourage they do it because I feel it makes their opinion better, stronger (LM#1).

This is one of the things that I'm finding myself getting a little bit more worked up about these days, of this issue of it was the laboratory that didn't express the extent and limitations of the testing. No, the lab is willing to do that, the lab wants to do that, all the rest of the system cut it off at the knees (LM#2).

Finally, when asked what participants would describe as the greatest challenges facing the pattern and impression evidence disciplines as it relates to examination and reporting methods, participants pointed to both cultural and resource challenges, the greatest factor being limited resources. One participant lamented that many of these scientific issues that have been at the forefront of debates seem to be trivial compared to the greater challenges of effectively managing the caseload and data management (LM#2). The other two participants referenced cultural and educational challenges (LM#1 and LM#3) as well as the inability for crime laboratories to actively engage in research given their limited resources and pressures to stay abreast of casework (LM#3). For example:

There is still a little bit of resistance that you're taking away the expertise [the experts] already have and supplanting it with something else. That, to me, I think is completely false if you agree to integrate them both together. … The other biggest reason is that [for] crime labs, it's not our mission to do research, unfortunately. I love research and it's wonderful, but we are under so much pressure to get casework done. We just don't have the time, energy or money to do it. It's unfortunate because we're really the best place to do it, but we just don't have the money to do it (LM#3).

Use of Algorithms:

Laboratory managers offered generally consistent perspectives as it relates to the use of algorithms in court and the benefits and limitations of them. All three participants expressed favorable viewpoints of using algorithms; however, participants were clear that the algorithms should be used to supplement the judgments of examiners and not to replace them. Participants recognized the value algorithms can provide by promoting greater objectivity and consistency in the results. One participant expanded on the utility of the algorithms to be a "force multiplier" to "build capacity" to help offset the limited analysts available and keep pace with caseload and throughput demands (LM#2). However, all three participants cautioned the urge to rely too heavily on the algorithms and supplant the expert, or to blindly rely on them without fully vetting them. All three participants viewed expert judgment, while subjective, as a valued asset that can account for factors that the algorithm cannot and to help interpret and convey the output of the algorithm to judicial stakeholders. For example:

> *I think that's an excellent thing to assist in better understanding why you came up with this opinion. But the danger is that people then rely too much on the number (LM#1).*

> *I think the greatest benefit on the algorithms is the relative consistency of the result case over case. … [However,] I think the biggest risk is becoming overly reliant and we just exchange the categorical certain answer from the spectacle nerd for now, an infallible algorithm (LM#2).*

When asked about concerns over how algorithms can be trusted for use in court, including issues concerning the disclosure of source code, participants largely pointed to validation. Two of the participants expressed views that source-code was unnecessary and requests for disclosure were legal tactics versus genuine efforts to evaluate the algorithm (LM#1 and LM#2); however, participants were willing to support disclosure if requested and all three participants stated they would factor source code disclosure as an element when selecting a commercial vendor. One participant took it a step further and suggested algorithms should include internal controls on every single application to help establish trust rather than simply rely on an initial validation prior to casework applications (LM#2). The third participant offered a slightly different view on these issues than the other two, expressing a stronger emphasis on disclosure. This participant, (LM#3), expressed the viewpoint that understanding the internal workings of the algorithm was key for establishing trust, and source code disclosure was a way to accomplish this. This participant pointed out that validations have limitations and, while informative and important, were not a complete substitute for understanding the innerworkings of the algorithm itself, which could be obtained through public disclosure and open explanations of the conceptual operations. For example:

> *I understand the concerns [of trust], but that just means we've got to do our job in showing these tools are valid before we actually apply them to the case. … I do believe that having appropriate validation data and showing that you don't have to see in the black box to see that it's reliable. … I think largely revealing source codes is just a tactic. ... It's a waste of time, but you know what, knock yourself out, here it is as long as it's protected (LM#1).*

The problem with validation is I don't have a perfect world [and] validation is subject to some limitations based on what I fed it. ... It doesn't mean the validations are not important. They are, but they are only black box validations. I don't know what's in the box. ... [That said,] I'm a big proponent of intellectual property, but that's not necessarily for courtroom use. ... [In] the perfect world, if you're dealing with people's lives in the courtroom, knowing everything about how decisions are made is a better approach (LM#3).

When algorithms are based on AI/ML, however, participants were receptive to the idea of using these, particularly if validation testing demonstrated superior performance. None of the participants expressed concern over the opaqueness of the algorithms and the inability to disclose source-code, provided there was adequate validation demonstrating its performance. One participant (LM#2) recognized the difficulties with truly understanding the full limits of a black box system; however, this participant's concerns were mitigated as long as "best efforts" were made to explore these issues during validation and the use of the system was confined by the limits of what was tested. Another participant (LM#3) expressed caution if the limitations are not fully understood. For example:

I can test the black box and show it's fit for purpose. ... Here's my acceptance criteria. I do my testing. It meets the criteria. It works. It's fit for purpose. ... So, you can't turn over source code, [well] I didn't really see that as being a real problem before. ... If it provides a better value of results, which I should show through my validation, my ongoing testing, I should always be picking the one that's better (LM#1).

I don't think using it is a bad thing, as long as you know the limitations. If we don't know those limitations, taking it to court then could cause more damage than good, and that's a problem. Those limitations have to be understood before it's actually used (LM#3).

When asked about regulation of algorithms, the participants recognized the need for better coordination and guidance to establish best practice and minimize duplication of efforts; however, they stopped short of suggesting full regulation. All three participants considered full-fledged regulation as potential overreach and causing other political and bureaucratic challenges. One participant considered the value of regulation, in theory, as similar to discussions around the requirement to license analysts and accredit laboratories, but questioned whether regulation of specific algorithms would work in practice (LM#2). Overall, participants seemed to express the view that regulation should come in the form of best practice recommendations and validation data that the legal system can consider within the course of case-by-case litigations. For example:

I feel that a weakness of our forensic science enterprise is that we don't have a cohesive, guidance mechanism as much as I think maybe we should. ... I think [full regulation] would probably be considered by many as an overreach, but the court system in a way should be self-regulating to a point. ... I think it's been fairly reasonable so far and I think the defense community is pretty well interconnected that when [issues] come out, they're on top of it and that information diffuses (LM#1).

I'm not sure I've got a good answer for that. ... I'd love to think [that an oversight regulatory body] was an advantage, but I've seen a lot of places where it gets to be a hindrance really quick (LM#2).

Finally, when asked what participants would describe as the greatest challenges facing the operational use of computational algorithms for court purposes, all three participants pointed to resources—specifically, resources to maintain current caseload requirements while enabling the examiners to gain the foundational training and education to fully understand the systems, validate the systems, and integrate them into day-to-day workflows. One participant (LM#2) offered a detailed description of the competing priorities and challenging decisions laboratory managers are faced with when choosing where to direct their focus. This participant went further by expanding on several other elements that would need dedicated resources to support the implementation of an algorithmic tool, such as the peripheral data management and infrastructure requirements. Another participant (LM#3) highlighted the challenges with developing the algorithms and ensuring they have the proper datasets to start with, which can be challenging given privacy issues preventing open sharing and coordination between public and private institutions. For example:

Resources. To stay on top of how quick things are developing, it's taking more and more resources. We all have backlogs and we're focusing on those. To take people off of [casework] to train them, then get these new things up to speed and implement them and then change people's minds [takes resources] (LM#1).

6.4.2 Prosecutors

Background & Experience:

Three prosecutors participated in the study—one male and two female. All prosecutors are actively working in large metropolitan jurisdictions in the United States and have between 17 and 40 years of experience litigating criminal cases involving forensic science. Each participant serves as the lead prosecutor specializing in litigating forensic science issues within their jurisdiction, including directing and training other litigators on issues related to forensic science. Participants' experiences span across a broad scope of disciplines, including both pattern evidence (e.g., fingerprints, handwriting, firearms), trace evidence (e.g., microscopy), and DNA, as well as across a range of different types of cases, such as street crime, sexual assault, and homicide. One participant expressed experience handling appeals related to forensic science all the way up to the Supreme Court. Participants' experience litigating algorithms primarily involved those related to probabilistic genotyping algorithms for DNA. Two of the three participants had experience litigating probabilistic genotyping algorithms as part of admissibility hearings. The third participant had experience litigating probabilistic genotyping algorithms "on paper" without an actual legal hearing.

Interpretation & Reporting Practices:

All three prosecutors expressed the perspective that categorical reporting in pattern evidence disciplines using terms such as "Identification" or "Individualization" was the most appropriate and preferred means of expressing conclusions and they disagreed with the claims that those terms imply "absolute certainty." Participants expressed the perspective that they are both appropriate and easily understandable. Two of the participants agreed that there should be limitations related to those claims, such as not asserting 100% certainty and "to the exclusion of all others" (P#1 and P#2); however, none of the participants expressed any reservations about forensic practitioners providing their opinion on matters related to source attribution (i.e., that a specific individual or item is the source of a questioned impression). For example:

> *I don't think saying identification implies absolute certainty (P#1).*
>
> *I don't have a problem with the use of a categorical response. It's easy to understand. It's easy for the jury to grasp, and I believe that it is the true opinion of the scientist who's giving us that opinion (P#3).*

Participants were not completely opposed to probabilistic reporting, in general, however. Participants' have been exposed to probabilistic reporting through DNA and they all feel it is appropriate in that context, primarily because there is a quantitative basis to the probability and the participants have a general conceptual understanding of how the numbers are produced. In pattern evidence, however, one participant was ambivalent and deferential to the practitioners (P#1), two participants expressed concern that probabilistic statements would be more confusing to interpret among fact-finders (P#1 and P#2), and one participant questioned whether there is a scientific basis to such probabilistic statements (P#3). For example:

> *So obviously probabilistic language has been used in reporting DNA results forever. ... I don't have any information or knowledge as to how something similar would be done in a pattern discipline. ... I would be open to considering it (P#1).*
>
> *A probabilistic conclusion is a lot looser and as a result is much less clear what that means (P#2).*

Overall, participants generally considered the benefits of categorical reporting as being its clarity and simplicity to express and understand. One participant added that an additional benefit is the certainty categorical expressions provide to the opinion, but also noted that it is just one small piece of the overall case (P#3). None of the participants expressed any significant risks to categorical reporting; however, two of the participants reasserted their concern over probabilistic reporting as creating additional complications to the conclusions. For example:

> *I think it gets messier the more you start complicating the conclusions in pattern matching disciplines (P#2).*
>
> *The benefit for categorical is the certainty of the opinion (P#3).*

When responding to concerns raised by practitioners as it relates to probabilistic reporting, participants agreed with the risk that it would be confusing to lay fact-finders and believed it was appropriate for them to take this into consideration when debating how to express their conclusions. For example:

> *I think that they should be worried about it to a certain extent. They should be cognizant of whether what they are saying at trial is an accurate description of their opinion (P#1).*

However, participants were less sympathetic to practitioners' expressing concerns that defense attorneys would use probabilistic reporting to create "reasonable doubt." Although one participant speculated the practitioners were concerned that defense attorneys would attempt to unfairly undermine their opinion with illegitimate attacks (P#1), which could be in the purview of the analyst to be concerned over, the other two participants expressed the perspective that practitioners should focus on what is scientifically appropriate and leave it to the litigators to argue their cases (P#2 and P#3). For example:

> *A defense attorney has an obligation to defend the interests of their clients. So, they can take anything in a case and try to create reasonable doubt. That's their job (P#2).*

When responding to questions raised about the role and duties of experts and the limits of their testimony, all three participants were clear that they expect the expert to accurately and impartially convey their opinion and limit their testimony to what is supported by the science. For example:

> *The roles and duties of forensic experts are to test the evidence and follow their rules and the best practices within their discipline and to accurately and impartially convey those opinions (P#1).*

> *A scientist, in my opinion, should give their opinion as to what the science can say (P#2).*

When asked about whether participants find it acceptable for experts to express their opinion in court without disclosing the underpinnings or statistical data to support those opinions, the participants were generally consistent in their response. Two participants responded by referencing governing evidentiary rules in their jurisdictions (P#1 and P#2) and all three participants suggested it is not required in their viewpoint, although it would not be the best practice to elicit the opinion without providing that foundation. For example:

> *There are specific rules of evidence that govern expert testimony in any jurisdiction, and they differ jurisdiction to jurisdiction. [In my jurisdiction], technically the expert doesn't even have to discuss the basis of their opinion. But they can be asked about it on cross (P#1).*

One participant expanded on this question by suggesting courts might tend to be more flexible when testimony is introduced as technical expertise versus scientific expertise and pointed out a growing debate as to whether pattern evidence might be better when presented under this framework. For example:

I think you're seeing a trend, particularly in microscopic toolmark evidence for firearms where the cases are being argued with technical expertise … and you're seeing some more challenges when it's being offered as scientific. So, it's an interesting question. It's a bigger question, I think, that is going on right now in the community is whether or not some of these pattern matching disciplines should be offered more as technical expertise rather than scientific experts to use, because both of them are legitimate to offer into evidence as expert opinion (P#2).

Finally, when asked what participants would describe as the greatest challenges facing the pattern and impression evidence disciplines as it relates to examination and reporting methods, the responses were varied—one participant pointed to understanding issues concerning the science (P#2) whereas the other two participants pointed to lawyers and other partisan attacks attempting to undermine forensic evidence overall (P#1 and P#3). For example:

I think it's a bigger issue that's happening in the community, is to understand what the conclusions are and what the limitations are, and to ensure that we're staying within those boundaries (P#2).

I think the challenge is that practitioners and people like you are attempting to appease the defense bar and that's never going to happen. ... You are never going to satisfy the defense bar because we are in an adversarial system. … So, I think that the challenge is trying not to fold in the face of that kind of pressure (P#3).

Use of Algorithms:

Prosecutors offered varying perspectives as it relates to the use of algorithms in court and the benefits and limitations of them. One participant objected to the use of algorithms in pattern evidence disciplines, claiming they did not believe algorithms were necessary and would unnecessarily confuse and complicate the testimony (making it more challenging for lay fact-finders to interpret) (P#1). Another participant was more skeptical, suggested algorithms could be useful to provide weight to analysts' conclusions, but cautioned against blind reliance on a computational algorithm without ensuring it is sufficiently valid and appropriate for the intended use (P#2). The third participant was more receptive to the use of algorithms, suggesting algorithms could be useful as a means of enabling the expert to be more efficient and delegate computational tasks to the algorithm that would otherwise be impractical to accomplish in a reasonable timeframe solely by the human, but questioned whether a computational algorithm similar to DNA is even possible for pattern evidence disciplines and expressed concern over how to effectively explain the algorithm to lay fact-finders. For example:

I think it would overly complicate things and I would not be in favor of it at this point (P#1).

[Algorithms] allow the scientists to do computations in seconds that would be undoable in a human timeframe, and so it gives you way more information and helps you weigh the

evidence. … I think it's working very well with the DNA [but] I do not see how we establish the numbers or the levels of confidence in pattern matching (P#3).

When asked about concerns over how algorithms can be trusted for use in court, including issues concerning the disclosure of source code, participants were generally consistent in their viewpoints. On the broader issues of trust, participants tended to be deferential to the forensic experts. On the issue of source code disclosure, although some participants did not feel it was necessary, they all expressed support for disclosure if requested by the defense under terms of confidentiality or protective order. For example:

[I]f it's scientifically valid and the scientific community is saying this is good science, then as a prosecutor, I'm behind it (P#2).

I'm all in favor of giving the defense every tool that they need to investigate the algorithm (P#3).

When algorithms are based on AI/ML, however, participants recognized the opaqueness of the algorithms as a potential issue. Although they generally believe AI/ML algorithms would be admissible under existing admissibility standards based on validation data, two participants recognized the potential challenges to admissibility on a constitutional dimension (P#1 and P#3). None of the participants, however, believed the algorithms would be wholly inadmissible, particularly if they were able to explain details about how the algorithms were developed (e.g., parameter selection, training data, etc.) and validated. For example:

Who am I going to call as a witness at a [admissibility] hearing to explain how this system works that I'm trying to show meets the admissibility standard for my jurisdiction (P#1)?

I would think that you would test that kind of algorithm the same way you do any other technology by using known samples. … I can see the confrontation issue. I don't see a due process issue, but I can see the argument that would be made (P#3).

When asked about regulation of algorithms, the participants were generally deferential to the forensic science community, but were conflicted on whether the legal system was an appropriate means of regulation. One participant believed the legal system was not the appropriate means of regulating algorithms (P#1). Another participant believed the legal system was an appropriate means of regulating algorithms, along with guidelines established by the scientific community (P#2). The third participant recognized the benefits of regulation, but expressed concern that many bodies composed of non-scientists often get "hijacked" by members with alternative agendas (P#3). For example:

I think [algorithms can be regulated] in the same way that forensic science is already being regulated. It's being regulated through best practice committees and through the court system, and I think that those are putting sufficient limitations around forensic science in general, and that would apply the same with algorithms (P#2).

I think that regulation in a reasonable way gives everybody confidence in the science. ... [However,] I'm not sure what that regulation would look like, and I'm not sure how, for lack of a better word, political, as opposed to scientific, that regulation would be (P#3).

Finally, when asked what participants would describe as the greatest challenges facing the operational use of computational algorithms for court purposes, the responses were generally consistent with one another and were concerned that algorithms might create additional challenges when presenting the evidence to lay fact-finders. Participants want to be sure examiners are comfortable and confident in their ability to explain in lay terms to the fact-finders the outcome of that evidence—the more complicated the computational methods, the more challenging it will be. For example:

I think it's getting stakeholders to understand. ... I think [algorithms are] very foreign to people in the entire forensic science community (P#2).

I think training the scientists within the labs, to validate it, and to understand it and have confidence in it. I'm not the scientist. I'm using the science and what I want is reliable science that is easy to understand and easy to explain to lay people (P#3).

6.4.3 Defense Attorneys

Background & Experience:

Three defense attorneys participated in the study—two male and one female. All defense attorneys are actively working in large metropolitan jurisdictions in the United States and have between 20 and 33 years of experience litigating criminal cases involving forensic science—primarily as public defenders. All three participants serve as the lead defense attorney specializing in litigating forensic science issues within their jurisdiction, as well as directing the work of other litigators on issues related to forensic science. One participant specializes strictly on post-conviction litigation. Participants' experiences span across a broad scope of disciplines, including both pattern evidence and analytical disciplines, such as drug identification, fingerprints, firearms, toxicology, dog scent, DNA, etc., as well as across a range of different types of cases, such as street crime, sexual assault, and homicide. Participant's experience litigating algorithms are varied and primarily involve probabilistic genotyping algorithms for DNA, as well as algorithms designed for investigatory purposes, such as "AI policing" and algorithms designed to detect and geolocate gunshots. The general focus of participants' litigation concerns is around issues concerning transparency, validation, and reliable applications of algorithmic tools.

Interpretation & Reporting Practices:

All three defense attorneys expressed a consistent perspective that categorical reporting in pattern evidence disciplines using terms such as "Identification" or "Individualization" is problematic, overstates the value of the evidence, and is not supported by the science. For example:

If you're going to make an association at all, it should never be categorical, and the association should always allow for the possibility of error or the possibility of a random match (D#1).

There's a tremendous amount of concern. Specifically, because there's essentially no scientific foundation for the claims of identification that are being made in almost all of the pattern disciplines (D#3).

Participants, however, did not necessarily view probabilistic reporting as superior to categorical reporting. The chief concern among participants is the extent to which the conclusions expressed are empirically supported, irrespective if they are reported categorically or probabilistically. Further, one participant expressed the concern that probabilistic reporting, without an adequate empirical foundation, would be misunderstood by fact-finders and misused by prosecutors (D#3). All participants were opposed to the use of probabilistic reporting using numerical references without empirical foundations as to what those numbers were based on. Rather than probabilistic reporting, especially in the absence of validated statistical methods upon which the numbers are based, two participants expressed the view that the optimal approach would be to report associations coupled with clear statements about error rates from black-box studies (D#1 and D#2). The other participant, however, expressed the view that probabilistic reporting would be marginally better (D#3). For example:

I think the move towards probabilistic language for any forensic discipline that doesn't have reliable rarity data is really problematic. (D#2).

There's a significant concern that jurors, number one, don't really understand probabilistic language and that prosecutors will misuse it. ... At the end of the day, if there were studies to support that type of language, and if there was some way to ensure that jurors understood what it meant and it was not misstated by either the examiner or by the prosecutor, I think probabilistic language is probably preferable (D#3).

Overall, participants generally considered the benefits of categorical reporting as the simplicity to express and understand what the expert is attempting to convey; however, all participants believe this is done at the cost of making inaccurate and exaggerated statements that are not supported. On the other hand, the participants generally considered the benefits of probabilistic reporting in that it explicitly conveys limitations, although the extent to which it accurately represents the limitations depends on the extent to which the statements are based on empirical studies. Without well-established validation studies to provide a foundation to probabilistic reporting schemes, especially when numerical quantities are included, could still be problematic since lay fact-finders tend to assume numerical expressions are based on empirical measurements. For example:

The positive is that [categorical statements] are easy to understand. … But it doesn't really accurately convey the weight of the evidence. ... I think very clearly categorical statements overstate the evidence, and that is always a significant danger. ... [On the other hand,] I think probabilistic statements they more accurately convey the weight of the evidence, [but] I think they are very difficult for judges, juries and litigators to understand (D#3).

When responding to concerns raised by practitioners as it relates to probabilistic reporting, participants generally agreed with the risk that it would be confusing to lay fact-finders and believed it was appropriate for them to take this into consideration when debating how to express their conclusions (although one participant [D#1] expressed the view that this is a reflection of the extent to which practitioners do not understand probabilistic concepts). For example:

> *I actually do think that the forensic science community does have some obligation for thinking through how information should be accurately reporting. I actually do think it is within their purview because I think that, again, that's something that for years has not been, either intentional or unintentional, but there have been overstatements made in every discipline for years and years and years (D#3).*

However, participants were quite critical of practitioners' expressing concerns that defense attorneys would use probabilistic reporting to create "reasonable doubt." Overall, none of the participants expressed a viewpoint that this would be appropriate for them to consider. One participant took it a step further and suggested this finding is indicative of a hidden bias in the criminal justice system (D#2). For example:

> *I think [forensic scientists] should stick to the science and let the lawyers worry about what we're going to say (D#1).*

> *I would call those results laughable if they didn't concern me so much. … Why are forensic examiners concerned about the outcome of the case? ... The fact that 80% of the examiners in a survey are concerned about case outcomes based on shifts of how we report language to me shows the power of the unconscious bias in the criminal justice system (D#2).*

When responding to questions raised about the role and duties of experts and the limits of their testimony, all three participants provided impassioned and consistent responses that forensic scientists base their conclusions on empirical data and be forthright about the limitations of their findings. Some participants went a step further by suggesting forensic scientists routinely fail to fulfill their ethical obligations, in their view (D#2 and D#3). For example:

> *The role and duty is to not overstate the science based on a subjective belief in it, or what you've been told by a mentor that isn't verified in science (D#1).*

> *Forensic experts have an ethical as well as a legal duty to accurately state the weaknesses and limitations of their forensic method. But forensic examiners don't take this duty seriously. In my 20+ years of litigating many forensic cases, I have never encountered a forensic examiner who took this duty seriously (D#2).*

When asked about whether participants find it acceptable for experts to express their opinion in court without disclosing the underpinnings or statistical data to support those opinions, all three participants were opposed to it. One participant stated a simple "no" without further elaboration (D#2). The other two participants went further to claim it is not legally admissible under existing admissibility standards (D#1 and D#3). One participant openly expressed frustration that such

testimony has been admitted in the past and pointed to poor education and poor performance by judges and defense attorneys in the past to have allowed such precedent to be established, but expressed optimism that judges are now beginning to take notice (D#3). For example:

> *No opinion should be entered into evidence without a thorough examination for the basis of it. The whole reason that we have a confrontation clause and cross examination is to examine the basis of the opinion (D#1).*

> *[Training and experience] are just not a legally sufficient basis for an opinion. ... [It's been admitted in the past because] for years and years and years, the defense bar really was, frankly, not educated and did not do a particularly good job of starting to bring to courts the problems with all of these disciplines. So, there's this whole body of case law that's based on either no litigation or very poor litigation (D#3).*

Finally, when asked what participants would describe as the greatest challenges facing the pattern and impression evidence disciplines as it relates to examination and reporting methods, all three participants pointed to the need to conduct the necessary research to provide empirical foundations to the evidence used in criminal cases. One participant (D#1) expressed an impassioned degree of frustration when expressing their viewpoint. This participant seemed to lament the impact of these divided perspectives across stakeholder groups and the lack of enforcement by the courts have had on indignant defendants, suggesting they are the ones that tend to bear the ultimate consequence for what should otherwise be straightforward scientific issues.

> *This digging in on the way that this has always been done because of subjective belief that there were no problems with it or because there haven't been tons of wrongful convictions associated with it, is sticking your head in the sand. ... The challenge is that courts will . . . . [well, . . .] I don't know, you know, actually, the truth is there may be no challenge, courts just may not care, because we don't care about the rights of the indigent defendants. In your typical criminal cases, the challenge is scientific integrity. The challenge is trying to claim science when you don't have any (D#1).*

> *In pattern matching, I would say it probably continues to be the lack of empirical research (D#2).*

One participant (D#3) went further and described their observation that research tends to be driven by the courts, based on what courts will or will not allow, and this is promoted by forensic scientists looking at court challenges to drive their research priorities. This participant expressed concern that this approach is unscientific and backwards—case outcomes where the admissibility of evidence is limited should not be the factor driving research agendas. Instead, this participant expressed the view that the research should be conducted without consideration of admissibility, then based on those results the courts determine whether the method is useful to the court.

> *It was stunning to me that the question that examiners would ask [litigators], essentially "what will the court allow?" And that is not how research should be conducted. It's not what the court will allow. It's what the research shows. ... And, then by that same token, I think that, at least in some of the disciplines right now, the research seems to be driven by*

the courts limiting the testimony. At least in firearms and toolmarks, what I've noticed is a court limits what a firearms examiner can testify to, and then there's a study that comes as a result of that limitation (D#3).

Use of Algorithms:

Defense attorneys offered generally consistent perspectives as it relates to the use of algorithms in court and the benefits and limitations of them. All three participants expressed significant caution to widespread adoption of algorithms, specifically over concerns of transparency, validation, and operational uses of algorithms. One participant summarized by stating "that's a complicated question" (D#3). Overall, all three participants were supportive of the use of algorithms, in theory, because, on the one hand they have the potential to provide an empirical basis to examiners' claims, to more accurately reflect the strength of evidence, to promote greater objectivity and consistency in examination results, and to enable examinations to be performed more efficiently. However, on the other hand, all three participants expressed concerns over transparency, validity, and reliability of algorithms when applied operationally. Participants' greatest concern was the lack of transparency surrounding the use of algorithms in criminal justice—specifically when algorithms are used from commercial vendors with proprietary software—which mask the underlying assumptions, parameters, and limitations of the algorithm. Without those details, participants' expressed concern that forensic scientists would apply algorithms operationally without fully understanding their limitations and the conditions upon which they might not be appropriate while at the same time "blindly" relying on the output as if it were factual. For example:

The greatest benefit would be is that you move away from unsupportable categorical claims into something that has some empirical basis to it and that you would actually have a number that's based on a valid statistical database, a population frequency database that is transparent and known. … [But,] I'm never not going to be concerned about proprietary software being used in these circumstances (D#1).

I think, when algorithms replicate the ability of human examiners in their interpretation, I'm much more comfortable with that use of an algorithm. … [However, I am concerned that] inevitably they will be used in the criminal justice system in a role that far exceeds what I'm calling for (D#2).

When asked about concerns over how algorithms can be trusted for use in court, including issues concerning the disclosure of source code, participants were consistent in their responses and renewed their calls for transparency and greater oversight. All three participants asserted that disclosure of source-code and access to the algorithm and underlying software application to enable them to test was key to gaining trust. One participant went a step further calling for the creation of an independent body of academic experts to assess the algorithm and oversee its operation in casework (D#2). None of the participants expressed a viewpoint that proprietary interests would be at risk if source-code were to be disclosed, particularly under conditions such as a protective order from the court, and each of the participants pointed to civil litigation as an example of courts applying disparate treatment of source-code disclosure in civil litigation versus

criminal litigation. One participant expressed the viewpoint that prosecutors shouldn't be using software for which they cannot give access to the source-code and underlying software (D#3). For example:

> *What would I need to be comfortable with widespread use and acceptance of an algorithm in the criminal justice system? First, I would need source code. … Developers should not work in any forensic space where the results of their algorithm operation are intended as evidence unless they are willing to publicly disclose their code. … Second, I would need some kind of oversight board—a team of neutral academic experts—provided with the time and resources to analyze the code, stress test it, and publish understandable reports about the assumptions underlying the code, the limits of operation based on stress testing, recommendations for improvement, and recommendations for testimony caveats based on their work. … Third, a pilot period of years, during which a limited deployment in casework is constantly reviewed by the neutral academic team to make sure that the system is being used as intended and that experts do not misstate the value of the evidence in court (D#2).*

> *If prosecutors are going to offer this service, then they should be prepared to turn over the discovery, and the discovery that I'm talking about in this context is the access to source code and the software, as well as all validation information and et cetera (D#3).*

When algorithms are based on AI/ML, however, one participant found it challenging to envision how these types of algorithms would be admissible (D#1). The other two participants, however, did not expressly object to the use of these types of algorithms, but re-enforced their concerns over the importance of transparency, accessibility, and oversight when these algorithms might be used (D#2 and D#3). For example:

> *You can't have somebody who just turns on the machine and you're coming in and testifying. If we don't know exactly how the machine works, why it works, what its error rates are, how it was developed and why, then it should never be used in criminal court. … It is, in my view, a sixth amendment violation, no matter what—if you were denied your right to confrontation, you were denied due process of law (D#1).*

> *I think [admissibility] would have to be on a case-by-case basis. ... I think the complication comes in when we try to find out what's behind the black box (D#3).*

When asked about regulation of algorithms, all three participants referenced the need for an independent oversight body responsible for assessing function, validation, operations, and testimony. One participant suggested it should be a neutral government entity, similar to the United States Food and Drug Administration (D#1). Participants also referenced standards set forth by the Institute of Electrical and Electronics Engineers (IEEE), suggesting a similar type of requirements should be established for the development and validation of software applications developed for criminal justice purposes. Finally, all participants expressed strong rejections to the idea of the legal system being an effective means of regulation. One participant went so far as to claim the legal system has "utterly failed" to regulate forensic science in general and therefore expressed no confidence it could not be trusted to effectively regulate algorithms (D#2). For example:

There should be independent bodies to assess their function, their validation, how they operate, who should be able to review training data, who should be able to require the appropriate caveats during testimony, who should be able to require that proper standards are used to develop [the algorithms], whether it's IEEE standards or others. … [The notion that the legal system could regulate algorithms is] really a laughable position. The criminal justice system has proven to be an utter failure as gatekeepers of forensic evidence (D#2).

Finally, when asked what participants would describe as the greatest challenges facing the operational use of computational algorithms for court purposes, participants referenced the need for increased investment in education for practitioners that will be expected to use the algorithms operationally, and for judges who will be expected to assess the admissibility of the algorithms. For example:

[The greatest challenge] is these non-scientists understanding what this machine is doing and the limitations of what the machine [and] results are. [Further,] having a forensic examiner, very few of which have a background in computational . . . anything, explaining accurately to these lay people what this machine is doing and the limitations of what this machine is doing (D#3).

6.4.4 Judges

Background & Experience:

Three judges participated in the study—one male and two female. One participant (J#1) is a sitting federal judge in a large metropolitan jurisdiction, having served for over 25 years as a federal judge and presiding over a wide range of criminal and civil cases, including issues concerning forensic evidence. Prior to being appointed as a federal judge, this participant served as both a federal prosecutor and a criminal defense attorney. Additionally, this participant serves as an adjunct professor at an Ivy League law school, has co-authored books, published numerous articles, delivered several presentations, and served on several professional committees, including those related to forensic science. Another participant (P#2) is a sitting state district court judge, having served six years of the current elected term[18]. Prior to being elected as a state district court judge, this participant served as a defense attorney, including experience as an assistance state public defender, with extensive experience litigating complex felony cases largely involving forensic science evidence—including issues related to the discovery of source code and admissibility of alcohol breath testing instruments. This participant has provided several presentations and trainings and has served on professional committees on issues related to the use of forensic science in courts. The third participant (J#3) is a former federal judge in a large metropolitan jurisdiction. This participant served as a federal judge for over seven years before stepping down in late 2018 to return to private practice and focus on issues in commercial litigation, including issues involving technology and artificial intelligence. Prior to serving as a

[18] This participant, (J#2), was first appointed by the state Governor in 2015 to fill a vacancy and elected to start a new term in 2016.

federal judge, J#3 served as a litigator in private practice for over 20 years and as the deputy assistant attorney general for the U.S. Department of Justice. While this participant has experience presiding over a wide array of criminal and civil cases, this participant has specialized experience on issues concerning artificial intelligence and algorithmic tools applied to the criminal justice system, having authored a book on the topic, provided several presentations and trainings, and served as an adjunct professor at a reputable law school on issues related to the use and presentation of quantitative methods by litigators, courts and policymakers as they advocate legal and policy positions.[19]

Interpretation & Reporting Practices:

The two participants who provided responses to these questions, (J#1) and (J#2), expressed the perspective that categorical reporting in pattern evidence disciplines using terms such as "Identification" or "Individualization" was challenging because it conveyed a degree of certainty that has not been well established.[20] These two participants suggested categorical reporting was akin to expressing an opinion "to a reasonable degree of scientific certainty," and expressed the concern that those statements do not have clear meaning to lay fact-finders and not only mask the level of subjectivity involved in the examination, but also convey a level of certainty that exceeds what can practically be achieved. One participant (J#1) goes further to suggest that the means by which forensic science conclusions are reported is a factor that has contributed to the erroneous conviction of innocent people. The other participant (J#2) expressed a view that categorical statements involving source attribution could be acceptable provided that the examiner could provide adequate foundation to support such a claim and the relevant uncertainties and limitations of the examination are conveyed. However, this participant goes further and openly questions whether it is practical to establish such a foundation and demonstrate that the uncertainty is such that a categorical statement of a source attribution is warranted. For example:

> *As I think many people know, bad forensic science has been an element in the conviction of innocent people. ... One of the reasons that those inaccuracies [in forensic science] came about [was] because the science itself was much more subjective than was represented to courts and to juries, [and] because they were presented as being certain conclusions. ... There's almost no part of science that can claim certainty. If you talk to physicists or chemists or whatever, they won't claim that. Yet here it is, in effect, being claimed by forensic science (J#1).*
>
> *I think it's very challenging to use [categorical statements] for purposes of how to report a result. ... How do I know that there's the foundational science to be able to say that, as we're doing this comparison, that I can make the statement, "yes, this impression came from this source?" We get into [things] like, "well, it's a match." Well, okay. It may be, [but] how do you know that (J#2)?*

[19] The majority of the interview with this participant, (J#3), focused on issues related to the broader topic of "computational algorithms" for court purposes. Many questions related to the broader topic of "interpretation and reporting" were omitted and, therefore, are not discussed.

[20] Participant (J#3) did not provide a specific perspective on issues related to this topic.

Participants considered probabilistic reporting as an improvement over categorical reporting; however, participants cautioned that it may not necessarily address all of the concerns. One participant (J#1) suggested probabilistic reporting is an improvement to categorical reporting, but expressed a concern that lay fact-finders would not be able to meaningfully interpret what was being conveyed or scrutinize the validity of the underlying statistical methodology upon which the probabilistic statement was based. Another participant (J#2) expressed initial reactions of being averse to probabilistic statements given the potential to be misunderstood. However, after reflecting on the issue more, this participant expressed a view that probabilistic statements could be advantageous to categorical reporting because they cause the fact-finders to pause and think through the nuances of what is actually being conveyed rather than relying on familiar colloquial definitions of terms that are often used when reporting categorically (despite such terms having a specific technical definition in the respective forensic discipline). This participant went further, however, to express the view that probabilistic statements should include numbers, and those numbers should be accompanied by a statistical model to provide the source of those values. For example:

> *Well, I think [probabilistic statements] would be an improvement, but I worry again about two things. First, the ability of judges and juries to really scrutinize, in a meaningful way, when someone says it's this probability or that probability. And secondly, the validity of the underlying statistical methodology used, which varies considerably. ... Nevertheless, I think expressing it as a probability would still be better than expressing it as a certainty. But I do think it still has a great potential to confuse (J#1).*

Overall, participants generally considered the benefits of categorical reporting as its simplicity of the statements; however, they also expressed the concern that such statements are not well-defined and are often interpreted to mean something that is not supportable. One participant (J#1) stated "the greatest risk with categorical is it's stated as a certain thing, and that's just not true." Another participant (J#2) believed such statements "do not always align with what lay persons' understanding of the definitions would be." Participants viewed probabilistic reporting as being an improvement over categorical reporting in the sense that probabilistic reporting is more defensible and easier to define, but participants still expressed concern. One participant (J#1) questioned whether statistical methods are appropriate when there is a high degree of subjectivity, and also noted "that the recipients, the judges and juries who are hearing these opinions are very rarely people of statistical sophistication and so they may give a greater weight than it really deserves." Another participant (J#2) cautioned that "probabilistic models have ways in which they can be misconstrued."

When one participant (J#1) was asked how such testimony should be permitted, the participant responded, "it varies from discipline to discipline." The participant elaborated by reference to a prior case opinion they authored:

> *The best way to answer that is by talking about an opinion I wrote, United States v. [REDACTED], where the question was whether there was a match between the marks on the bullet and cartridge from the gun. ... Originally, I asked the expert, "what's your error rate?" and he said "zero." I said "zero?" And he said, "because I've never testified in a case in which the defendant wasn't convicted." ... Put[ting] aside that non-sequitur for the*

moment. More to the point, in the end, what I allowed in that case was for the expert to show great big blow ups of the marks on the bullet and cartridge and the marks on the gun, and to point out some of the similarities between those and to then express the opinion that it was more likely than not that this came from the same gun. That's as far as I felt one could go without misleading the jury. I'm not sure today I would even go that far because I've seen many more examples of wrong, inaccurate forensic science, but I certainly wouldn't go any further than "it's more likely than not in my opinion that this bullet came from that gun." Of course, it depends on the forensic discipline. When you're talking about, for example, microscopic hair analysis, the error rate is extremely high and I wouldn't allow that in. I might have back at the time of [this case] considered allowing it in the modified way I indicated, but no longer. So, it varies from discipline to discipline. [For fingerprints in particular,] they are not bad forensic science, but they're not DNA either. ... I think I would not exclude it. ... I think that the evidence is there that fingerprint evidence is not junk science and that with proper limitations, it can be received in evidence. [For example, I would probably allow] the expert to blow up pictures of the two fingerprints to be shown to the jury and point out some of the similarities between those and then express the opinion that it was more likely than not that this [print] came from the same [individual]. ... I [also] think maybe you should require as part of [the expert's] direct testimony, to say, "now I've arrived at that [opinion] through experience, not through some sort of scientific formula" (J#1).

When responding to concerns raised by practitioners as it relates to probabilistic reporting, participants agreed that probabilistic reporting would be more confusing to lay fact-finders, but they did not express the view that the issues were insurmountable.[21] One participant (J#1) suggested that the risks for confusion, which probabilistic reporting might entail, would be less worrisome than the view fact-finders often take with categorical reporting. The other participant (J#2), while recognizing the potential for confusion, expressed the view that, on the other hand, probabilistic reporting might be useful to cause people to pause and think through the nuances rather than rushing to judgment based on colloquial uses of terms that experts use categorically. For example:

Well, I do think there is a potential for confusion, but it's not as bad as the view that the jury will take otherwise, that it's an absolute fact. When the jury hears the opinion it's a match, their natural reaction is to say, "okay, it's been scientifically found that it's a match. Period." (J#1).

At first, when I started working with them, I was like, this is way too confusing and there's no way we're going to be able to do this in a way that's meaningful to people, but in some ways, I think there are some things about it that makes it more approachable (J#2).

However, participants were less sympathetic to practitioners' expressing concerns that defense attorneys would use probabilistic reporting to create "reasonable doubt."[22] Participants disagreed with the practitioners' concern and expressed concern that it would be a factor taken into consideration. One participant (J#1) suggested that this indicates practitioners do not have faith in

[21] Participant (J#3) did not provide a specific perspective on issues related to this topic.
[22] Participant (J#3) did not provide a specific perspective on issues related to this topic.

juries and offered a reminder that the determination of reasonable doubt is what the judicial system is all about. The other participant (J#2) suggested that this indicates a general fear practitioners have for defense attorneys. For example:

> *I'm not sure what is meant by the objection that this might create a reasonable doubt. Well, that's what the system is all about, is finding out whether there is, or is not, a reasonable doubt. It sounds like those respondents didn't have much faith in juries (J#1).*

> *I think we would need to stop being afraid of defense attorneys. I really do think that we just need to stop that nonsense. These numbers can be misused by everybody because they aren't being understood properly. I don't think a lot of it is even intentional. I just think that it is what it is. So, I think misuse happens for all sorts of reasons and it doesn't have to do with what side you're on. So no, I don't think that it should be a reason that we should not look at [probabilistic reporting] (J#2).*

When responding to questions raised about the role and duties of experts and the limits of their testimony, participants expressed the view that results should be reported in an accurate manner with appropriate foundation to base such conclusions and that the experts should be forthright about the error rate and limitations of the findings. As one participant noted, without being forthright about this information "the jury is deprived of information that is available, that is out there" (J#1).

When asked about whether participants find it acceptable for experts to express their opinion in court without disclosing the underpinnings or statistical data to support those opinions, one participant (J#1) stated "no" without further elaboration. Another participant (J#2) admitted to have struggled with this question, stating that the rules of court require the expert to provide the foundational support for their opinions, but experts should be answering the questions put to them by the lawyers. Instead, this participant, suggested that experts should be more proactive about disclosing these foundational issues earlier, such that it is laid out before the court process, such as on the report that is provided to both parties, which, in turn, would enable either party to further discuss during court as they deem appropriate. The third participant, (J#3), stressed throughout the interview that "the means to the end matter"—both as it relates to expert testimony and the use of algorithms—and that an opinion that is expressed without the reasons for that opinion would be considered *ipse dixit* and cannot be relied upon. For example:

> *My view is that [would be] called ipse dixit—"it is because I said it is," and, under the Daubert standards, the Supreme Court standard for the admissibility of an expert opinion, that's not allowed. ... Every judge should require that an opinion be backed up by the reasons for the opinion and that, if an expert gets up there and says, "based upon my experience, this is just the way it is," ... I would say that that's an unreliable opinion (J#3).*

Finally, when asked what participants would describe as the greatest challenges facing the pattern and impression evidence disciplines as it relates to examination and reporting methods, participants pointed to multiple issues.[23] One participant, (J#1), responded with the need for "good, blind, scientific testing" to strengthen the scientific rigor underlying many forensic science

[23] Participant (J#3) did not provide a specific perspective on issues related to this topic.

disciplines. This participant, (J#1), elaborated that the "greatest failing" is that many forensic sciences, with the exception of DNA, have been developed by police as investigative tools and began to be introduced as hard evidence without subjecting it to serious testing. The other participant, (J#2), expressed the view that the greatest challenge is to ensure, irrespective of how those results are reported, that everyone understands how to properly interpret the value of the evidence.

Use of Algorithms:

The judges offered generally consistent perspectives as it relates to the use of algorithms in court and the benefits and limitations of them. All three participants expressed views that algorithms can be helpful—particularly for purposes of augmenting the expert to reduce the degree of subjectivity in the analysis and performing tasks that humans would otherwise be incapable of doing. However, participants also expressed caution about the desire to rely on algorithms without ensuring that there is transparency into how the algorithms operate and clear understanding of the limitations of the systems. One participant (J#1) expressed concerns citing the lack of transparency, logistical, and financial challenges often prohibiting defense counsel to meaningful scrutinize algorithms used in the criminal justice system. Another participant (J#2) expressed the view that the lack of transparency around these algorithms not only creates the opportunity for misuse, but also perpetuates a culture of distrust that already pervades the adversarial system, which ultimately "erodes confidence in the analysis as well as potentially in the system itself." The third participant (J#3) suggested there needs to be a national conversation on how to create trustworthy and reliable algorithms, and what that means, as it relates to uses for individual liberty determinations. For example:

> *I think algorithms can be helpful, to a degree, if they are totally transparent. ... I think really good algorithms could reduce the subjective portion of the analysis. ... [However,] some companies are obscuring inquiry through trade secrecy laws, but even where that doesn't operate it's very hard for even defense counsel [to review]. ... Even in those states where the trade secrecy law objection is overruled, they have to hire an expert ... [but often] there's no money available to hire that kind of expert (J#1).*

> *I think that algorithms are here to stay. ... There's a great potential [with algorithms], [if] done correctly, to create criminal justice reform to a degree that we've never seen before. ... [T]hey have an ability to take out some of the human biases that have plagued the criminal justice system. ... [B]ut there are certain risks. ... What we need is a national conversation on what that means and how to create trustworthy and reliable algorithms that can be used for individual liberty determinations. That's where the rubber meets the road (J#3).*

When asked about concerns over how algorithms can be trusted for use in court, including issues concerning the disclosure of source code, participants were consistent in their views, echoing their prior concerns about transparency and asserting the need for access to source code. Participant (J#3) expanded on the concept of trustworthiness by pointing not only to reliability testing, but also whether the design of the system corresponds to a concept of "fairness." This

participant argues, on a Constitutional basis, that "the means to the end matter" and the "means" are contained within the source code. For example:

> *I think [source code] absolutely should be disclosed in every case. I don't see how you can tell the judge, let alone the defense lawyer, [they] can evaluate whether it's a good algorithmic approach or not if you don't know how what went into the source code and what its components were, how they were arrived at it, and so forth. And, give me a break about trades secrets. I appreciate that companies like to make money, but we're talking about human liberty here, and that has to trump any concerns over trade secrets (J#1).*
>
> *I personally think that it should be open source codes, period. … I respect the fact that there's intellectual property issues and so forth that's around that, but I think that we have mechanisms to assist in protecting that (J#2).*
>
> *I think that what it means to be trustworthy is very close to what it means to be reliable, but I think it incorporates something else. Reliability is simply, "does the tool work as it is intended to work?" … Trustworthy certainly incorporates that, but it [also] incorporates something else, which is a concept of fairness. … In my view, if an algorithm is going to be used for a liberty-based decision, a criminal defendant is entitled to have access to the source code, and I would say for an adequate defense, just as a criminal defendant is entitled to the experts that he or she can demonstrate are needed to put on an adequate defense, that same individual is entitled to an expert who can then help them analyze the algorithm (J#3).*

When algorithms are based on AI/ML, however, participants were not completely opposed to their use; however, they did express views that were even more cautious given the lack of transparency. When asked whether the opaqueness of these types of algorithms could present an issue from a Constitutional dimension, such as Due Process or Confrontation, two participants (J#1 and J#2) did not believe, in general, it would be wholly excluded, but did express concern over their use nevertheless. The third participant, (J#3), expressed the view that understanding the design of the algorithm is absolutely critical, and in the absence of such information the evidence generated by the algorithm should be excluded. Ultimately, this participant was unwilling to accept that the conceptual innerworkings and design of the system is incomprehensible, despite the apparent black box nature of the source code file itself that is often the case with AI/ML algorithms, and expressed the view that giving up the ability to understand these issues would be giving up important Constitutional principles. For example:

> *At a minimum you need to know what the error rate is. ... But, also, I'm a little suspicious about any notion in the legal system where we say, "we don't know why X causes Y, but we know it does." … I think a lot of scientists, a lot of lawyers, would be very skeptical about the use of that because ultimately the law depends on reason, not on assumptions. ... So, I am skeptical of the black box approach (J#1).*
>
> *They fascinate me and scare me all at the same time. I can't say that access to the source code is the "be all and end all" of anything. … [B]ut I don't even know how to begin to assess that stuff. … I really think that if we're going to start using them, that we need to*

figure out what it is that we do need for purposes of making sure that there's essentially buy-in from everybody, that this is why this is working and that we can have some check on the fact that it is working in the way that we believe that it's working (J#2).

Understanding how the instrument was designed is absolutely critical to understanding the calibration of the instrument and the choices. … [Ultimately,] I think there are serious due process issues with a defendant being denied access to understanding information that underlies a tool being used for liberty decision (J#3).

When asked about regulation of algorithms, the participants expressed views that spanned across the forensic sciences more broadly, not just algorithms, that there should be regulation. Although participants had different views on who and how that regulation should be done, participants did not feel the legal system was effective as-is. For example:

Yes, [but] not just algorithms. I think there is a real need for an Institute of Forensic Science staffed by a high-level scientists who could tell us with the neutrality that we deserve, this is good forensic science, this is bad forensic science, this is possible forensic science but it has to be improved and here's how to go about improving it. … I don't think the legal system, ultimately, is well positioned to regulate forensic science. Judges know beans about science. Lawyers know beans about science. The natural thing when you have that kind of problem is to turn it over to the people who do know about science, the scientists (J#1).

Yes, [but] the by whom and how is a much harder question. ... [Whether the legal system is an appropriate means of regulating forensic science,] no, [but] I will also say I'm not sure the federal government is the place to regulate it either (J#2).

In my view, there should be a form of regulation that is for any liberty-based decision. It's a broad question in terms of algorithms and any kind of forensic science, … [but] if it's going to be used for a liberty-based decision for a human being, then they need to meet the constitutional standards, so they should be regulated. … The, how, I think, is extraordinarily complicated, but I don't accept that it can't be done (J#3).

6.4.5 Other (Academic Scholars)

Background & Experience:

Three "other" stakeholders (i.e., academic scholars) participated in the study—two male and one female. One participant (O#1) has over 30 years of experience performing research in forensic science, with the specific aim to provide a more structured foundation to case assessment, evaluation, and interpretation, and served for several years in a chief government role establishing policy governing forensic science practices on a national level[24]. Another participant (O#2) has over 30 years of experience as an academic scholar at an Ivy League university, primarily focused on research involving human judgment and decision making from a multidisciplinary perspective,

[24] This participant, (O#1), was the only non-U.S. centric participant.

including law, psychology, biology, and statistics. The third participant (O#3) has over 35 years of experience as an academic scholar at an Ivy League university, primarily focused on physical sciences, mathematics, and general scientific issues of public interest. This participant has also served as the president of a large scientific organization. All three participants are respected in the general scientific community, have doctoral degrees in scientific disciplines, have numerous scientific publications, and have experience serving in senior advisory roles on issues affecting forensic science practices on a national scale.

Interpretation & Reporting Practices:

All three participants expressed the perspective that categorical reporting in pattern evidence disciplines using terms such as "Identification" or "Individualization" was inappropriate and conveyed a level of certainty that was unsubstantiated and outside the realm of what scientific principles can support. One participant recognized the effort that would be involved with promoting such a transition and expressed a perspective that categorical reporting, in the interim, should be accompanied by statements about the limitations of such claims (O#1). The other two participants expressed a much more rigid perspective, suggesting such claims were not scientifically justified and were an overstatement of what can be empirically supported (O#2 and O#3). One participant took it a step further and expressed the view that such claims violated the trust that fact-finders place in forensic scientists and was "immoral" if they made such claims under the auspice of "science" (O#3). For example:

> *I think it's clearly not justified scientifically. It's an overstatement of the value of the evidence. We know it's simply not plausible for a discipline, like fingerprinting, that a trained examiner can determine the rarity of the set of features observed [based solely on human judgment] with the precision necessary to know whether it's probability in the population is low enough to support the claim that it's a unique observation (O#2).*

> *I think it is wrong. I think it's immoral to stand in front of a jury and make categorical statements if you are a forensic scientist because the word "scientist" confers in the minds of the jury that you are, well, one way that I heard it expressed is that the words have totemic power. I think it's wrong to abuse that level of trust. ... Look, the way I view it, we can make categorical statements, but don't claim it's backed up by science (O#3).*

Participants were not completely consistent with endorsing probabilistic reporting, however. One participant expressed strong views that probabilistic reporting was the path forward (O#1). However, another participant seemed to support probabilistic reporting simply because of the lack of any reasonable alternative and that categorical reporting was not acceptable (O#2). This participant seemed to accept probabilistic reporting as the path forward, but was more interested in how to most effectively articulate probabilistic results to lay fact-finders to maximize their comprehension of the information—a topic that this participant believes still requires more research. The third participant, however, expressed views that seemed to reject both categorical reporting (as it is traditionally practiced) and probabilistic reporting (O#3). This participant expressed concern that probabilistic reporting, albeit superior than categorical reporting from a scientific standpoint, would not be well understood by fact-finders. Instead, this participant

suggested black-box testing of examiners' performance was the optimal approach, so that examiners' conclusions can be accompanied by an empirical measure of certainty based on error rate data (O#3). For example:

> *I strongly believe that [probabilistic reporting] is the appropriate approach to take. ... It is much more scientifically correct and defensible to acknowledge that uncertainty in a probabilistic form (O#1).*
>
> *I have problems with [probabilistic reporting] too, but the problems don't lie on the side of the forensic science community, it lies on the side of the triers of fact. [For example,] I know for a fact, most people don't understand fractions … So, I'm not sure if probabilistic is better, but I know a lot of people are in favor [of it] (O#3).*

Overall, participants generally considered the benefits of categorical reporting as its simplicity to express and understand; however, they all acknowledged that ease of understanding is at a cost of being scientifically valid and transparent about the uncertainty. Participants viewed probabilistic reporting, on the other hand, as being scientifically more defensible, but at the same time, more challenging for lay people to understand and at an increased risk of erroneous interpretations.

When responding to concerns raised by practitioners as it relates to probabilistic reporting, all three participants were sympathetic to the concern that probabilistic reporting would be confusing to lay people. Although one participant (O#3) responded in a way that suggested probabilistic reporting was not the ideal path forward (versus black box testing to derive empirical error rates), the other two participants (O#1 and O#2) did not believe the confusion that would accompany probabilistic reporting was insurmountable or a strong enough reason not to pursue it. For example:

> *I think they are right. It may be confusing to a lot of people, but I don't think that's a sufficient reason to go back to an unjustifiable alternative form of reporting (O#2).*
>
> *[I agree,] just ask someone on the corner and say, "I have this problem with fractions. I want you to solve it" and see what kind of reaction you get. So that informs me that for the average person who finds themselves on the jury, a deep understanding of probability is it's like asking them to solve Einstein's equations. It's just not going to occur (O#3).*

Participants were also understanding of practitioners' expressing concerns that defense attorneys would use probabilistic reporting to create "reasonable doubt;" however, participants did not view it as a reason to oppose probabilistic reporting. To the contrary, participants suggested it bolstered the reason to pursue probabilistic reporting if it more effectively represented the certainty of the findings. One participant (O#1) expressed concern that this indicates a deeper cultural challenge that forensic scientists are averse to talking about anything that might undermine the certainty of their findings. Another participant (O#2) noted the irony in the question and highlighted the fact that it is the very job of defense attorneys to highlight anything that should cause fact-finders to

doubt the evidence—particularly if the doubt is "reasonable"[25]. The third participant (O#3) agreed with the practitioners' concerns recognizing probabilistic reporting creates an opportunity to for defense attorneys to abuse it and bolster their arguments, but also suggested categorical reporting that does not acknowledge the uncertainties also creates opportunities for prosecutors to abuse it to bolster their arguments. Considering the risk for both parties to abuse each type of reporting methods, this participant, (O#3), echoed their perspective that empirical measures of accuracy through black box testing is a way to put boundaries around these issues. For example:

> *I think they don't want doubt introduced, [and] it scares me actually. It scares me that forensic scientists don't feel confident to talk through uncertainties and anything that is below a hundred percent. We, as scientists, should be comfortable in talking about the limitations of our analysis as much as the strengths of our analysis. It's the job of defense attorneys to introduce reasonable doubt, but it's our job to be sufficiently transparent to allow them to scrutinize the evidence (O#1).*

> *[First of all,] creating reasonable doubt is what defense lawyers are supposed to be doing. If there's some reasons to doubt the finding, then the jury should know about them. ... [Second,] from my perspective, this portrays a mindset, which is that the goal of forensic science is to produce convictions and anything that gets in the way of producing convictions is a bad thing. I just have a totally different perspective on this (O#2).*

When responding to questions raised about the role and duties of experts and the limits of their testimony, all three participants expressed the view that experts' number one priority should be ensuring their results that are reported are scientifically defensible. Two of the participants define this in terms of transparency about the uncertainty that might exist to ensure the court has the requisite information to make an informed decision (O#1 and O#2). The other participant (O#3) defines this in terms of ensuring testimony is grounded by measures of repeatability and reproducibility. For example:

> *I think the role of a forensic science expert is to assist the court, not the prosecution or the defense but the court, in its evaluating evidence and to use their skill and knowledge that lay people don't have to help evaluate the scientific findings in a way that is helpful to the court—that is transparent about strengths and limitations. ... I think it is the role of the court to conduct that final reasoning in the light of the uncertainty that exists (O#1).*

> *I think the first duty is to get it right—to say things that are justified scientifically [and] to not go beyond their expertise and not claim more than the science will support. That's duty number one. Do not make unjustifiable claims. Then duty number two is, once you've identified the various claims that might be justifiable, try and choose among them in a way that promotes better understanding for a wider range of people. When in doubt, maybe present the evidence in multiple alternative ways and focus on transparency and a fair characterization of uncertainty (O#2).*

[25] This participant noted the awkwardness of the question to suggest the doubt be "reasonable." The wording of the question was intentional and correctly represented how it was phrased in the survey to practitioners—as "reasonable doubt." See [51] for the wording of the question as phrased to practitioners.

When asked about whether participants find it acceptable for experts to express their opinion in court without disclosing the underpinnings or statistical data to support those opinions, two participants flat out stated "no" without further elaboration or exception (O#2 and O#3). The other participant (O#1) expressed the view that disclosing the underpinnings of the expert opinion is important, but also recognized the dynamics that affect testimony in a court setting. Nevertheless, this participant suggested the foundations for the expert's opinion should be disclosed in the case file so that it is documented and available, if needed. For example:

> *I think it is really important to disclose the basis of your opinion. I think when it comes to the actual courtroom, [however,] it depends on so many things—what you actually say in testimony. When it comes to your written statement of evidence and your case file, that contains all your notes, [however,] I think that underpinning has got to be disclosed so at least it should be available for scrutiny by whoever in the court process wants to scrutinize it. I think that when we just give unqualified opinions, it is almost impossible to challenge really, because if you're not giving a reason for your opinion then it just comes down to, "well, that's my opinion" (O#1).*

Finally, when asked what participants would describe as the greatest challenges facing the pattern and impression evidence disciplines as it relates to examination and reporting methods, participants' responses were quite varied. One participant (O#1) pointed to an on-going narrative that forensic sciences are "in crisis" and implications that they are useless unless perfect. This participant expressed the view that such an aspiration of perfection is unrealistic and fails to recognize the value that many pattern evidence disciplines can give, provided that there is transparency around the limitations and imperfections of the disciplines. Another participant (O#2) pointed to the need for on-going validation of the examination methods and recognition of the limitations of those methods as revealed by validation studies. This participant expressed the view that these validation studies should be on-going and ideally be incorporated into routine casework through blind testing. The other participant (O#3) pointed to resources as the greatest challenge facing the forensic sciences. This participant suggested that the conditions that many forensic scientists are working under is conducive to errors, and calls for greater investment and support of the forensic science community to provide the resources necessary to perform at the level that society expects and needs. For example:

> *The greatest challenge that I've observed is actually resources. … I have had a chance to see the conditions that real forensic scientists work under. They're not the conditions that Hollywood tells the public about. The real conditions are often overworked people [and] under-resourced people with no time to get the results out. I mean, that's the real world. To me, that's the greatest challenge to forensic science, to convince our society to put in the resources so that people can do the best job, so that this intuitive expertise that I [believe forensic scientists have], is actually allowed to work without having the pressure that can induce errors (O#3).*

Use of Algorithms:

The academic scholars offered generally consistent perspectives as it relates to the use of algorithms in court and the benefits and limitations of them. All three participants expressed favorable views of algorithms, in general, but with caveats. One participant (O#1) expressed very favorable views of algorithms for which the underlying operation is understandable and explainable; however, this participant expressed extreme caution when the algorithms are not well understood. This participant went further to question whether it is even practical to fully validate algorithms that are not well understood, or if there is a sufficient legal basis for which to introduce those types of algorithms. Another participant (O#2) recognized the value of algorithmic approaches over human judgment, but conditioned that support on whether the specific algorithm in question was "validated and appropriate," including assessments in case specific applications. The other participant (O#3) was supportive of algorithms provided they were free of any ties to demographic factors or large characterizations of populations and pointed to algorithms used in "predictive policing" as an example, where the algorithms can perpetuate systemic biases. For example:

> *I think algorithms may well be preferable to human examiners giving opinions based upon experience because the use of the algorithm reduces the chances for bias and it may allow better estimation and calibration of that strength of the evidence. … [However,] these models tend to be very complicated and difficult to assess. Algorithms have advantages, but it's going to require a whole new realm of expertise to evaluate them (O#2).*

When asked about concerns over how algorithms can be trusted for use in court, including issues concerning the disclosure of source code, participants all pointed to validation data as the key to demonstrating the performance of the system under various conditions that are representative of the facts of the present case. As part of validation, participants expressed strong views that there needs to be clear understanding of the boundary conditions for which the algorithm performs well, and the circumstances (or combination of circumstances) for which the algorithm might begin to fail. As long as those conditions are well understood, participants suggested that there is reason to trust the output of the algorithm. Participants recognized the importance of transparency in building trust, and disclosure of source code is a key element of transparency. Although all participants encouraged the disclosure of source code to promote greater transparency, none of the participants expressed strong views that the source code *must* be disclosed before an algorithm can be trusted. Rather, participants pointed to the existence of validation studies and conceptual descriptions of how the algorithm operates, along with its limitations, as well as having access to the algorithm for independent testing as being more useful to the typical expert. For example:

> *Well, two things: transparency and performance testing. Transparency, because if I were in some sort of legal situation where an algorithm played a role in determining my freedom or even more consequentially my life, I would want my attorney(s) to have the ability to bring their experts to look at the algorithm [and] to make sure that I wasn't a victim of bias. So, transparency is for me, the first thing. The second thing is [that] I want these algorithms tested on a regular basis, looking for failure modes. I want the reliability testing as part of the use of it (O#3).*

When algorithms are based on AI/ML, however, participants expressed views that suggested they were skeptical of whether these algorithms could be validated in a way that fully understood their boundary conditions and limitations in such a way that would be appropriate for court. Although two of the participants did not explicitly reject the idea of AI/ML algorithms (O#1 and O#2), one participant (O#3) opposed the idea altogether. All three participants expressed similar concerns that the level of effort to truly understand the boundary conditions and limitations of the algorithm through performance testing would be impractical to accomplish. For example:

> *I think if [the algorithm] is not understood to the developers and it's a total black box, then I struggle to see on what basis that there is fair transparency in the [legal] proceeding (O#1).*

> *Theoretically, it could be acceptable to use these systems if we have reliability testing. [The problem is], the testing has to be large and broad because you don't know where the failure modes are[.] ... [That said,] is this type of reliability testing practical? What I've talked about is the ideal. I don't think the idea is actually practical [and] realizable. I don't think you could actually implement it (O#3).*

When asked about regulation of algorithms, the participants recognized the need for oversight, but offered slightly different perspectives. One participant (O#1) suggested that the regulation should be focused on the method—not just the algorithm, which is a narrow part of the overall method—such that regulation addressed the validation of the algorithm as well as the use of the algorithm (including the training and competency of the people, the inputs, and testimony of the results). Another participant (O#2) expressed the need for regulation by an independent oversight body. This participant lamented the current situation where regulation is left to the legal system and expressed strong concerns that the legal system is ineffective at regulating forensic science overall, much less algorithms. The other participant (O#3) felt unqualified to address this question and was cautious to offer an opinion from a professional capacity; however, when asked from a personal capacity, as a citizen and potential consumer of forensic science evidence that could be based on algorithmic tools, this participant stated clearly that they were opposed to the use of any algorithm in court based on machine learning that was a total "black box."

> *It's not the algorithms that need to be regulated, it's the methods, and the methods include the people, the algorithms, the data, and everything else (O#1).*

> *Yes, I still think it would be nice if we had a national institute of forensic sciences contemplated by the NAS report in 2009. ... Right now, we're stuck with the regulatory authority being exercised by judges who, for the most part, have not shown a willingness to apply rigorous quality control with regard to validation of forensic science. ... So, I'd like to see more federal involvement with agencies that have the ability to make some scientific assessment and set regulations on their own. I think that would be appropriate (O#2).*

Finally, when asked what participants would describe as the greatest challenges facing the operational use of computational algorithms for court purposes, participants offered very different

viewpoints. One participant (O#1) highlighted the need for clear understanding of what type of algorithms is being considered because the benefits and risks vary widely, and to ensure scientific debates about the validity and appropriateness of algorithms are done in a scientific setting outside of a specific legal hearing. This participant also pointed out the need for improving education and training for both forensic science practitioners and legal stakeholders on these issues. Another participant (O#2) discussed the need for the development and validation of robust algorithms, but highlighted the challenges associated with their implementation. This participant went a step further and suggested that the move toward algorithms might also necessitate changes around recruitment and selection of forensic science practitioners to include stronger backgrounds in mathematics, statistics, and hard physical sciences that might provide greater exposure and receptivity to algorithmic tools. The other participant (O#3) expressed concerns of the potential for the quality and reliability of algorithmic tools to degrade over time if their development and validation are left to commercial entities with financial interests. For example:

> *I think we need to really work on education of practitioners and our legal colleagues in terms of fundamentals of probabilistic [concepts], in terms of what it means to be transparent and to disclose limitations, and how we work with these kinds of new technologies (O#1).*

> *We need to be realistic about how easy it is to implement them. ... I think we need to think seriously about, given our movement toward these algorithms, the way we train forensic scientists and select them. So, picking people who have higher levels of mathematical and statistical aptitude training might be really important. At the same time, I think we need to be sensitive to current practitioners who are math phobic and, kind of ease them in and select more of those practitioners who have degrees in math and statistics, or the harder physical sciences and, thus, may be capable of moving into the new world with a greater degree of facility than we may see from the typical pattern matching person (O#2).*

6.5 Discussion

This study explored the perspectives of key criminal justice stakeholders, including laboratory managers, prosecuting attorneys, defense attorneys, judges, and other academic scientists and scholars on issues related to: (i) interpretation and reporting practices (with or without algorithmic tools) and (ii) the implications of the use of algorithms in legal settings as a means of calculating the probabilistic values assigned to the evidence. Participants offered a rich and diverse set of perspectives on these issues; however, we caution against generalizing these perspectives too broadly. We cannot suggest, nor do we believe, these perspectives are representative of the different stakeholder groups more broadly. Rather, we believe these perspectives are representative of a small sample of individuals that have been vocal and actively engaged in steering forensic science policy and practice over the last several years. Thus, while we must be careful not to over-generalize these individual viewpoints, we believe they provide valuable insights into the different perspectives affecting the current discourse in forensic science. Ultimately, we hope these insights provide a foundation for stakeholders to navigate a path forward that is cognizant and respectful of those different views, and generally amenable across all stakeholder groups. In the discussion that follows, we present a summary of the responses

compared across the different stakeholder groups along with salient observations and key points of view related to these issues.

6.5.1 Interpretation & Reporting Practices

Participants offered different perspectives related to the validity and/or appropriateness of reporting results categorically versus probabilistically. Prosecutors expressed views that categorical reporting was most appropriate, with most participants citing the ease of understanding and one participant (P#3) citing the benefit of categorical reporting as the certainty it conveys to fact-finders. Defense attorneys, academic scholars, and judges, however, expressed views suggesting that the certainty it conveyed was the very issue of concern, that categorical reporting conveyed a degree of certainty that was outside the realm of what can be scientifically supported and, therefore, was unsubstantiated and inappropriate. Laboratory managers, on the other hand, were more ambivalent to the issue. While laboratory managers recognized the concerns that have been raised related to categorical reporting, specifically, the propensity for categorical reporting to mask the underlying uncertainty in the conclusion, they found categorical reporting acceptable if practitioners caveated the claims as their opinion.

Reporting results probabilistically, however, was not embraced *carte blanch*e by any stakeholder group. All stakeholder groups expressed concerns that probabilistic reporting would be confusing and easily misunderstood by lay fact-finders. While prosecutors expressed the greatest hesitation to probabilistic reporting, all other stakeholder groups expressed views suggesting that probabilistic reporting was superior, in theory, to the alternative (of categorical reporting as it is traditionally expressed); however, probabilistic reporting would need to be carefully implemented to ensure the uncertainties and limitations of such conclusions were appropriately conveyed. Among those stakeholder groups that were receptive to probabilistic reporting, defense attorneys were most concerned about the extent to which the conclusions would be empirically supported by validated statistical methods and the risks that probabilistic expressions would be misused by prosecutors to imply greater certainty than warranted. Judges questioned the extent to which lay fact-finders and other legal actors would be able to meaningfully scrutinize the validity of the underlying statistical methodology, but recognized its utility to cause people to pause and carefully think through what is being conveyed. Laboratory managers acknowledged the benefits of probabilistic expressions and numerical references to provide stronger foundations to expert opinions; however, they suggested probabilistic statements should not stand-alone. Academic scholars offered the least consistent views, with one scholar expressing strong views in favor of the transition to probabilistic reporting (O#1), another scholar expressing a more ambivalent perspective, suggesting there was no other better alternative (O#2), and the third scholar, aligning most closely with defense attorneys, pointing to the need for black-box testing to assess applicable error rates related to the performance of practitioners overall as the most immediate need.

Overall, all stakeholder groups viewed the benefits of categorical reporting as the clarity and simplicity to convey and understand such statements. These findings were not surprising and generally consistent with social science literature on lay understanding of statistical references (e.g., see [108]). Except for prosecutors, who did not express concern of any risks associated with

categorical reporting, particularly under the auspice of an opinion, all other stakeholder groups suggested benefits were counterbalanced by the risk of making statements that were not scientifically valid or defensible (even under the auspice of an opinion) without some explanation around the uncertainties and limitations of the conclusion. On the other hand, most stakeholder groups viewed the benefits of probabilistic reporting as providing a means of conveying the uncertainties and limitations associated with the conclusion. However, all stakeholders noted that the extent to which those uncertainties and limitations are accurately represented depends on the extent to which such statements are based on empirical studies. None of the stakeholders expressed comfort with practitioners expressing conclusions probabilistically using numerical references without such numerical values being based on a validated statistical method. The chief concern being that the numerical values imply a level of precision and statistical basis to the assessment that cannot be substantiated. Thus, the so-called "subjective probabilities" approach, in which numerical values expressed are derived from subjective judgment rather than empirical measurement does not seem to be widely supported by stakeholders in the United States. That said, except for the majority of prosecutors, when sufficient statistical data is not available many of the other stakeholders responded in ways that suggested they were receptive to practitioners expressing conclusions probabilistically using qualitative statements without numerical references. These findings are generally consistent with guidelines set forth by the European Network of Forensic Science Institutes (ENFSI) in their Guidelines for Evaluative Reporting [110] as well as by the United Kingdom Forensic Science Regulator (UK FSR) in their Codes of Practice and Conduct: Development of Evaluative Opinions [127]. The ENFSI encourages numerical values be based on appropriate published statistical data, although as a "last resort" permits them to be based on subjective judgment [110]. The UK FSR, on the other hand, only permits numerical values be expressed if they are based on appropriate statistical data. In the absence of appropriate statistical data, the results shall still be expressed probabilistically but without numerical values [127]. Although the ENFSI and UK FSR advocate for likelihood ratios specifically, research has begun to explore how qualitative probabilistic statements should be phrased to ensure coherent interpretation by lay fact-finders (e.g., [107, 128]). Overall, though, this approach seems to be generally acceptable as an alternative to categorical claims and intermediary until validated statistical methods become accessible.

When presented with the findings from a recent study characterizing practitioners' perspectives related to probabilistic reporting, which found that approximately 80% of practitioners cited concerns that probabilistic reporting would be confusing to lay people and would be misused by defense attorneys to create "reasonable doubt" [51], the different stakeholder groups had mixed reactions. On the former issue, nearly all participants across every stakeholder group agreed that probabilistic reporting would be more confusing to lay people and agreed practitioners should take this into account when debating ways to express their conclusions; however, none of the stakeholder groups suggested the confusion would be insurmountable or was sufficient of a reason to completely oppose probabilistic reporting altogether. On the other hand, to the latter issue, all of the stakeholder groups were critical that practitioners would bear such a concern. Some even suggested that's the very purpose of the legal system, for example, as noted by one judge, "that's what the system is all about, is finding out whether there is, or is not a reasonable doubt" (J#1), and as noted by one scholar, "creating reasonable doubt is what defense lawyers are supposed to be doing" (O#2). Some participants suggested this finding illustrates a cultural bias that is believed to underlie many forensic science disciplines. Others, particularly

laboratory managers, while they did not personally support such a concern, recognized it to be an additional barrier that would need to be overcome if probabilistic reporting were to be adopted more widely by practitioners. How to overcome that concern, though, remains an open question. Separation of forensic science laboratories from law enforcement controls, as recommended by the NAS [3] could be a step toward mitigating such undercurrents, but greater understanding of the sources of such biases and the extent to which they can detract from sound scientific practices is needed.

One of the chief complaints with categorical reporting is that such expressions mask the uncertainties and limitations inherent in the interpretation of the evidence. Given the current discourse between categorical versus probabilistic reporting (e.g., [51]), we inquired what participants viewed as the roles and duties of forensic experts and the limits of their testimony. Admittedly, this question was attempting to elicit perspectives on a more technical nuanced issue, such as whether participants believed it was appropriate for experts to convey a statement about a proposition given a set of observations (i.e., a posterior probability about the source of an impression, or a decision that one proposition is true, such as "the two impressions were made by the same source") or whether experts' testimony should be more limited to a statement about the observations given a set of propositions (e.g., the observations provide strong support for the proposition the two impressions were made by the same source, and weak support for the proposition the two impressions were made by different sources). However, given the intentional broadness of the question to be careful not to unintentionally steer participants toward a particular response, we found that participants incidentally offered a similarly broad response. Not surprisingly, the responses across every stakeholder group were generally consistent—participants repeatedly echoed the need for forensic experts to accurately and impartially convey their findings and limit the testimony to what is supported by the science, ensuring that the conclusions are neither overstated or understated. Not a single individual disagreed with this sentiment; however, what was most interesting is that there seems to be little agreement as whether practitioners are adequately fulfilling these duties and how they should be conveyed. Defense attorneys expressed very explicit frustration that practitioners rarely take these duties seriously and elaborate on the full scope of the limitations. For example, one defense attorney stated outright: "In my 20+ years of litigating many forensic cases, I have never encountered a forensic examiner who took this duty seriously. … It is always a game of hide and seek for examiners" (D#2). On the other hand, laboratory managers expressed the desire to be transparent about the limitations but expressed challenges in doing so most effectively, and also pointed to litigators and the courts as a factor that makes it even more challenging to convey these details during testimony. One academic stakeholder, however, suggested that the issue might be more deeply rooted in culture, commenting: "I think we just need to be so careful not to try and be so helpful to the court in helping them to get rid of the uncertainty that they don't like [such] that we stray beyond what we can robustly and scientifically say. It's something that I would say I've observed anecdotally over the years. … I think it's dangerous" (O#1).

Related to this issue of disclosing limitations of their examinations, when participants were asked whether they find it acceptable for experts to express their opinion in court without disclosing the underpinnings or statistical data to support those opinions, the responses were divided across stakeholder groups. Although all participants suggested it strengthens the testimony when experts explain the basis for their conclusions, prosecutors and laboratory managers did not

believe it was necessary. Prosecutors pointed to their interpretation of statutory requirements as the guiding factor for their responses, and laboratory managers pointed to their past experiences. However, defense attorneys, academic scholars, and judges expressed counter views. Defense attorneys claimed such testimony would effectively be *ipse dixit* without such disclosure and is inadmissible under existing standards (despite courts allowing it in the past). Academic scholars recognized that different dynamics might affect testimony in a court setting but responded that such foundation was necessary from the perspective of sound scientific practice. Judges pointed to Daubert factors and prevailing admissibility standards suggesting such testimony should not be admissible; however, they recognized that many judges tend to admit it in anyways, referencing external pressures and past precedent. As one judge (J#3) explained, in general, "I think that some judges don't like to exclude. They'd rather let in than exclude and let it go to the jury. If there's an arguable basis for the jury to have accepted something, civil or criminal, then they [tend to] let it go to the jury. And that's a relatively safe place for them to be. If they exclude, they're subject to a reversal for an erroneous exclusion." To illustrate this even further, another judge (J#1) pointed to the *United States v. Llera-Plaza* decision [129], where the judge after only two months reversed his earlier decision that fingerprint evidence did not meet the Daubert standard. For example:

> *In his first opinion, [the judge] concluded [the fingerprint evidence] did not pass the Daubert standard. ... In the second [opinion], Judge Pollack withdrew his earlier objections. I think frankly, under intense pressure, and that's not a good thing. I hope I'm not being unfair to judge Pollak, but not much had really changed between the first opinion [and the] second opinion. He said things in the second opinion, like, "well, I've learned since that it's accepted by the courts of Great Britain, and, so I'm going to accept it." Who cares whether it's accepted by the courts of great Britain, they're not doing a Daubert analysis. The question is whether it meets Daubert or doesn't meet Daubert. So, I thought that was a cop-out and sort of revealing of the pressure he was under after his groundbreaking first opinion. [That aside,] I think that the evidence is there that fingerprint evidence is not junk science and that with proper limitations, it can be received in evidence (J#1).*

When asked what types of pressures might judges find themselves under, this judge offered an elaborated response pointing to political pressures, professional incentives, and biases to their own prosecutorial experiences. For example:

> *I will speculate. I should tell you, though, the statistics are quite striking. Daubert challenges succeed in civil cases frequently. They succeed in criminal cases almost never. And that shows, I think, that there is a double standard operating. So, why is that? One factor is that in most states trial judges are elected, and if they have to face re-election on the basis they are "soft on crime" because by God, I wouldn't even allow fingerprint evidence in, they're in trouble to be re-elected or even to be renominated by the party of their choice. So, election is an element, but I think a more subtle element is going on in most of these cases. The stakes are so much higher and judges, having seen the other evidence in the case, may think "yea, he's probably guilty, but you never know what a jury is going to do. If I keep out this evidence, maybe there won't be a conviction, and I really think it would be unfair to the prosecutor not to at least be able to present this evidence to the jury and they can take it for what it's worth." I think that is a wrongful attitude. I think*

I'd say a dereliction of duty and really ignores what Daubert is all about or even Frye for that matter. But, I do think that's a common traditional attitude: "I don't want to be responsible for this guy being acquitted, when, what I've heard so far, he's probably guilty." And of course, forensic evidence carries great weight. It has an aura of neutrality that you don't have from testimony of accomplices, for example. So, I think judges are reluctant to keep it out. I'll mention a third factor, which is that most criminal court judges are former prosecutors. Relatively few are former defense lawyers. So, there's also, "oh yeah, of course. I always let this in, I used to do it myself. This is just routine. I recognize this" (J#1).

This participant went further to provide another example and criticize a state Supreme Court decision in *Johnson v. Commonwealth* [130] in Kentucky that relied on judicial notice to admit microscopic hair analyses simply because it had been admitted previously without challenge. For example:

Some courts, well, I will take the liberty of criticizing a court with apologies, which is the Supreme Court of Kentucky, which had for many years the Frye standard, then it adopted Daubert. Then the question came along, whether microscopic hair analysis met a Daubert challenge and a federal court in Oklahoma had already held that it did not. So, the defense lawyer in the case, Johnson v. Commonwealth [130], a murder case, said we want to keep out this evidence, or at least we want a hearing, and the trial judge denied both and let in the evidence without a hearing. It went all the way up to the Supreme court of Kentucky, which held, with only one descending judge that, "well, all those years it came in under Frye and no one ever challenged it, so it must be good science." I think that's bad logic. So, they went so far as to say that a court in Kentucky can take judicial notice of the fact that microscopic hair analysis is good science, which is a terrible decision (J#1).

Finally, in wrapping up the broader topic of issues related to interpretation and reporting, we asked participants what they believed were the greatest challenges facing the pattern and impression evidence disciplines as it relates to examination and reporting methods. This question was intended to be a broad "catchall" question to allow participants to summarize what they believe to be the greatest need to support the pattern evidence disciplines moving forward. In response, participants pointed to a range of issues, often encompassing a scope much broader than just examination and reporting. Overall, however, defense attorneys, academic scholars, and judges all pointed to the need for more robust research establishing stronger empirical foundations and scientific rigor for many pattern evidence disciplines, including a better characterization of the limitations of those methods. Laboratory managers pointed to the need for additional resources to survive increasing caseload demands and to support foundational education and training needs for practitioners related to statistical issues and algorithmic tools that are being proposed. Prosecutors, on the other hand, pointed to partisan "attacks" from individuals or institutions attempting to undermine forensic evidence. Interestingly, this concern from prosecutors manifested throughout the interview and yielded an apparent contradiction. For instance, when responding to various questions throughout the interview, prosecutors were very deferential to scientists as to what they considered scientifically valid and appropriate as it relates to examination and reporting methods. For example, one prosecutor said quite explicitly:

[W]hat drives my decisions here is what is legitimate science and what are the scientists saying? Not as much of what are the lawyers saying about it? What are the scientists saying about it" (P#2)?

However, when prosecutors were asked if they were deferential to the scientists who have expressed concern over the validity and reliability of many forensic science methods, such as the President's Council of Advisors on Science and Technology (PCAST) [7], among others (e.g., [3, 8, 9]), prosecutors were quick to rebut the credibility of those reports and the individual authors. For example:

... like the PCAST report, which, as you can probably guess, I think is not worth the paper it was written on" (P#1).

I found that virtually everything about that [PCAST] report was suspect. I don't have trouble with the statement that forensic scientists should be conservative and careful. ... I think that seems self-evident, but if it was in the PCAST report, I don't think the report was honestly done (P#3).

These views expressed from the litigators during the interviews, however, are not completely unexpected and are generally consistent with those that have been expressed by prosecutors and defense attorneys more broadly. Shortly after the PCAST report was published, it stimulated a flurry of responses from professional organizations involved in the criminal justice system. The National District Attorney's Association (NDAA), for example, published a response representing 2,500 elected and appointed District Attorneys across the United States claiming "the NDAA takes issue with, and has substantial concern about, the logic of the [PCAST] report and the manner in which it portrays several forensic disciplines," citing "the pervasive bias and lack of independence apparent throughout the report" [131]. Similar responses were made by other professional forensic science organizations, including the American Society of Crime Laboratory Directors (ASCLD) [132] and the Association of Firearms and Toolmark Examiners (AFTE) [133], among others, which disagreed with several of the conclusions issued by the PCAST. Defense attorneys, on the other hand, welcomed the report with open arms; for example, the National Association of Criminal Defense Lawyers (NACDL) stated the report "offers further evidence of the pervasive use of flawed analysis erroneously presented as grounded in science" [134], and the Innocence Project claimed the report "provided a blueprint for fixing one of the most critical problems plaguing the criminal justice system" [135].

This sharp contrast between prosecutors and defense attorneys is not only evident from the published responses but has also been noted through anecdotal observations by academic scholars who have looked into the forensic sciences from neutral, outside perspectives. For example, one scholar, (O#1), commented during the interview about this "ongoing narrative of forensic science in crisis." This participant, (O#1), lamented that such narrative is "unhelpful," and forensic science "can still be of assistance to the courts, … but we must be honest about its limitations" (O#1). Another scholar, (O#3), noted "it's principally in the prosecuting community that I see the most resistance. … I understand why, I understand what you're saying, but you're not being completely honest with the jury if you say that" (O#3). When asked why this participant, (O#3), considers

prosecutors as the most resistant, the participant simply pointed to the adversarial nature of the criminal justice system. For example:

> *What you have is a back and forth between two sides, presenting evidence. The point of the exercise is to convince the majority of the triers of fact that my side has done better on the argument than yours. So, if you have a tool in that process of back and forth that lends more credence to the points that we're making than the other side, then you're not going to want to give that tool up. The way that forensic science is currently structured, mostly that tool is something that prosecuting attorneys can use (O#3).*

Related, during the interview, when one judge, (J#1), was asked their view on a comment made by one prosecutor, (P#3), that they view the greatest challenge facing the pattern evidence disciplines as "trying not to fold in the face of that kind of pressure [from people] attempting to appease the defense bar," (P#3), this judge, (J#1), responded by elaborating on the nature of the adversarial system and the emotion that runs high in the criminal justice system which only exacerbates such contrast between prosecutors and defense attorneys. For example, participant (J#1) stated:

> *Well, I don't know [that prosecutor] means other than [the prosecutor] thinks [they are] always right, and those defense counsel exercising the right of the defendant under the Constitution of the United States are evil people who are trying to pervert justice. But, it is of course true that every prosecutor has sooner or later a case in which they think a guilty person was wrongfully acquitted, and the nature of the adversary system, unfortunately, is you always impute the worst motives to your opponent. So, even in civil litigation, I'm confronted repeatedly, "judge, you won't believe what that guy on the other side did! It's outrageous! It's immoral! It's illegal! It's wrong!," and, usually it's some little squabble over nothing, but the emotions grab you. So, when the stakes are as high as they are in criminal cases, the emotions run even higher and you are very quick to impute to your adversary, "[they] only won that case cause [they] pulled the wool over the jury's eyes or whatever. One great privilege I've had as a judge is to talk to the jurors after each case, and I've had more than three hundred jury trials. There are some civil, but more criminal, and I am constantly impressed by how carefully juries take their obligation. They know the stakes in criminal cases are real and they take it very seriously. When I asked them, "well, how did you arrive at that decision?" They almost always give me good reasons. Occasionally I'll disagree with them. Like most judges, I'm more inclined to convict than to acquit if I were on the jury, but it's not because they're not giving me good reasons for the acquittal. And of course, acquittals are still a tiny, tiny fraction of the cases. To me, what should be bothering the prosecutors is the now indisputable proof that the system sometimes convicts innocent people. And who's responsible for that, Mr./Ms. Prosecutor, if not you? So, the very familiar word of Justice Jackson, when he was Attorney General of the United States, I still think should ring in every prosecutor's ear, which is [paraphrasing] "your job is to do justice, not to convict, not to exercise hunches, but to make sure that you have analyzed every case carefully and then go forward if you can objectively say that you have proved beyond a reasonable doubt. Not to view it as this is a competition, a game, an adversary process." Hard to avoid that in an adversary system, but I still think that's the right attitude for a careful prosecutor (J#1).*

From these results we see that the pattern evidence disciplines are facing a myriad of perspectives from various stakeholders across the criminal justice system as they relate to interpretation and reporting methods in the pattern evidence disciplines. Overall, it appears that prosecutors' perspectives represent one extreme end of a spectrum and defense attorneys' perspectives represent the other extreme, particularly as they relate to the validity and appropriateness of traditional practices. Broadly speaking, prosecutors expressed the desire for practitioners to adhere to good scientific practices and argue that existing methods are appropriate. On the other hand, defense attorneys argue that existing methods go beyond the standards of good scientific practice, are invalid, and are inappropriate. Of course, this opposition is not completely unexpected given the adversarial nature of the American legal system. Responses from laboratory managers, judges, and academic scholars seemed to be less extreme, but still represented an affinity toward one side of the spectrum compared to the other. Perspectives from laboratory managers tended to align more closely with prosecutors in the sense that they maintained perspectives that traditional practices were acceptable (although maybe not ideal); however, judges and academic scholars tended to align more closely with defense attorneys in the sense that they were more overt with their concerns as it relates to traditional practices. Comparing these results to those of pattern evidence practitioners (e.g., see [51]), we would conclude that practitioners' perspectives align most closely with laboratory managers. For example, from [51], we see most practitioners tend to maintain the perspective that traditional reporting methods using a categorical framework are appropriate and defensible, although there is a growing minority that believe probabilistic methods are a more suitable alternative. All stakeholders, however, including practitioners (i.e., see [51]), expressed concern that probabilistic reporting methods will bring new challenges that have yet to be fully explored.

Despite the nature of the discourse and diverse perspectives on this broader issue of interpretation and reporting, it seems that there were some areas in which stakeholders offered shared perspectives, particularly in relation to the benefits and limitations/risks of categorical reporting versus probabilistic reporting. Nearly every stakeholder recognized the need for forensic conclusions to be scientifically defensible *and* easily interpretable. The major critique of categorical reporting is that it is not scientifically defensible (at least, how it is traditionally expressed), but it is easily interpretable. The major critique of probabilistic reporting is that it can be more scientifically defensible[26]; but it is not as easily interpretable. Although we recognize that no single approach will satisfy all stakeholders, perhaps an immediate next step for the community to consider is a combination of the two. Admittedly, given the discourse on this subject to date, going into the interviews we held a belief that probabilistic reporting was going to be considered the ideal by many stakeholders, particularly defense attorneys. Although most stakeholders did express the superiority of probabilistic reporting over categorical reporting, the responses left us skeptical as to whether the superiority was truly because of the benefits of probabilistic on its own, or merely because it was the better of the two without any other alternative when presented in a binary context. Recognizing that defense attorneys represent an extreme end of the spectrum in terms of the various perspectives, suggesting traditional categorical claims were inappropriate, interestingly, they did not seem to wholly endorse probabilistic reporting outright as the preferred alternative. Instead, defense attorneys were most concerned about ensuring the limitations of the

[26] Provided the basis for the probabilistic statement is disclosed (e.g., human judgment versus statistical methods), along with those limitations and numerical references are empirically demonstrable.

methods (and all sources thereto[27]) are clearly explained. In that sense, probabilistic expressions that account for the strength of evidence are relevant, but information related to empirical measures of error rates[28] were equally, if not more, important in their view.

6.5.2 Use of Algorithms

Participants offered generally consistent perspectives related to the use of algorithms in court and the benefits and limitations of them. All stakeholders were receptive, at least in theory, to the use of algorithms, and pointed to several benefits algorithms could provide, such as better means of reflecting the strength of evidence, promoting greater objectivity and consistency in examination results, and enabling examinations to be performed more efficiently. Stakeholders differed, however, as to how they viewed the limitations of algorithms. Prosecutors, while generally receptive to algorithms, questioned whether they were truly necessary compared to traditional methods in pattern evidence disciplines (versus their necessity for DNA interpretation, for example). Prosecutors seemed to be most concerned whether algorithms would unduly complicate reporting and testimony, making it more difficult for lay fact-finders to understand the testimony. Laboratory managers, defense attorneys, academic scholars, and judges, on the other hand, were most concerned about the transparency surrounding these systems, the underlying validity of the systems, and the risks of analysts and lay fact-finders blindly relying on the output of algorithmic tools without fully understanding and accounting for their limitations. Laboratory managers were concerned about delegating decision-making responsibilities from the analyst to the algorithm, suggesting that algorithms would be most useful as tools to supplement their judgment rather than supplant their judgment. Defense attorneys were most frustrated about the lack of access and proprietary protections that have been placed around algorithms in the past preventing their disclosure of the underlying source code. Academic scholars were most receptive to algorithms provided they were sufficiently validated but speculated about whether algorithms can be thoroughly validated such that all limitations and boundary conditions are known, particularly as the algorithms become more complicated and less transparent. Judges were most concerned about ensuring algorithms were trustworthy, reliable, and fair, and that defendants are afforded the opportunity to challenge the evidence against them and exercise their due process rights granted under the Constitution.

A salient theme in the conversation about algorithms was how they could be trusted for use in court thereby having an impact on human liberties. All stakeholders suggested that trust requires that the algorithm be validated, and validation requires that the algorithm be shown to be "reliable" through performance testing. Academic scholars, defense attorneys, and judges, however, suggested that trust also requires that the algorithm be shown to be "fair," which may not necessarily be determined through performance testing alone. These stakeholders, in addition to some laboratory managers, went a step further and pointed out more nuanced details that are required, in their views, for an algorithm to be appropriately validated. The design and conceptual operations of the algorithm must also be understood to ensure that the validation testing is appropriately designed and that the boundary conditions for which the algorithm is able to

[27] In this sense, we distinguish between the uncertainty of an association based on a coincidental match and the error rate based on the performance of the expert.
[28] Error rates in this context refer to blind performance testing of examiners under normal casework conditions.

appropriately function can be established, such that the conditions for which the algorithm is expected to work well and the circumstances, or combination of circumstances, for which the algorithm begins to fail are known. Defense attorneys, judges, and academic scholars all noted that these details lie within the source code. Academic scholars recognized the value of source code but held back from suggesting source code was the only means by which that information could be ascertained. Defense attorneys and judges, on the other hand, argued that the source code and the software application containing the algorithm were critical to permit an independent evaluation and testing under conditions they consider appropriate given the circumstances of the case at hand. Consequently, these stakeholders expressed strong views that disclosure of source code is necessary to enable criminal defendants to mount an adequate defense, and failure to provide access to these materials could be considered an infringement of the defendant's Constitutional right to due process. As one judge (J#3) commented: "When we're dealing with due process and equal protection under the United States Constitution, we are now in a world where 'the means to the end' matter, [and] the means are contained within the source code." Concerns about the capacity to meaningfully scrutinize algorithmic tools in the absence of source code and its impact on criminal defendants' Constitutional rights is not limited to these participants. It is a perspective that is held more generally (e.g., see [114, 115, 117, 118]). Although prosecutors, laboratory managers, and academic scholars did not express such a strong view on the necessity for disclosure of source code as defense attorneys and judges did, all stakeholders suggested they would be amenable to the disclosure of the source code if desired by the defendants and they were in a position to disclose it.

The issue with disclosure, however, is that some commercial vendors of algorithms have exerted trade secret protections to prevent such disclosure, and some courts have therefore been faced with balancing these countervailing positions. When stakeholders were asked how courts should address these issues, nearly every single participant pointed specifically to protective orders or described a level of protection that is comparable to a protective order. Many stakeholders, particularly prosecutors, defense attorneys, and judges, were quick to dismiss trade secret protections as even being an issue given that there are existing mechanisms for protecting intellectual property, but also suggested the ideal situation is that these algorithms are open and publicly accessible without the need for such court orders, given that they are being used for human liberty decisions. Defense attorneys and judges both pointed to civil litigation as examples of established precedent and procedures for how to permit disclosure while still protecting intellectual property concerns from commercial vendors. Given these perspectives, when judges were asked why courts have failed to mandate disclosure, some judges were openly critical of those rulings. One judge (J#3) went so far as to claim they were "wrongly decided," for example:

> *I think they're wrongly decided. ... We do know how it's done, and there's a whole body of case law that can be utilized from the civil side and transferred over. So, it is possible that what we're seeing is just a lack of experience by some of the state court criminal judges with the disclosure of the super-secret stuff. The federal judges ought to know how it's done, because we did it all the time, and they do it all over (J#3).*

Interestingly, one defense attorney (D#3) suggested that prosecutors should not proffer evidence from an algorithmic tool unless they can disclose the source code under discovery. When presented with this perspective, prosecutors claimed that if they did have access to the source code,

then it would be disclosed under existing discovery rules. However, they often do not have access to it but also don't believe it is critical to possess before proffering such evidence. When asked how prosecutors can assure the evidence they are proffering from an algorithmic tool is trustworthy without having access to the source code, the prosecutors often pointed to the forensic scientists for such assurances based on their validation testing. When laboratory managers were asked whether disclosure of source code was a factor that they considered when procuring an algorithmic tool, they all claimed it would be taken into account, but was not a governing factor in the decision. Laboratory managers tended to be more focused on the performance characteristics and capabilities offered by the algorithmic tool versus issues related to disclosure but recognized the benefits of procuring an algorithm for which the vendor was willing to disclose the source code when requested.

Having discussed what stakeholders considered were necessary for algorithms to be trusted and the role of source code in that assessment, stakeholders were asked their opinion about the use of algorithms based on AI/ML methods, which are often "black boxes," even to their developers, and that human interpretable source code is effectively nonexistent. While the specific responses varied between individuals both within and between stakeholder groups, most individuals across all stakeholder groups expressed even more caution and skepticism with the use of AI/ML algorithms compared to their existing concerns related to "traditional" algorithms based on straight programming and rule-based approaches (i.e., non-AI/ML-based algorithms). Although most individuals were receptive to the idea of using AI/ML algorithms in theory, participants from each stakeholder group expressed a number of concerns about their transparency and whether they can be sufficiently tested such that the boundary conditions are known to permit an appropriate validation for practical application. Laboratory managers recognized the additional complexity of these types of algorithms but seemed to be the most receptive to their use provided they were sufficiently validated and demonstrated superior performance characteristics. Academic scholars were the most concerned about the practicality of performing all the testing that would be necessary in order to fully trust the algorithm and whether such testing was practical, for example, as one scholar commented: "I don't think the idea is actually practical [and] realizable. I don't think you could actually implement it" (O#3). Prosecutors, defense attorneys, and judges expressed an additional layer of caution with reference to potential concerns on a Constitutional dimension, such as whether the application of these algorithms could be an infringement on due process and confrontation rights. Ultimately, these stakeholders suggested that the admissibility of these types of algorithms would require careful consideration on a case-by-case basis depending on what information was available about the algorithm, such as design, inputs, parameters, weightings, training data, validation data, etc., and how those details relate to the circumstances in the case at hand. The various responses from these participants across all stakeholder groups, and general hesitation concerning the use of AI/ML algorithms overall, illustrate that many of these issues have yet to be fully fleshed out. Indeed, it is a novel subject that legal scholars are just beginning to discuss (e.g., see [118]). For example, as one judge (J#3) made clear:

> *What we need is a national conversation on what that means and how to create trustworthy and reliable algorithms that can be used for individual liberty determinations. That's where the rubber meets the road. ... The greatest risk is that we allow complex design and complex tools to just snow us a little bit ... [and] that we don't have these conversations as to what fairness means and what fair design is and what trustworthiness is in time (J#3).*

Given the concerns that have been expressed about the use of algorithms, participants were asked their opinion about whether they should be regulated, and, if so, how. This question stimulated several diverse responses, and some stakeholders took this opportunity to express their views on the regulation of forensic science more broadly. Prosecutors and laboratory managers tended to be deferential to the forensic science community to establish applicable guidelines in a centralized and coordinated fashion, but then leave it to the legal community to enforce those guidelines, where appropriate, on a case-by-case basis. These participants seemed to be generally satisfied with how the legal system has regulated forensic science practices more broadly and believed the legal system would be similarly effective with the regulation of algorithms. Defense attorneys, academic scholars, and judges, however, rejected the idea that the legal system could be effective at regulating algorithms. For example, these participants went so far as to claim the legal system "has proven to be an utter failure" (D#2) and "defective" (J#1) in its ability to regulate forensic science more broadly. Instead, these participants suggested algorithms, and forensic science overall, should be regulated by an independent entity with both oversight responsibilities and approval authorities. Participants, however, were not as aligned on whether the entity should be part of the federal government. Academic scholars also pointed out that such regulation should address the entire method rather than just the algorithm itself (e.g., inputs, personnel, testimony, etc. associated with the use of the algorithm). The perspectives from defense attorneys, academic scholars, and judges are not limited to just these individual participants; instead, they generally align with one of the key recommendations from the 2009 NAS report [3], which is to create an independent Institute of Forensic Sciences staffed by high level scientists, which one judge (J#1) lamented never received enough traction to materialize. Although academic scholars recognized the OSAC as a step in the right direction, there were mixed perspectives as to whether the OSAC is able to provide the central coordination desired by some or assess the appropriateness and rigor behind the validation of forensic methods. To illustrate the concerns related to the topic of regulation more broadly and to highlight the need for greater consistency across the forensic science community as it relates to resources and practices, when one academic scholar (O#3) was asked whether they trust forensic science overall today based on what they have observed as an "outside scientist" over the last several years, they responded:

> *It depends on where I am. Literally. It literally depends on my physical location, because if I'm in a location where I have some confidence that the forensic scientists are appropriately supported with the proper amount of resources to perform at the highest level, yes, I would trust them. [But,] if I'm someplace where that's not the condition, [then] no, I'm not going to trust them (O#3).*

Finally, in wrapping up this broader topic related to the use of algorithms, we asked participants what they believed were the greatest challenges facing the operational use of computational algorithms for court purposes. This question was intended to be a broad "catchall" question to allow participants to summarize what they believe to be the greatest need to support the pattern evidence disciplines as algorithms become more available. In response, participants pointed to a range of issues, often encompassing a scope much broader than just the use of algorithms. Overall, all stakeholder groups (except for judges)[29] pointed to the need for greater investments in foundational education and training for the forensic science and legal

[29] Due to time limitations, judges (J#1, J#2, and J#3) did not provide a specific response to this question.

communities—specifically practitioners who will be expected to use the algorithms and judges who will be expected to assess the admissibility of the algorithms. Prosecutors also pointed out the need for ensuring the algorithms are developed in a way that can be effectively explained in lay terms to fact-finders, recognizing that the more complicated computational methods become, the more challenging it is to present scientific evidence in court. Laboratory managers expanded on the need for better training and echoed their prior concerns related to lack of resources to support the validation and implementation into day-to-day practice. Academic scholars also pointed to the need to be clear about what type of algorithms are being considered (i.e., traditional rule-based programmed algorithm vs. AI/ML-based algorithm) so that stakeholders have a common understanding of the varying benefits and risks surrounding their use, the need to consider changing recruitment and selection of forensic practitioners to those who have higher aptitudes in physical sciences and mathematics, and the need for safeguards, standards, and oversight to be placed around the use of commercially developed algorithms to prevent financial interests from impacting the quality of their development and validation.

From these results we see that the use of computational algorithms in court is a complicated issue. While all stakeholders across the criminal justice system were welcoming of the potential benefits that algorithms can provide, they all expressed caution about the risks associated with them and the need to carefully consider the more nuanced details around their development and implementation—the central issue being how algorithmic tools can be trusted for court purposes that can directly impact human liberty decisions. We find that trust is a complicated and multi-dimensional concept, and stakeholders have similar but inconsistent and incomplete perspectives on what that entails. Overall, stakeholders held a variety of perspectives on these issues related to the use of algorithms, but all expressed a shared desire to ensure these systems are developed and implemented in a responsible and practical manner that upholds the values of fairness and equal justice under the law. How this can be done in a structured and consistent way requires a broader national dialogue. In recent work, we have begun to explore this in greater detail based on perspectives that have been raised in the literature thus far and provide some initial recommendations [53]. The perspectives expressed in the present study provide greater breadth and depth to these issues. While we believe they align with those that have been raised thus far in the literature, this study reinforces the need for this conversation to occur sooner rather than later. It will only be a matter of time until these algorithmic tools are introduced for court purposes, and it is critical that we have a shared perspective and mutually agreeable framework for how to address these issues before we find ourselves in a legal quandary.

It should also be noted that participants' experiences related to algorithms were widely variable. Although all participants had direct knowledge and experience dealing with algorithms in the criminal justice system in one capacity or another (e.g., related to their use, development, validation, or litigation), many of the questions related to this issue required participants to speculate in general terms without a single specific algorithm to point to, and were focused on the use of algorithms for court purposes which have direct impacts on decisions impacting human liberty. Some participants noted a distinction with the use of algorithms for other purposes in the criminal justice system, such as for investigatory leads or general purposes to augment traditional policing practices, and recognized that their perspectives might vary as it relates to algorithms designed for those purposes, since their benefits and risks can be very different. Although issues concerning the use of these types of algorithms were outside the scope of this evaluation, it is

relevant to note that some stakeholders suggested the risks associated with the use of algorithms for those purposes can be much lower compared to the risks associated with the use of algorithms for human liberty decisions. Other stakeholders, particularly defense attorneys, however, asserted that algorithms used for investigatory purposes should be held to the same standards as algorithms intended for court. Although this perspective was not broadly shared across other stakeholder groups, the primary concern expressed from defense attorneys is that these types of algorithms eventually make their way into court and once they do, they can significantly influence fact-finders' decisions that impact human liberty. Considering the nuances that often impact stakeholders' perspectives on these issues, additional research is needed to explore the implications of algorithms used for purposes other than court, such as investigatory purposes, to better understand whether, and under what circumstances, they could be used that are generally amenable across stakeholders.

6.6 Conclusion

Over the last decade, there have been increasing calls for the introduction of probabilistic reasoning and validated statistical methods into forensic practice—particularly in the pattern evidence disciplines—to formally recognize and articulate the uncertainties inherent in forensic interpretation and reduce the heavy reliance on subjective judgment. While probabilistic reasoning can be achieved without the need for sophisticated technology, computational algorithms are often a means by which empirical measurements are made and probabilistic values are assigned to the evidence. In recent years, various approaches have been proposed. However, reactions to probabilistic reporting and the use of computational algorithms in forensic science have been mixed. Some commentators have argued that probabilistic reporting and computational algorithms promote more scientifically defensible reports and provide more objective and greater scientific capabilities to the evaluation of forensic evidence. Others, however, have argued probabilistic approaches unduly complicate the issue, and the opacity of algorithmic tools makes it challenging to meaningfully scrutinize the evidence. Consequently, the forensic community has been left with no clear path forward on how to navigate these mounting concerns as each proposed solution seemingly has countervailing benefits and risks. In order to better understand these issues, this study elicited the perspectives of key criminal justice stakeholders, including forensic laboratory managers, prosecuting attorneys, defense attorneys, judges, and other academic scientists and scholars on issues related to (i) interpretation and reporting practices (with or without algorithmic tools) and (ii) the implications of the use of computational algorithms as a means of calculating the probabilistic values assigned to forensic science evidence in the American legal system. This study was conducted as one-on-one semi-structured interviews of fifteen individuals (three from each stakeholder group) resulting in over twenty hours of recorded interviews and over three hundred pages of written transcripts capturing their perspectives on these issues. Although the number of individuals from each stakeholder group prevents broad generalizations, these individuals are considered prominent in their fields and have various marks of distinction, such as occupying senior level roles in their disciplines, served on boards and committees steering policy and practice recommendations, and are influential in the practices of others across the broader community, either directly through supervision or indirectly through training and continuing education activities. Participants' responses were rich with information illustrating their diverse

viewpoints on various issues and providing valuable insights into the different perspectives affecting the current discourse in forensic science.

As it relates to interpretation and reporting practices, we found that the pattern evidence disciplines are facing a complex myriad of perspectives that has effectively stifled the ability to find consensus on nearly every issue. Generally speaking, prosecutors' perspectives often represented one extreme end of a spectrum and defense attorneys' perspectives represented the other extreme. Perspectives from laboratory managers tended to align more closely with prosecutors in the sense that they maintained perspectives that traditional practices were acceptable (although maybe not ideal); however, judges and academic scholars tended to align more closely with defense attorneys in the sense that they were more critical and expressive of their concerns as it relates to traditional practices. Nearly every stakeholder recognized the need for forensic conclusions to be scientifically defensible *and* easily interpretable. However, stakeholders differed on how that should be accomplished. Further, although stakeholders generally agreed on the roles and responsibilities of experts and the importance of ensuring opinions expressed during testimony are accompanied by the underpinnings or statistical data to support those opinions, they differed in their views related to whether forensic practitioners are adequately fulfilling those roles and responsibilities and whether disclosing that information is necessary from scientific and legal perspectives.

As it relates to the topic of the use of computational algorithms in court, we found that stakeholders recognize their potential benefits and, in theory, were receptive to their use. Generally, stakeholders pointed to the benefits algorithms provide as being a better means of reflecting the strength of evidence, promoting greater objectivity and consistency in examination results, and enabling examinations to be performed more efficiently. Stakeholders differed, however, how they viewed the limitations of algorithms. Prosecutors seemed to be most concerned whether algorithms would unduly complicate reporting and testimony making it more difficult for lay fact-finders to understand the testimony. Defense attorneys, judges, academic scholars, and laboratory managers, on the other hand, were most concerned about the transparency surrounding these systems, how to ensure the underlying validity of the systems, and the risks of analysts and lay fact-finders blindly relying on the output of algorithmic tools without fully understanding and accounting for their limitations. These concerns highlight the need to carefully consider the more nuanced details around their development and implementation—the central issue being how algorithmic tools can be trusted for court purposes that can directly impact human liberty decisions. However, we find that trust is a complicated and multi-dimensional concept, and stakeholders have similar but inconsistent and incomplete perspectives on what that entails. Overall, despite stakeholders having a variety of perspectives on these issues related to the use of algorithms, they all expressed a shared desire to ensure these systems are developed and implemented in a responsible and practical manner that upholds the values of fairness and equal justice under the law. How this can be done in a structured and consistent way requires a broader national dialogue to occur sooner rather than later. In our view, computational algorithms are now beginning to be introduced for court purposes, and it is critical that we have a shared perspective and mutually agreeable framework for how to address these issues before we find ourselves in a legal quandary.

Looking forward, participants pointed to several challenges facing the forensic science community. First and foremost, there is a need for more robust research establishing stronger

empirical foundations and scientific rigor for many pattern evidence disciplines, including a better characterization of the limitations of those methods. As this research develops, and computational algorithms become more accessible, however, the challenges will become even more complex. As we consider the use of computational algorithms, we need to be sensitive to the diverse perspectives related to their use from different stakeholders operating within the criminal justice system. Overarching all else, there is a need for greater investments in foundational education and training for the forensic science and legal communities—specifically practitioners who will be expected to use the algorithms and judges who will be expected to assess the admissibility of the algorithms, as well as greater allocation of resources for forensic laboratories to support these investments while maintaining the caseload and throughput demanded of them. Second, we need to be conscientious that these algorithms need to be understandable and explainable to lay fact-finders, recognizing that the more complicated computational methods become, the more challenging it is to present scientific evidence in court, and that starts with how the algorithms are designed and developed. Third, we need to be clear about what type of algorithms are being considered (i.e., traditional rule-based programmed algorithm vs. AI/ML-based algorithm) so that stakeholders have a common understanding of the varying benefits and risks surrounding their use. Fourth, we need to consider changing recruitment and selection of forensic practitioners to those who have higher aptitudes in physical sciences and mathematics. Finally, we need for policy-safeguards, standards, and oversight to be placed around the development, validation, and application of forensic science methods, including algorithmic tools. Overall, these growing concerns and diverse perspectives illustrate a need for additional research and a national conversation to continue across the criminal justice community on how to navigate a path forward most effectively in a manner that is both cognizant and respectful of the different views and generally amenable across all stakeholder groups. Until that occurs, we can expect growing divisiveness and continued frustration amongst different stakeholders as we seek a more effective administration of justice.

7 Implementation of Algorithms: Challenges, Considerations, and a Path Forward

This chapter presents a manuscript entitled "Implementation of Algorithms in Pattern & Impression Evidence: A Responsible and Practical Roadmap" (Swofford & Champod, 2021) [53] published in *Forensic Science International: Synergy* that discusses challenges, considerations, and a path forward for the implementation of algorithms in pattern and impression evidence domains. The paper explores human-algorithm interactions and seeks to understand *why* practitioners (in general) tend to oppose algorithmic interventions and *how* their concerns might be overcome. Further, it addresses issues concerning to human-algorithm interactions in both real-world domains and laboratory studies as well as issues concerning the litigation of algorithms in the American legal system. With these considerations in mind, the article proposes a strategy for approaching the implementation of algorithms, and a taxonomy describing the various ways algorithms can be implemented, in a responsible and practical manner.

Implementation of Algorithms in Pattern & Impression Evidence:
A Responsible and Practical Roadmap

Swofford, H. and Champod, C.

School of Criminal Justice, Forensic Science Institute, University of Lausanne, Switzerland

7.1 Abstract

Over the years, scientific and legal scholars have called for the implementation of algorithms (e.g., statistical methods) in forensic science to provide an empirical foundation to experts' subjective conclusions. Despite the proliferation of numerous approaches, the practitioner community has been reluctant to apply them operationally. Reactions have ranged from passive skepticism to outright opposition, often in favor of traditional experience and expertise as a sufficient basis for conclusions. In this paper, we explore *why* practitioners are generally in opposition to algorithmic interventions and *how* their concerns might be overcome. We accomplish this by considering issues concerning human-algorithm interactions in both real world domains and laboratory studies as well as issues concerning the litigation of algorithms in the American legal system. Taking into account those issues, we propose a strategy for approaching the implementation of algorithms, and the different ways algorithms can be implemented, in a *responsible* and *practical* manner.

Keywords: Forensic science, Pattern evidence, Algorithms, Statistics, Models, Automation

7.2 Introduction

Over the years, the forensic science community has faced increasing criticism from scientific and legal scholars, challenging the validity and reliability of many forensic examination

methods that rely on subjective interpretations by forensic practitioners [1-9]. Of particular concern is the lack of an empirically demonstrable basis to substantiate conclusions from pattern and impression evidence, which has led to calls for reform through the development and integration of tools to evaluate and report the strength of forensic evidence using validated statistical methods [3, 7-9]. Some, such as the President's Council of Advisors on Science and Technology (PCAST), have gone so far as to suggest forensic analyses should be fully objective such that "they can be performed by either an automated system or human examiners exercising little or no judgment" [7]. As illustrated by the PCAST, algorithms and automation are often proposed as a natural solution to the limitations of human judgment. Although concerns over subjective interpretation and lack of statistical evidence span across most pattern and impression evidence domains, the practice of friction ridge examination is often a focal point of debate due to its long-standing history and ubiquitous practice. In the friction ridge discipline in particular, there have been a number of notable efforts by researchers for which algorithms have been introduced to provide quantitative or statistical approaches to the analysis and evaluation of evidence [17-43, 50]. Despite the proliferation of proposed methods, however, the practitioner community has been reluctant to apply them operationally. Reactions toward the intervention of statistical methods, even statistical *concepts*, have ranged from passive observation and skepticism to outright opposition [46-48, 51]. Reasons cited for these reactions are expansive—often resulting in an excuse for the whole-sale rejection of available methods in favor of traditional experience and expertise [51]. Fueled by perceptions that algorithms can only be implemented as an "all or nothing" approach (*either* the human *or* the algorithm), this has led to controversy and opposing viewpoints between many in the scientific and practitioner communities which have ultimately created a stalemate. On the one hand, opponents of algorithmic interventions might point to anecdotal instances in which the algorithms have failed as proof that technologies are inferior and highlight concerns of new challenges that have yet to be understood and fully explored when algorithms are implemented without proper scrutiny, training, oversight, and quality controls [136]. On the other hand, proponents of algorithmic interventions might too easily brush off negative reactions toward algorithms and characterize them as irrational or hyperbolic seeking to maintain traditional practices and preserve autonomous decision making.

What has become clear over the last decade is that calls for algorithms in forensic science are unlikely to subside and challenges to implementation are unlikely to be solved by improvements to technology or the mere proliferation of the tools alone. The problem is much more complicated and requires careful consideration of different issues. First, we need to take a step back and better understand *why* practitioners are hesitant to rely on algorithms and *how* their concerns might be overcome to increase receptivity. This will require us to look outside forensic science in other domains where algorithms have been introduced and consider the issues through the lenses of psychology and behavioral sciences as it relates to human-algorithm interactions. Then, we need to consider the environment in which forensic science operates and to which the algorithms will ultimately be applied—the criminal justice system—and the impact algorithms can have on sensitive decisions impacting life and liberty of citizens. Finally, with these contexts in mind, we need to consider how to mitigate the concerns of forensic practitioners and criminal justice stakeholders and navigate a way forward for the implementation of algorithms into forensic science in a *responsible* and *practical* manner. In most circumstances, hastily jumping from no algorithmic influence, which represents the current state of forensic science today, to complete automation, as envisioned by PCAST, without a clear roadmap and consideration of the complex

and dynamic issues at play is both irresponsible and impractical. To this end, in this paper we will (i) outline the foundations that need to be in place from a quality assurance perspective before algorithms should be implemented, such as education, training, protocols, validation, verification, competency, and on-going monitoring schemes, and (ii) propose a taxonomy of six different levels of algorithm implementation ranging from Level 0 (no algorithm influence) to Level 5 (complete algorithm influence) describing various ways in which algorithms can be implemented. In levels 0 through 2, the human serves as the predominant basis for the evaluation and conclusion with increasing influence of algorithms as a supplemental factor for quality control (used *after* the expert opinion has been formed). In Levels 3 through 5, algorithms serve as the predominant basis for the evaluation and conclusion with decreasing influence from the human. We note that this taxonomy is distinct from levels of technology readiness often used to describe the maturity of technology for operational deployment (e.g., see [137]); the levels outlined in our proposed taxonomy applies to algorithms that have been validated and are ready for operational deployment. This taxonomy, therefore, not only provides a common foundation to communicate what it means for an algorithm to be implemented and the degree to which algorithms influence the overall outcome of the evaluation at each level, but it also provides a framework for forensic science disciplines to implement algorithms in a deliberate and progressive way that is considerate of the implications algorithms will have on traditional examination practices as well as the criminal justice system and its stakeholders.

In the discussion that follows, we take an agnostic viewpoint of any specific method and instead frame the issue on the topic of integrating algorithms (in general) into domains that are largely driven by human judgment. For these purposes, the term "algorithm" is used to broadly describe any evidence-based prediction method, such as statistical models, decision rules, and other mechanical processes used for forecasting, predictions, statistical evaluations and decision making. We approach this discussion in five parts. In Part I, we start by taking a retrospective look at the challenges faced with the initial introduction of algorithms into the scheme of clinical decision making, with particular emphasis on medical practitioners—a domain we consider a reasonable proxy for exploring issues related to human-algorithm interaction in forensic science. In Part II, we discuss issues concerning human-algorithm interactions more generally and summarize key research findings from psychology and behavioral sciences regarding the tendency for people to rely on algorithms and factors that are believed to increase or decrease those tendencies. In Part III, we consider the generalizability of the research findings in the context of two real-world domains that have traditionally relied on human judgment based on intuition and experience and where human-algorithm interactions have naturally begun to take shape: medicine and autonomous vehicles. In Part IV, we discuss specific challenges related to the introduction of algorithms into the American legal system. Finally in Part V, we build on the discussion from prior sections and propose a path forward for the integration of algorithms into forensic practice that is believed to increase the likelihood for adoption across all stakeholders and lead to an overall stronger foundation and improvement to the quality and consistency of forensic science. We note that in the various parts throughout this review, we provide several (sometimes lengthy) quotations from key papers. This is done to ensure we do not distort the authors' original positions or views related to specific issues discussed and to enable readers to discern the similarities and applicability to the current state of forensic science.

7.3 Part I: The Introduction of Algorithms in Clinical Decision Making

Leading up to the 1950's, there were growing debates in the scientific and medical communities on the superiority of predictions made on the basis of clinical judgment (e.g. subjective, experience-based) vs. statistical methods (e.g. algorithmic, actuarial). Theoretical arguments divided the two communities (often down parting lines of clinicians vs. statisticians) and proponents for each paradigm asserted the answers were 'obvious' [138]. In 1954, Paul Meehl, a clinical psychologist, explored this issue and published a landmark book entitled *Clinical versus Statistical Prediction: A Theoretical Analysis and Review of the Evidence* in which he considers the theoretical arguments from both sides and reviews results of twenty different forecasting studies across diverse domains, including academic performance and parole violations. This was the first known empirical study comparing the relative performance of clinical judgment versus statistical methods (e.g. linear models) for prediction tasks. Meehl finds that predictions based on statistical methods consistently outperformed those based on the judgment from skilled human counterparts [138]. Shortly after publication, Meehl's findings were met with skepticism by other clinical experts. In his book *Thinking Fast and Slow*, Kahneman recounts "[f]rom the very outset, clinical psychologists responded to Meehl's ideas with hostility and disbelief" [139]. In the years that followed, Meehl's work stimulated a proliferation of research on the topic of clinical versus statistical methods for prediction tasks. Study after study, researchers repeatedly reported the superiority of algorithms versus humans [140, 141]. Grove et al. (2000) provides a meta-analysis on 136 studies over the last four decades and finds overwhelming evidence demonstrating statistical methods performing on par with or better than human judgment across a variety of domains, including medical diagnoses, mental health, psychology, academic success, parole violations, business operations, personnel decisions, and more [141]. The authors summarize their findings as "[o]n average, mechanical-prediction techniques were about 10% more accurate than clinical predictions" and "[s]uperiority for mechanical-prediction techniques was consistent, regardless of the judgment task, type of judges, judges' amounts of experience, or the types of data being combined" [141]. Thirty years after his original publication, Meehl published a commentary regarding his original 1954 publication in which he recounts the reactions of his fellow clinicians suggesting he was "fomenting a needless controversy" and offers his views following three decades of reflection [142]. In his commentary, Meehl notes [142]:

> *There is no controversy in social science that shows such a large body of qualitatively diverse studies coming out so uniformly in the same direction as this one. When you are pushing 90 investigations, predicting everything from the outcome of football games to the diagnosis of liver disease and when you can hardly come up with half dozen studies showing even a weak tendency in favor of the clinician, it is time to draw a practical conclusion, whatever theoretical differences may still be disputed. Why, then, is such a strongly and clearly supported empirical generalization not applied in practice, particularly because there are no plausible theoretical reasons to have expected otherwise in the first place?*

With mounting evidence demonstrating the superiority of algorithms over subjective judgment, it would seem logical for people to welcome algorithms for these tasks with open-arms. However, they often *don't*. Some of the anecdotal reasons cited for the reluctance to rely on algorithms include the presumed inability of algorithms to incorporate qualitative data [140], the

notion that algorithms cannot properly consider individual circumstances [140], the notion that algorithms are dehumanizing [140, 143], the inability of algorithms to learn [143], concerns about the ethicality of relying on algorithms to make important decisions [143], the desire for perfection [143, 144], and the presumed ability of humans to improve through experience [144]. In their article, Grove & Meehl [140] challenge common objections from clinicians regarding the use of algorithms in practice and suggest "some of the sociopsychological factors that may help to explain this remarkable resistance to argument and evidence" [140] include: "[f]ear of technological unemployment," "self-concept" (perceptions of self-worth), "attachment to theory" (an idea that theory-mediated predictions do not contribute beyond what an atheoretical algorithm could produces cognitive dissonance), "misperception of the actuarial method as dehumanizing to clients or patients," "general dislike of computers' successfully competing with human minds," and "poor education" [140]. On the topic of "poor education," Grove & Meehl [140] elaborate:

> *Poor education is probably the biggest single factor responsible for resistance to actuarial prediction; it does not involve imputation of any special emotional bias or feeling of personal threat. In the majority of training programs in clinical psychology, and it is surely as bad or worse in psychiatry and social work, no great value is placed upon the cultivation of skeptical, scientific habits of thought; the role models—even in the academy, more so in the clinical settings—are often people who do not put a high value upon scientific thinking, are not themselves engaged in scientific research, and take it for granted that clinical experience is sufficient to prove whatever they want to believe.*

Grove & Meehl [140] ultimately conclude their discussion with the following appeal to policymakers:

> *[P]olicy makers should not accept a practitioner's unsupported allegation that something works when the only warrant for this claim is purported clinical experience. Clinical experience is an invaluable source of ideas. It is also the only way that a practitioner can acquire certain behavioral skills ... [but] ... [i]t is not an adequate method for settling disputes between practitioners, because they each appeal to their own clinical experience. ... All policy makers should know that a practitioner who claims not to need any statistical or experimental studies but relies solely on clinical experience as adequate justification, by that very claim is shown to be a nonscientifically minded person whose professional judgments are not to be trusted. ... To use the less efficient of two prediction procedures in dealing with such matters is not only unscientific and irrational, it is unethical. To say that the clinical-statistical issue is of little importance is preposterous.*

What is interesting is the reactions of clinicians even *after* the evidence had mounted and was generally incontrovertible. This phenomenon is demonstrated by the Evidence Based Medicine (EBM) movement initially introduced in the 1990s.

On November 4, 1992, the Evidence-Based Working Group, chaired by Gordon Guyatt, published a consensus article describing the newly coined term *evidence-based medicine* as "a new paradigm for medical practice" [145]. The working group explained that "[e]vidence-based medicine de-emphasizes intuition, unsystematic clinical experience, and pathophysiological rationale as sufficient grounds for clinical decision making" [145]. The core concept of EBM was

to transition clinical decisions from relying solely on a subjective, experience-based foundation to a more objective, evidence-based foundation by emphasizing the integration of research evidence and recognition of uncertainty into the scheme of clinical decision making [145]. Over the next two decades, the medical community slowly began to embrace the EBM paradigm as the standard of clinical practice. Today, nearly every medical school teaches EBM principles and numerous textbooks and journals have become devoted specifically to the topic creating in a new generation of physicians for which EBM is the default method for clinical practice. Looking back, it is considered by many "difficult to exaggerate the impact of EBM on the medical world" [146].

Although EBM ultimately became the standard for clinical practice, it was not without resistance and criticism—oddly reminiscent to the reactions from clinicians in years prior. Originally termed "scientific medicine," Dr. Guyatt recounts the unsympathetic responses by many of his colleagues—"Those already hostile were incensed and disturbed at the implication that they had previously been 'unscientific'" [147]. Although the EBM moniker was more palatable, clinicians were openly disparaging and critical—particularly on the role of experienced-based opinions in the realm of the new paradigm [148, 149]. The source of this criticism was often thought to be brought on by those within the profession looking to preserve autonomous, professional jurisdiction of decision making [148]; however, it highlighted the importance of a synergy between evidence-based principles and individual judgment. This was considered particularly important in situations where the context of an individual case is under-represented in the statistical data [150] and therefore insufficient to provide a meaningful contribution to the decision. Although EBM supporters recognize the value and necessity for individual judgment, it remained a point of debate as to how experience should be integrated with EBM evidence [150].

The EBM movement provides an important, and more recent, case-study illustrating the challenges with introducing algorithms into a domain that has traditionally been driven by human judgment. By the time EBM was initially introduced in the mid-1990s, the debate concerning the general superiority of algorithms was well established and the evidence was heavily favoring algorithms and statistical methods as superior to intuition and experience across a number of different domains. However, clinical practitioners continued to be reluctant to embrace it. Even more, the reasons cited for the resistance were not novel or specific to a unique aspect of EBM. On the one hand, these reactions can seem irrational, unscientific, and unethical from the perspective of proponents to the new paradigm (e.g., consider the responses by Grove & Meehl [140] in the discussion above). On the other hand, these reactions suggest the issue is more complex. As irrational as it might seem (to choose an inferior method between two options), there is clearly more to consider in order to understand *why* people tend to be averse to algorithms and *how* it can be overcome. In the section that follows, we take a step back and explore the dynamics of human-algorithm interactions more generally from perspectives of psychology and behavioral sciences to consider the tendency for people to rely on algorithms and factors that are believed to increase or decrease those tendencies.

7.4 Part II: Human-Algorithm Interaction in Laboratory Studies

As concepts and applications in artificial intelligence and machine learning bolstered the prominence of algorithms in the 1990s, it became clear that in order for the superior capabilities

of algorithms to be realized, people had to be willing to rely on them. This phenomenon in which people tend to remain resistant to using algorithms and, when given the choice, often opt to rely on predictions made by a human compared to a superior algorithm—ultimately dubbed *algorithm aversion*—became an important focal point of research in psychology and human behavior. Rather than continually demonstrating the superiority of algorithms, the emphasis shifted to understand *why* people tend to be averse to algorithms and factors that may increase or decrease that tendency. In 2014, Dietvorst et al. note that prior literature on this topic has been limited to anecdotal experiences given the dearth of empirical evidence [151]. In an influential study involving incentivized forecasting tasks, Dietvorst et al. find that algorithm aversion, in part, hinges on people's experience with the algorithm [151]—specifically:

> *[S]eeing algorithms err makes people less confident in them and less likely to choose them over an inferior human forecaster. This effect was evident in two distinct domains of judgment, including one in which the human forecasters produced nearly twice as much error as the algorithm. It arose regardless of whether the participant was choosing between the algorithm and her own forecasts or between the algorithm and the forecasts of a different participant. And it even arose among the (vast majority of) participants who saw the algorithm outperform the human forecaster.*

Dietvorst et al. note that the resistance to algorithms is, at least partially, due to a greater intolerance for error from algorithms than from humans and that "people are more likely to abandon an algorithm than a human judge for making the same mistake" [151]. Although these studies do not provide clear insights into how algorithm aversion can be overcome, the findings suggest proposed solutions will need to consider how to counter the apparent tolerance imbalance.

Following the work by Dietvorst [151], Logg et al. suggests the concept of algorithm aversion is not as straightforward as prior literature suggests and highlights boundary conditions for empirical evidence supporting algorithm aversion [152]. They argue that people are not necessarily universally averse to algorithms (particularly, prior to receiving any performance accuracy feedback); rather, they suggest that there is more to unpack on this topic and reliance on algorithms is likely to depend on several factors, such as *who* is relying on *which* advice and for *what* purpose [152]. For basic prediction tasks by lay people, such as non-incentivized numerical predictions based on visual stimuli (e.g., predicting the weight of an individual from a photograph) and forecasts about the popularity of songs and romantic attraction, Logg et al. show that "lay people adhere *more* to advice when they think it comes from an algorithm than from a person"—an effect coined as *algorithm appreciation* [152]. They note, however, that algorithm appreciation decreased when people chose between an algorithm's estimate and their own (versus a different human judge) and when the people had expertise in the domain [152]. These findings suggest observations related to algorithm aversion may have also been bolstered by people's excessive appreciation of their *own* opinions—a phenomenon well established in the literature and referred to as "overconfidence bias"—which has demonstrated that individuals treat their judgment as superior to that of others [152]. Further, and perhaps more concerning, they find that individuals bearing domain expertise were *least* likely to recognize the value of algorithmic advice [152]. Specifically, Logg et al. state: "Paradoxically, experienced professionals, who make forecasts on a regular basis, relied less on algorithmic advice than lay people did, which hurt their accuracy. ... These results might help explain why pilots, doctors, and other experts are resistant to algorithmic

advice. Although providing advice from algorithms may increase adherence to advice for non-experts, it seems that algorithmic advice falls on deaf expert ears, with a cost to their accuracy" [152].

The findings by Dietvorst et al. [151] and Logg et al. [152] are significant in that they demonstrate that challenges persist in present day related to people's willingness to rely on algorithms in certain conditions, despite general familiarity and presence of algorithms across nearly every industry. Although Logg et al. did not explicitly note the findings from Arkes et al. [153] regarding the impact of expertise three-decades prior, their observations were remarkably consistent and equally alarming. In their paper Arkes et al. [153] expanded on the popular *clinical vs. statistical* debate in the 1980s and provided one of the first empirical studies regarding specific conditions impacting people's willingness to rely on algorithms vs. human judgment for prediction tasks. In a series of experiments, Arkes et al. evaluated willingness to rely on an algorithm (i.e., a simple classification rule) versus human judgment when three different conditions were manipulated: incentivization, instructional warning, and expertise [153]. Arkes et al. made three important observations: (1) "incentive for high performance resulted in less use of the decision rule whether the incentive was given for each correct judgment or for the best performance among a group of judges ... [which] actually resulted in poorer performance," (2) "warning subjects of the counterproductive results of abandoning the rule caused the subjects to use the rule more," and (3) "those with expertise (or those who judged themselves to have expertise) were less likely to use a decision rule than those with less expertise ... [and] [b]y choosing not to use the rule, such 'experts' performed worse but had higher confidence in their performance than the nonexperts" [153]. They note (as did Logg et al. [152]) the likely impact of overconfidence bias as a cause for experts being less willing to rely on the algorithm decision aid. Specifically, Arkes comments: "One of the dangers of overconfidence is that one feels that no assistance is needed. If one assumes that his or her judgment is quite good, decision aids would be entirely superfluous. Indeed, in [our experiment] the more knowledgeable subjects were less likely to use the rule, which resulted in inferior performance" [153]. Arkes et al. conclude with an important implication of these observations: "Note that in both the psychological and medical diagnosis scenarios described above, there exist well-meaning diagnosticians with high motivation, high expertise, and few constraints on innovative tendencies [e.g., lack of discipline to adhere to a decision rule]. These are the conditions under which decision aids are less likely to be used—to the detriment of those being served" [153].

In 2016, as a follow-up study to their initial observations related to algorithm aversion (the phenomenon that people often fail to use evidence-based algorithms *after* learning that they are imperfect), Dietvorst et al. investigate strategies to reduce algorithm aversion by allowing people to exert some influence over the algorithm output [154]. Dietvorst et al. hypothesize that "[i]f people's distaste for imperfect algorithms is in part driven by an intolerance of inevitable error, then people may be more open to using imperfect algorithms if they are given the opportunity to eliminate or reduce such errors. Thus, people may be more willing to use an imperfect algorithm if they are given the ability to intervene when they suspect that the algorithm has it wrong" [154]. In evaluating this, Dietvorst et al. recognize that "[a]lthough people's attempts to adjust algorithmic forecasts often make them worse, the benefits associated with getting people to use the algorithm may outweigh the costs associated with degrading the algorithm's performance" [154]. In a series of incentivized (non-expert) forecasting tasks in which participants could choose

between using their own judgments or forecasts of an algorithm built by experts, Dietvorst et al. observe evidence supporting their hypothesis: "Participants were considerably more likely to choose to use an imperfect algorithm when they could modify its forecasts [even after seeing it err] ... [and] ... the preference for modifiable algorithms held even when participants were severely restricted in the modifications they could make" [154]. Further, Dietvorst et al. note that participants who were able to modify the imperfect algorithm's forecasts "reported higher satisfaction with their forecasting process and thought that the algorithm performed better relative to themselves compared with participants who could not modify the algorithm's forecasts" [154]. In closing, Dietvorst note that participants' intervention "did often worsen the algorithm's forecasts when given the ability to adjust them. However, we may have to accept this error so that, overall, people make less error" [154].

The discussion above provides, in our view, important insights into the complex relationship between human-algorithm interactions. Despite the abundance of evidence in these disciplines demonstrating algorithms generally outperform humans, people tend to discount them in favor of their own judgments—even when their own judgments are known to be inferior—often resulting in lower accuracy than relying on the algorithm alone [140, 141]. This phenomenon is most pronounced when the individuals have a high motivation to be accurate and possesses domain expertise in the prediction task (e.g., physicians performing medical diagnoses) [152, 153]. Some researchers point to several sociopsychological factors based on anecdotal observations as potential causes [140]; others suggest it is a manifestation of overconfidence bias [152, 153]. More recently, researchers found that this phenomenon is exacerbated when people are presented with the performance of the algorithm, and thus, the inevitable susceptibility to err, which often results in worse performance and impacts to business and society can be costly [151]. In an effort to explore solutions to mediate the impacts of algorithm aversion and increase the likelihood people are willing to rely on algorithmic advice, Dietvorst et al. find that allowing people to intervene and modify the algorithm's output, even under limited conditions, tend to result in higher satisfaction, greater belief in the superiority of the algorithm, and higher likelihood to commit to using algorithms in subsequent tasks [154]. In a general sense, these observations illustrate an interesting paradox: to reduce error, we may need to accept error[30]. Allowing for even just the *potential* for such intervention, despite constrained circumstances, appears to cater to people's desires to incorporate their own judgments and *feel* they have some control over the outcome, thus resulting in a higher likelihood for adoption. Consequently, people are likely to be more satisfied with the process and performance will likely increase overall compared to the alternative in which people are more prone to reject the algorithm altogether in favor of their own judgments. Dietvorst et al. [154] note the operational implications of these findings and suggest:

> *[F]raming the decision of whether or not to use an algorithm as an all-or-nothing decision is likely to be counterproductive. People are unlikely to commit to using an algorithm's forecasts exclusively after getting performance feedback or learning that it is imperfect. Furthermore, forcing employees into a regime in which they have to use an imperfect algorithm's forecasts exclusively may lead them to become dissatisfied or push for a change. However, asking people to commit to an algorithm's forecasts that they can modify by a limited amount seems much more palatable. People will be much more likely*

[30] In other words, to achieve greater performance overall through algorithms, we may need to tolerate reduced performance through human intervention in order to increase the tendency for people to rely on them.

> *to choose to use an imperfect algorithm if they can modify its forecasts, and employees will not necessarily be dissatisfied if they are partially constrained to an imperfect algorithm's forecast. ... If for some reason having employees making constrained adjustments to an algorithm's forecasts is not possible, [our study] shows that having employees make unconstrained adjustments to an algorithm's forecasts can also substantially improve their forecasting performance.*

These findings and recommendations have implications across a broad array of domains which are faced with increasing human-algorithm interactions, particularly in those domains traditionally dominated by human judgment (based on expertise and experience) and for which there is high motivation to be accurate. The empirical evidence is telling; however, what is more interesting is to reflect on human-algorithm interactions in distinct yet relevant real-world circumstances involving these conditions in context of these findings. For this purpose, we look at how things have unfolded in medicine as well as how things are presently unfolding in autonomous driving—a domain that *everyone* can relate to. These examples allow us to explore how people and society have grappled with the dynamics of human-algorithm interactions in these contexts and what seems to have ultimately proven to be successful (or at least palatable) over time. Indeed, as we discuss in the next section, it seems that irrespective of the specific domain, allowing human-algorithm *integration* is associated with an increase in people's initial willingness to consider algorithms enabling them to ultimately grow more trusting, reliant, and accepting of the algorithms to influence the decision outcome.

7.5 Part III: Human-Algorithm Interaction in Real World Domains

In medicine, we have had the luxury to look back onto two (somewhat overlapping) eras and see how the issues have played out over time. As the evidence mounted demonstrating the superiority of algorithms, the idea that it was a dichotomous choice of one or the other was quickly met with resistance and criticism. Simply put, for various cited reasons (e.g., see [140, 143, 144]), clinicians were not willing to yield their clinical decision-making responsibilities to an algorithm—even though algorithms have been shown to provide superior performance. Over time, however, as the debates ensued, clinicians began to trend toward an integrated approach to bring algorithms into their scheme of clinical decision making (as opposed to wholesale outsource which often resulted in wholesale rejection) and the concept of "clinical *and* statistical" was introduced and slowly emerged as a workable solution [155]. Around that same time, when EBM was initially introduced in 1992, the authors of the Working Group stressed the importance of integrating research evidence into the decision-making scheme. At the outset, EBM was not presented as a dichotomous choice, but many clinicians clinging to the value of clinical judgment still reacted with criticism and outright rejection as if it were (see EBM 25-27]). Once again, the importance of judgment and experience in the scheme of clinical decision making was highlighted by those concerned it would go to the wayside. Although viewpoints were polarized, as the initial reactions subsided, clinicians naturally trended toward a reasonable middle ground and clinical decision making became an integrated and multi-faceted approach of "clinical *and* statistical." As Coen et al. highlighted: "Perhaps EBM should be renamed 'methods of incorporating epidemiologic evidence into clinical practice' ... but this is quite a cumbersome moniker" [150].

Looking retrospectively, we see that within the domain of medicine, what started out as a "clinical *versus* statistical" debate naturally transitioned into a "clinical *and* statistical" integrated solution. Interestingly, one may argue that the present-day research related to human-algorithm interactions offers a reasonable explanation. Indeed, physicians possess high expertise, are highly incentivized and motivated to provide accurate decisions, and operate fairly autonomously. As noted previously: these are the conditions under which people are most likely to rely on their subjective judgment and least likely to accept algorithms [152, 153]. The solution proposed by Dietvorst et al. [154] appears to have naturally taken shape. By structuring the scheme of clinical decision making as an integrated approach based on statistical evidence and subjective judgment, the clinician maintains the ability to exert some influence on the overall outcome—whether that is by adjusting for idiosyncratic factors that are shown to be under-represented by the statistical evidence or by relying on the statistical evidence as an additional pillar to support the overall foundation of the decision. While this approach seems to provide the conditions that are most appealing for practitioners in terms of their willingness to adopt, there is concern that clinicians are too quick to find "exceptions" in the statistical data and adjust in favor of their subjective judgment [156]. How those "exceptions" can and should be moderated remains an open question.

In autonomous driving, the issues of human-algorithm interaction can be viewed through the public's willingness to embrace automation to either supplement or supplant their own driving tasks—tasks for which drivers generally consider themselves to have expertise based on specialized knowledge and experience and for which there are high stakes and serious safety concerns for inaccurate decisions. For background, in 2014, SAE International first published the standard J3016 Levels of Driving Automation which defined six levels of vehicle automation ranging from Level 0 (no automation) to Level 5 (full automation), transitioning gradually from "driver support features" to "automated driving features" [157]. The SAE J3016 Levels of Driving Automation taxonomy was subsequently adopted by the United States National Highway Traffic Safety Administration (NHTSA) in 2016 as a formal taxonomy for describing increasing levels of automation and shifting roles from the human to machine for executing dynamic driving tasks. For context, the levels of the J3016 standard are: Level 0 (No Automation), Level 1 (Driver Assistance), Level 2 (Partial Automation), Level 3 (Conditional Automation), Level 4 (High Automation), Level 5 (Full Automation) [157]. In levels 0 through 2, the human maintains full control with increasing assistance from technology and in levels 3 through 5, the system is in control with decreasing need for human intervention [157]. The levels codified by the J3016 standard provide a useful framework for considering people's willingness to engage in progressive levels of vehicle automation and shifting responsibility and control from human to machine for various driving tasks.

First, from the perspective of safety it is important to consider the necessity of moving toward automation in consumer vehicles. For example, in 2016, the United States Department of Transportation released the NHTSA fatal traffic crash data on American roadways and the results were startling: human choices were linked to 94 percent of serious crashes resulting in a call to "promote vehicle technologies [to] ... help reduce or eliminate human error and mistakes that drivers make behind the wheel" [158]. These results alone demonstrate that the case for vehicle automation is clear and people should embrace automation with open arms—after all, it will improve safety and save lives. However, once again, they often *don't*. In 2016 and 2017, surveys were conducted regarding consumer interest in automation and the highest level of automation in

vehicles they would be willing to consider. In 2016, Shoettle and Sivak analyzed 618 survey responses from participants throughout the United States and found that 45.8% of respondents preferred no self-driving capabilities, 38.7% preferred partially self-driving capabilities, and only 15.5% preferred full self-driving capabilities [159]. When asked about the preferences for controlling completely self-driving vehicles, Shoettle and Sivak found that "[n]early all respondents (94.5%) would want to have a steering wheel plus gas and brake pedals (or some other controls) available in completely self-driving vehicles [159]. In 2017, Abraham et al. released a follow-up to a survey completed in 2016 to explore changes in perception from one year prior [160]. This was due in large part because shortly after the initial data were collected from the 2016 survey, the world saw the first fatality related to a highly automated driving feature [161]. In their follow-up survey, Abraham et al. analyzed 2,976 survey responses from participants throughout the United States and found "a significant decrease in the proportion of respondents who were comfortable with the idea of a fully self-driving car and an apparent shift toward more limited automation in the form of 'features that actively help the driver while the driver remains in control.' Similarly, there was a proportional decrease in those who were comfortable with features that periodically take control of driving" [160]. Further, Abraham et al. found that among participants who reported they would never purchase a self-driving car, "[t]he most cited hesitation was discomfort with the loss of control; other commonly mentioned factors included not trusting the technology, a disbelief that it would be robust enough to rely on exclusively, and a feeling that self-driving cars are unsafe" [160]. Abraham et al. conclude with the following [160]:

> *The perception that self-driving cars need to work perfectly to be acceptable, combined with present and past experiences of low-risk technology failure both in and out of vehicles, may lead many consumers to believe the technology will never be good enough such that they can trust it with their lives. The difficulty here is that it remains an open question as to how safe a self-driving vehicle needs to be in order to become socially acceptable as a mobility option. ... Encouraging the appropriate use of driver assistance and other human-centric automated vehicle systems by investing in educational resources that consumers prefer may be an important stepping stone to improving consumer interest, confidence, and trust in self-driving technology.*

Around this same time, in 2017 the RAND Corporation released their report "The Enemy of the Good: Estimating the Cost of Waiting for Nearly Perfect Automated Vehicles," which explored this very issue [162]. The RAND Corporation open their report with the current quandary:

> *[A] key question for the transportation industry, policymakers, and the public is how safe [highly automated vehicles (HAVs)] should be before they are allowed on the road for consumer use. From a utilitarian standpoint, it seems sensible that HAVs should be allowed on U.S. roads once they are judged safer than the average human driver so that the number of lives lost to road fatalities can begin to be reduced as soon as possible. Yet, under such a policy, HAVs would still cause many crashes, injuries, and fatalities—albeit fewer than their human counterparts. This may not be acceptable to society, and some argue that the technology should be significantly safer or even nearly perfect before HAVs are allowed on the road. Yet waiting for HAVs that are many times safer than human drivers misses opportunities to save lives. It is the very definition of allowing perfect to be*

> *the enemy of good. ... The lack of consensus on how safe HAVs should be before they are allowed on the road for consumer use reflects different values and beliefs when it comes to humans versus machines.*

To explore this issue further, the RAND Corporation conducted a series of analyses comparing road fatalities over several decades under different theoretical policies in which HAVs are deployed when they are just 10% better than the average human driver (*Improve10*) or wait until they are 75% better (*Improve 75*) or 90% better (*Improve90*) than the average human driver. From these analyses, the RAND Corporation found [162]:

> *In the short term (within 15 years), more lives are cumulatively saved under a more permissive policy (Improve10) than stricter policies requiring greater safety advancements (Improve 75 or Improve90) in nearly all conditions, and those savings can be significant—hundreds of thousands of lives. The savings are largest when HAVs under Improve10 are adopted quickly. ... In the long term (within 30 years), more lives are cumulative saved under an Improve10 policy than either Improve75 or Improve90 policies under all combinations of conditions we explored. Those savings can be even larger—in many cases, more than half a million lives.*

These data demonstrate the value of moving toward vehicle automation sooner rather than later. Despite these findings, people remain in disbelief and reluctant to accept the technology. In their review of the literature related to consumer acceptance of automated vehicle technology between 2013 and 2019, Jing et al. note that despite the rapid development of the technology, public acceptance of automated vehicles is one of the major factors affecting widespread distribution. Specifically, the term "safety" was the most frequently occurring word in all of the collected literature: "Some respondents even estimate autonomous driving is not as safe as human driving. Hence, they are more willing to accept [automated vehicles (AVs)] with manual driving options than fully AVs without steering wheels. The deaths of AV accidents reported in recent years may intensify public suspicion about the safety issues, and safety concerns have proven to be a potential deterrent to the acceptance of AVs" [163].

The issues concerning autonomous driving once again illustrate the pervasive impact of people's reluctance to accept imperfect algorithms and disjointed expectations that algorithms need to be nearly perfect before accepting them. Ironically, despite the evidence that automation in vehicles is *more*-safe, the most commonly cited reason for peoples' hesitation to adopt is due to concerns that they are *less*-safe. The barriers to improved safety and performance are, once again, rooted in peoples' reluctance to rely on the algorithms—particularly after news of an accident where the technology was involved. Despite the apparent aversion, the evidence shows that people will be more willing to accept autonomous vehicles *if* they still have the option to maintain control and can rely on their own judgment and decisions [159]. Within the domain of vehicle automation we see again that the findings from Dietvorst et al. appear to generalize well—people are averse to relying on algorithms after seeing them fail (even though the algorithms' overall performance is better than human judgment alone) and people tend to hold algorithms to higher standards than their human counterparts, demanding near perfect performance before personally embracing them [151]. Further, we see that allowing the human to have some control over the outcome of the

driving task tends to increase their willingness to work *with* the algorithm—a possible solution to reduce the effects of algorithm aversion proposed by Dietvorst et al. [151].

Human-algorithm interactions in both medicine and autonomous vehicles are not too different from one another, despite the apparent orthogonal relation of the two domains. The issue ultimately boils down to trust and confidence with the algorithms. People naturally trust themselves and their *own* judgment (however flawed it may be) over other sources, particularly when they are the ones ultimately responsible for the outcome or have some inherent incentive to be accurate. By introducing increasing levels of automation designed to *supplement* the human as opposed to immediately *supplant* the human, people tend to be more willing to incrementally accept the increased intervention of automation and slowly become more comfortable and trusting in the technology. As comfort and trust in the technology evolves, reliance on the algorithms will increase resulting in improved performance and safety over time.

7.6 Part IV: Algorithms and the American Legal System

As algorithms have advanced and automated decision systems have become more accessible, researchers, advocates, and policymakers are debating when and where these systems are appropriate—including particularly sensitive domains such as criminal justice [124]. Questions have been raised on how to fully assess the short and long-term impacts of these systems and the appropriateness of their applications given many operate as "black-boxes" [124]. In an effort to keep pace with these types of issues, the first G7 Multi-stakeholder Conference on Artificial Intelligence was held in Montreal, Canada in December 2018 with the overarching theme of "Enabling the Responsible Adoption of AI" [164]. Over 200 experts in artificial intelligence (AI) attended the conference, representing all of the G7 countries and beyond, as well as key multi-stakeholder perspectives from industry, academia, civil society, and government. Among those in attendance was Geoff Hinton, world-renowned computer scientist, industry leader in AI, and developer of the "Google Brain." During an interview at the conference, when prompted about AI's eventual role in decision making, Hinton responded [165]:

> *I'm an expert on trying to get the technology to work, not an expert on social policy. One place where I do have technical expertise that's relevant is [whether] regulators should insist that you can explain how your AI system works. I think that would be a complete disaster.*
>
> *People can't explain how they work, for most of the things they do. When you hire somebody, the decision is based on all sorts of things you can quantify, and then all sorts of gut feelings. People have no idea how they do that. If you ask them to explain their decision, you are forcing them to make up a story.*
>
> *Neural nets have a similar problem. When you train a neural net, it will learn a billion numbers that represent the knowledge it has extracted from the training data. If you put in an image, out comes the right decision, say, whether this was a pedestrian or not. But if you ask "Why did it think that?" well if there were any simple rules for deciding whether an image contains a pedestrian or not, it would have been a solved problem ages ago.*

In a follow-up question, when asked about how should people trust these algorithms, Hinton responded [165]:

> *You should regulate them based on how they perform. You run the experiments to see if the thing's biased, or if it is likely to kill fewer people than a person. With self-driving cars, I think people kind of accept that now. That even if you don't quite know how a self-driving car does it all, if it has a lot fewer accidents than a person-driven car then it's a good thing. I think we're going to have to do it like you would for people: You just see how they perform, and if they repeatedly run into difficulties then you say they're not so good.*

Hinton's remarks during this interview were immediately met with criticism, challenging the notion that scientists working to develop algorithms can separate themselves from downstream implications resulting from algorithm applications. For example, Dr. Heather Roff from the University of Cambridge responded [166]:

> *This is a dangerous position to take. An expert on technology who feels themselves divorced from social or policy implications does not understand that technology is not value neutral, and that their decisions—even seemingly basic ones on how many gradient descents to take in a system—have socio-political implications. If one thinks they are only Scientists doing Science, but then simultaneously think that regulators should take an interest has fundamentally misunderstood their role as scientists engaging in socially and morally important questions. If your work requires legislation then you should think about that at the design stage … period.*

As illustrated by the exchange above, it would be a naïve viewpoint to consider the issue of algorithm implementation in a specific domain complete without consideration of the environment for which the algorithms are ultimately applied and the implications of such applications. Within the broader criminal justice context, law enforcement leaders are strategizing how to leverage the benefits that algorithms provide within various aspects of policing and criminal justice, but in doing so have stressed the importance of maintaining public trust and upholding societal values by ensuring algorithms are characterized by fairness, accountability, transparency, and explainability [167-169]. In the case of forensic science, an important consumer of the forensic results is the legal system, which bears the ultimate responsibility for ensuring all people receive fair and equitable justice under the law. Although algorithms have demonstrated remarkable potential to provide advanced scientific capabilities and promote objective foundations to the ultimate issues in question, they do so often at the cost of transparency and explainability [114-120]. In some cases, algorithms may operate as a black-box due to trade secrets or other legal protections asserted by the manufacturer. In others, they may manifest as a black-box due to their computational complexity. In either situation, legal actors have expressed concern that the opacity of algorithms can stifle meaningful scrutiny and accountability of the systems thereby infringing on criminal defendants' Constitutional rights (e.g., see [114, 115, 117, 118]). These issues are exacerbated by examples in which algorithms have indeed perpetuated historic inequities (e.g., see [119, 124, 170, 171]). Faced with these concerns, courts have found themselves arbitrating complicated legal questions forcing them to grapple with issues concerning the admissibility of algorithms and their implications to the law. Legal scholars have begun to

explore these issues in various contexts within the American legal system—most notable and relevant to our discussion are those by Imwinkelried [114] and Nutter [118], which are briefly summarized below. Our intent here is to be illustrative, not exhaustive, of the importance to consider broad downstream legal implications of algorithms when deciding when and how to apply them to a particular (and sensitive) domain, such as forensic science for criminal justice purposes. Specific technologies and circumstances concerning their applications within the criminal justice pipeline may create additional implications and it would be impractical to cover them all in this discussion. Our discussion is intentionally generic in terms of the specific legal issues and narrowly focused on the application of algorithms developed for purposes of augmenting traditional forensic science methods which rely predominantly on human judgment and expertise. Although we borrow examples from probabilistic genotyping for illustrative purposes, our focus is directed toward pattern and impression evidence disciplines and is not meant to apply to all types and applications of algorithms that have been, or could be, introduced into litigation. Finally, we do not consider the issue to be *whether* algorithms should be implemented into forensic practice for criminal justice purposes—we consider the issue to be *how* to implement them in a way that is cognizant of the legal issues and increases the likelihood legal stakeholders will be willing to consider them within their own regulatory framework.

The legal issues concerning the application of algorithms to pattern and impression evidence has yet to be fully explored. Only recently legal scholars begun to unpack the issues and consider how the legal system can adapt to the inevitable application of algorithms while maintaining their gatekeeping function. In 2016, Imwinkelried considers the issue in the context of the expanded use of algorithms for probabilistic genotyping software introduced in 2009 using TrueAllele software from Cybergenetics, Inc. [114]. Probabilistic genotyping software analyzes DNA mixtures and provides a statistic that helps assess whether or not a particular defendant was one of the contributors. Although TrueAllele is not the only probabilistic genotyping software available, it has received attention in the United States due to its use and the manufacturer's assertion of trade secret protections when criminal defendants have requested its source-code to examine its reliability. Ultimately, courts have largely rejected defendant's requests for disclosure and independent review of source-code and ruled in favor of admissibility, which has led to an outcry by defense litigators (e.g., see [114-117]). After reviewing prior case law admitting testimony based on technologies such as TrueAllele while denying defendants access to the source-code, Imwinkelried presents a critical analysis of the legal issue. In particular, Imwinkelried addresses the question of whether the prosecution should be permitted to introduce expert testimony based on a computerized technique without presenting foundational testimony about the validity of the program's source code controlling the technique and, if so, whether there are circumstances in which the defense ought to have access to the source code despite the trade secret assertions protecting such disclosure [114].

The first issue considered by Imwinkelried is the admissibility of a computerized technique without providing foundational testimony about the validity of the source-code controlling the technique. In other words, whether the evidence produced by the system should be admitted without first revealing the underlying code controlling the operation of the algorithm and the algorithm itself. In the United States, a minority of non-federal jurisdictions still rely on the 1923 *Frye* standard of "general acceptance" [172]. Under this standard, the proponent need not demonstrate foundational validity at all; rather, as Imwinkelried describes, the proponent only

needs to demonstrate whether the "theory or technique has gained a certain degree of popularity—'general acceptance'—within the relevant scientific fields" [119]. In 1975, however, the Federal Rules of Evidence took effect, which led to the 1993 ruling in *Daubert v. Merrell Dow Pharmaceuticals, Inc.* [173], the first of a trilogy of Supreme Court decisions on the admissibility of expert testimony (*Daubert v. Merrell Dow Pharmaceuticals, Inc.* [1993] [173], General *Electric Co. v. Joiner* [1997] [174], and *Kumho Tire Co. v. Carmichael* [1999] [175]—collectively referred to as the "*Daubert* standard"). Under the *Daubert* standard, the proponent must demonstrate that the theory or technique rests on adequate validation for which trial judges bear that gatekeeping responsibility. Today, federal and the majority of non-federal jurisdictions rely on the *Daubert* standard as a framework for admissibility. It is under this standard that many criminal defendant's assert that a computerized technique, without access to the underlying source-code and algorithm itself, should be inadmissible due to the inability to demonstrate its validity. However, as Imwinkelried describes, courts have ruled that the burden of demonstrating the validity of a technique can be met "by presenting testimony about the validation studies investigating the accuracy of the software, ... [t]he very purpose of a validation study is to investigate whether the theory or technique does what its proponent claims" [114]. Imwinkelried argues (in the context of TrueAllele) [114]:

> *Federal Rule of Evidence 901(b)(9) [176] captures the essence of the "authentication" or validation of a scientific technique. In the words of 901(b)(9), the essential foundation is a "showing that [the process or system] produces an accurate result" [citing [176]]. Validation studies summarizing the results of tests of the technique and showing that the technique yields accurate results satisfy that standard. As a matter of logic, the court should treat the studies as adequate validation under Daubert. The proponent can shoulder the burden of Daubert without making a further, separate showing about the source code of the software controlling TrueAllele. The lack of testimony about the source code might increase the degree of uncertainty in the expert's final opinion, but post-Daubert, the expert need not vouch for his or her opinion as a certainty. In short, many courts have reached the correct result that prosecution evidence based on TrueAllele can be admitted, even without testimony about the source code.*

Ultimately, Imwinkelried argues that, in the context of probabilistic genotyping software in particular, courts have rendered appropriate decisions from a legal standpoint as to the admissibility of algorithms without the requirement that the source-code (and algorithm thereto) be released [114].

Admissibility, however, is only one of the legal issues to consider. As Imwinkelried notes, "[e]ven when the proponent's item of evidence is admissible, the opponent has the right to attack the weight or believability of the evidence. ... The U.S. Supreme Court has held that in criminal cases, the defendant's right to attack the weight of the prosecution's evidence is of constitutional dimension under the Sixth Amendment Confrontation Clause" [114]. Indeed, when handing down the ruling, the *Daubert* court noted: "Vigorous cross-examination, presentation of contrary evidence, and careful instruction on the burden of proof are the traditional and appropriate means of attacking shaky but admissible evidence" [173]. This leads to the second issue—in the absence of the underlying algorithm and source-code, are defendant's deprived of the ability to challenge the credibility of the evidence? This question is more complicated. Although Imwinkelried

recognizes that prior courts have pointed to the existence of validation studies as the means to enable opponents to evaluate the accuracy of the system, Imwinkelried also notes: "The answer does not turn on the mere existence of validation studies or even their availability to the defense. Rather, the answer depends on the number of studies, their quality, and a comparison between the test conditions and the conditions in the instant case" [114]. Consequently, if these factors are not well established, it might seem that the source code is warranted; however, such a decision has to be considered in light of the countervailing argument of the proponent's assertion of trade secret protections and "[f]aced with competing legitimate interests, a trial judge must attempt to strike a rational balance" [114]. Ultimately, Imwinkelried proposes a judge could accomplish this by proceeding in two steps [114]:

> *First, a judge should assign to the accused seeking discovery the burden of showing that the facts of the instant prosecution exceed, or are at the margins of, the validation range of the empirical studies relied on by the prosecution. More specifically, the defendant must convince the judge that the available studies do not adequately address the effect of a specified, material variable or condition present in the instant case. The most clear-cut case would be a fact situation in which none of the available studies relied on by the prosecution experts tested the application of the technique to fact situations involving the condition. ... [However,] [t]he judge should certainly not accept the ipse dixit assertion of the defense counsel that the omitted condition is material in the sense that its presence could affect the outcome of the test. ... Rather, the judge ought to demand that the defense present expert testimony explaining why it is plausible that that condition could change the test result.*
>
> *Assume that in the first step, the judge concludes that the defense has met its burden. Even then the judge should not automatically require the manufacturer to furnish the defense with a printout or electronic version of the source code. Instead, the judge could give the manufacturer a choice to: either (1) allow the defense to test the application of the program to a fact situation including the material condition or variable omitted from the validation studies, or (2) provide the defense with the source code. ... By enabling the defense expert to conduct a new validation study testing that application, the manufacturer would afford the defense expert a fair opportunity to investigate the merit of the criticism ... [and ultimately] determine whether the inclusion of the additional condition could actually—not merely theoretically—affect the outcome of the use of the automated forensic technique.*

Imwinkelried concludes: "Until courts guarantee the defense [the ability to challenge trustworthiness], source code will continue to be a source of controversy and doubt about the marked trend toward the automation of forensic analysis in the United States" [114].

In 2019, Nutter considers the issues with an added layer of complication when the evidence is the product of machine learning algorithms [118]. The distinction is that the source code for machine learning algorithms is practically uninterpretable, even for the manufacturer. Thus, contrary to the prior discussion, disclosure of source code in this context would not materially advance the a party's interests of ensuring the reliability of the algorithm. This discussion is relevant to illustrate the legal issues when the algorithm is truly a "black box" and a legal order for source code disclosure is not a practical solution. In the discussion below, although Nutter

considers the issues in the context of general "machine learning algorithms," the issues are analogously applicable to any algorithm that operates or manifests as a black-box (due to legal protections outside of the Court's realm of control or due to computational complexities); thus, for purposes of this discussion, references to "machine learning algorithms" are considered synonymous with "black-box algorithms" in general.

Nutter addresses the issues of machine learning evidence (e.g., black-box algorithms) in criminal prosecution from a prospective standpoint, recognizing that it is only a matter of time until courts will be required to grapple with these issues. In doing so, Nutter "aims to look ahead to possible evidentiary issues when, not if, the output of machine learning algorithms is used as substantive evidence in criminal prosecution" [118]. This context is particularly important as it could enable proponents of forensic algorithms to consider these issues *a priori* when the algorithms are developed and ultimately implemented into forensic practice in a way that recognizes the concerns from legal stakeholders and promotes judicial efficacy.

Like Imwinkelried [114], Nutter first explores the legal issues concerning evidence generated from machine learning algorithms from the perspectives of admissibility [118]. Both Nutter [118] and Imwinkelried [114] share similar perspectives on the issue of admissibility under Evidence Rule 702 [177] and *Daubert* [173]—although there is nothing inherently inadmissible, proponents of the algorithm will need to ensure the validation of the system is applicable to the circumstances of the existing case. However, Nutter takes the discussion a step further and also considers the implications of algorithms under the Constitution. Nutter notes that "[s]everal constitutional provisions may be implicated by machine learning identification in criminal prosecutions. Defendants may cite the Fifth Amendment's Due Process Clause [178] or the Sixth Amendment's Confrontation Clause [179]" under the premise of "guilt by black-box" [118]. Under such an argument, Nutter suggests defendants might claim the lack of transparency and explainability of how the algorithm arrived at the particular conclusion deprives the defendant the ability to challenge its credibility and disclosure of source-code is not the effective remedy [118]. On the basis of a Due Process argument, with reference to analogous past precedent, Nutter ultimately concludes that the Court would most likely "find due process satisfied when (1) the defendant can at least challenge the data that go into the algorithm (a requirement that can be addressed with procedural rules of discovery wholly within the Court's control) and (2) the algorithm possesses some sufficient level of accuracy, which can come to light at a *Daubert* hearing on admissibility or cross-examination at trial" [118]. Consequently, and furthermore, Nutter argues that "it is likely that the Sixth Amendment's Confrontation Clause would require an expert to testify in-person and be subject to cross examination" [118]. Taken together, Nutter ultimately suggests that despite the black-box nature of the algorithm, neither Constitutional provision will categorically bar machine learning evidence; however, the weight of such evidence may be impacted because of the machine learning's distinct unexplainably rendering it difficult, if not impossible, to explain how the algorithm makes a particular conclusion [118]. It is this inherent unexplainably that will present the greatest challenge to proponents of algorithms that operate or manifest as a black-box—despite their admissibility.

Similar to Imwinkelried [114], Nutter recognizes that although there is nothing inherently inadmissible about black-box algorithms, questions remain regarding the weight of such evidence at trial [118]. Accordingly, there will be "considerable onus on trial counsel to persuade the trier

of fact to discount the weight that the evidence should be assigned ... [and] [j]urors might be cautious to assign much weight to machine learning evidence because of its peculiar property that it is often not explainable" [118]. Nutter explains [118]:

> *It is an entirely open question the extent to which, in open court, jurors would trust the validity of unexplainable machine learning evidence. Indeed, this question is ripe for empirical research by psychologists and legal scholars of scientific evidence. Developers understand that the extent to which a person trusts a machine in everyday life is highly variable and context-dependent. Outside the courtroom, an individual's trust in a machine ranges from none or little (for a variety of reasons, one of which is often because it is a machine [referencing [151]]), to passive trust in machines without so much as a second thought. [Further] . . .[r]esearchers find trust in machines to be highly variable and influenced by different factors like belief about the functionality of the technology, belief that the technology is helpful, and belief that the technology is reliable. ... Inside the courtroom, how jurors will respond to machine learning output is very difficult to predict. ... Additionally, inextricably linked to the credibility of the machine is the credibility the jurors extend to the testifying expert him- or herself. That human credibility would likely affect credibility that jurors would extend to the underlying machine, especially as the scientific evidence at issue is particularly complex for laypeople. In that case, the prosecution or defense would surely already be familiar with the usual tactics to use to attack the expert's credibility.*

This last point raised by Nutter [118] brings us to our final point of concern when considering the issue of introducing algorithms into the legal system. From the discussion above, we see that the most significant issues are less about whether the algorithms would be admissible or not—provided they were adequately validated in a way that are representative of the circumstances for the case at hand, then they are likely to be found admissible. Rather, the issues are more so the extent to which fact-finders will be receptive of the evidence generated by the algorithm and afford it the appropriate weight. Thus, in addition to being considerate of issues that might be raised concerning admissibility, proponents of the algorithms will also need to concern themselves with factors that might increase or decrease jurors' and judges' willingness to trust the results of the algorithm (and by extension, human-algorithm combination). This, in turn, causes us to think about two additional issues and their implications to practice: (1) the extent to which the expert will need to be knowledgeable about the underlying algorithm and method employed when faced with such challenges, and (2) *how* the algorithm is implemented at the laboratory and used by the expert.

To expand further on the first point, the implementation of an algorithm will require more than a mere policy change. Such decisions will need to be accompanied by robust training and education to ensure experts are able to be responsive to questions and challenges raised during testimony. Implementation without proper education and training could detract from the overall credibility of the evidence and undermine the benefits it is intended to provide. The depth of that knowledge, however, may depend on how the algorithm is implemented and the extent to which the final conclusion was dependent upon the algorithm. To expand further on the second point, in some situations, such as probabilistic genotyping, the use of the algorithm is necessary to derive information that is otherwise difficult to interpret by the human; thus, the algorithms provide a

capability that was otherwise non-existent. The output of the algorithm is the *sole basis* of the information. Laboratories (and experts) are much more limited in how they use the algorithms in these contexts and fact-finders have little choice but to rely on the algorithm or discount the information altogether. In other situations, such as traditional pattern evidence disciplines, the use of algorithms can be done in parallel with the human to assist with quantifying the value of impressions independent from the value assigned by human judgment alone. The output of the algorithm is a *supplemental basis* of the information. Thus, the algorithms could be applied to impressions that would normally be considered "no value" through subjective interpretation alone (e.g., see [61]) thus providing additional information for the courts to consider, *or* they could be applied to impressions for which experts believe have associative value through their subjective interpretation, but unable to substantiate empirically. It is this latter condition that we are particularly interested in exploring further since it characterizes the most immediate point of concern among scientific and legal scholars calling for algorithms—the need for empirical substantiation so that conclusions do not rely solely on human judgment [3, 7-9].

In circumstances where the algorithm is not a precondition for interpretation, we have flexibility to consider different strategies for how an algorithm could be implemented within the broader examination methodology and the pros and cons of one approach over another—both in context of practitioners' willingness to adopt the algorithm and fact-finders' willingness to rely on evidence generated by the algorithm. In the section that follows, we discuss these issues further and ultimately propose a path forward for the implementation of algorithms into forensic practice that is believed to increase the likelihood for adoption across all stakeholders and lead to an overall stronger foundation and improvement to the quality and consistency of forensic science in general and pattern evidence examination in particular.

7.7 Part V: A Path Forward for Forensic Science

Over the years, several forensic science disciplines have been encouraged to adopt algorithms (i.e., statistical methods). The perceived benefits of algorithms are wide-ranging, but the immediate advantage (particularly for the pattern evidence domains) is to provide an empirical foundation to the evaluation of forensic evidence [3, 7-9]. Although the calls for algorithms in forensic science have prompted researchers to propose numerous potential technology solutions, none have addressed the fundamental questions or strategies of how algorithms should be (or could be) implemented operationally. In the preceding discussions, we have explored the benefits algorithms provide as well as issues of human-algorithm interactions in several different ways. Collectively, these explorations have enabled us to characterize key challenges and consider strategies to reduce the barriers for algorithms to be implemented within forensic science. In this section, we consider *how*, not *if*, algorithms could be implemented into operational practice in such a way that forensic practitioners and other legal and scientific stakeholders are likely to accept. With the context of prior discussions in mind, we first explore different ways that algorithms could be implemented operationally within the examination methodology and implications of those approaches to future practice. Then, we outline a path forward for laboratories to consider as a strategy for implementing algorithms operationally and progressively moving toward ensuring evidence is presented with stronger scientific foundations.

In Parts I through III, we found that the implementation of algorithms into domains traditionally dominated by human judgment is often fraught with resistance [140, 142-144, 148, 149]. People tend to exhibit a general aversion to algorithms and prefer to rely on their own judgment—often despite knowledge that their own judgment is typically inferior to that of algorithms [151]. This phenomenon is exacerbated when people possess domain expertise [152, 153], are faced with high-stakes decisions [152, 153], and are presented with an algorithm that is susceptible to err [151]. Although the actual source of these reactions has not yet been fully understood, some researchers have pointed to various sociopsychological factors [140], overconfidence bias [152, 153], and a general lack of trust in algorithms' abilities to account for idiosyncratic factors [140] as possible explanations for the behaviors. Finally, both anecdotal observations of human-algorithm interactions in different domains and recent research have suggested that people tend to be more receptive to algorithms if they are integrated as a factor that *supplements* as opposed to *supplants* human decision making and the human retains some amount of influence on the ultimate outcome [154, 155]. The above provides important context when considering the implementation of algorithms into forensic science. Indeed, forensic science has the major conditions for which algorithm aversion is most pronounced: (i) forensic examination results (in the pattern evidence domains particularly) are traditionally based entirely on subjective judgment, (ii) forensic examiners possess expertise, and (iii) forensic conclusions involve high-stakes decision-making. Thus, we have no reason to expect the reactions and behaviors of forensic practitioners to be substantially different than what has been observed in research and other domains explored. In fact, to some extent we have already observed similar behaviors manifest. Practitioners' reactions to the mere notion of implementing statistical approaches have been met with criticism and opposition from practitioners [46-48, 51]. Further, when given the opportunity to incorporate algorithms into their decision-making, practitioners tended to disregard them in favor of their own judgments [180]. In an appendix to a discussion regarding the presentation of probabilities in a moot-court exercise related to fingerprint evidence, Langenburg addresses a list of fears that he has commonly heard from practitioners as he has traveled around various jurisdictions providing training [103]. The anecdotal reactions outlined by Langenburg [103] in the context of fingerprint examinations are eerily similar to those addressed by Grove & Meehl [140]. The comparison and recognition of similarities between non-forensic and forensic domains related to reactions to algorithmic interventions and human-algorithm interactions are important because they allow us to understand and be responsive to the perspectives of forensic practitioners and consider strategies for implementation such that practitioners might be more willing to embrace.

In addition to characterizing the anticipated concerns from forensic practitioners, we also need to be considerate of the needs of the legal system as it relates to the implementation of algorithms. Ultimately, the legal system is concerned with ensuring defendants receive fair and equitable justice under the law. Accordingly, courts will need to consider the admissibility of algorithms against existing legal standards and ensure they are used in a way that does not infringe on defendants' Constitutional rights. In Part IV, we found that this can be particularly challenging given the "black-box" nature of many algorithms and, in some cases, the countervailing legal protections against disclosure of the actual algorithm and source-code. Defendants will often argue the opacity of algorithms fail to demonstrate reliability under Evidence Rule 702 and *Daubert* standards for admissibility. Further, defendants might claim that "black-box" algorithms deprive them of their Fifth Amendment right of Due Process and Sixth Amendment right of

Confrontation. Ultimately, legal scholars have opined that algorithms are likely admissible under existing evidentiary rules and standards; however, (i) they will likely need to be introduced as part of expert testimony, (ii) experts will likely face challenges as a proxy to the algorithm, and (iii) the weight fact-finders give to the evidence could be impacted in unpredictable ways. At times, jurors may be more receptive to the evidence because it is the product of an algorithm. In others, jurors may be more skeptical because of their lack of trust and understanding of the system and deflection of any negative perceptions they may have of the creditability of the expert. Accordingly, experts will need to have sufficient familiarity with the algorithm and be able to answer to the challenges under cross examination. Additionally, besides testifying to the overall result, experts will need to be able to help educate fact-finders on issues related to the validation of the algorithm, conceptual operation of the algorithm, how the algorithm is factored into the overall examination methodology, and the extent and manner in which the algorithm influences the overall interpretation of the evidence. These details are important as they allow us to be responsive to the perspectives of legal stakeholders and consider strategies for implementation such that the legal actors are more willing to embrace

From the above discussion, we see that the implementation of algorithms into many forensic science disciplines are likely to face considerable headwind from practitioners and will require careful consideration of the legal issues and resulting implications. These anticipated challenges, however, we believe are outweighed by the perceived benefits algorithms can provide to the overall evaluation of the evidence. To some extent, as hinted by PCAST, ignoring the calls for algorithms and failing to implement them as a means of empirically substantiating subjective judgment could be inevitably consequential to the enduring validity and admissibility of forensic evidence [7]. However, blindly implementing without careful planning and preparation could disastrous. For that reason, we turn our attention to *how* algorithms can be implemented into forensic practice in a *responsible* and *practical* manner. A responsible implementation requires consideration of issues from a quality assurance perspective to ensure the appropriate foundation has been laid out to support the implementation. Oftentimes, focus is directed toward whether the candidate method has been "validated" or "fit for purpose." In our view, this is too narrow of a focus and, without further context, too broad of a question (i.e., what is the intended purpose and what is considered "fit" for such purpose?). A proper foundation requires a formalized quality management system be in place to ensure conformance with requisite requirements related to: education, training, protocols, validation, verification, competency, and on-going monitoring schemes. Specific guidelines related to each of these elements are available in other sources and within the context of specific examples, such as algorithms for DNA mixture interpretation (e.g., see [181-183]). For purposes of this discussion, each of these key topics are discussed in a more generic sense below.

The first pillar for a responsible implementation is to ensure practitioners and other stakeholders have foundational *education* related to the principles and theory underpinning the algorithm and quantification of the forensic observations, such as probability, statistics, uncertainty, and logic and reasoning. This education is distinct from training on the application of a specific algorithm and should be applied broadly to both practitioners, criminal justice and legal stakeholders, and, if possible, the public at large. For practitioners, this education should enable them to understand and articulate the epistemic limits of the evidence to which the algorithm is being applied and inferences that can be formed during evaluation. For legal stakeholders, this

education should enable judicial actors to understand how algorithms could be applied, how to evaluate the reliability of a given method (e.g., through key performance characteristics represented in validation and verification materials), and the extent to which the algorithms can and should inform their ultimate judicial determinations. For the public at large, this education should expose the public to the realities of forensic evidence interpretation and the role algorithms can play in that process so they have an understanding of the strengths and limitations of forensic evidence for which the algorithms are applied.

The second pillar for a responsible implementation is to ensure practitioners have proper *training* on the algorithm, including appropriate applications of the algorithm. Specifically, to the extent possible, practitioners should understand and explain how the algorithm works, such as what features are taken into account, how they are accounted for, and how the output is calculated and the extent to which outputs might change as inputs vary. Additionally, practitioners should understand the key performance characteristics of the algorithm and the contexts under which those were tested to ensure the data are representative of real-world applications and the circumstances for a given case. In situations where algorithms operate as black-boxes, practitioners may not have a complete understanding of the innerworkings of the algorithm, but should have, at a minimum, a conceptual understanding of the details outlined above in order to understand the applicability and strengths and limitations of the system.

The third pillar for a responsible implementation is to ensure written *protocols* are in place to ensure the algorithm is applied correctly, consistently, and appropriately to evidence in a given case. Protocols related to the standard operations of the algorithm, interpretation guidelines, reporting standards, technical review, and adjudication of conflicts between practitioners' subjective assessments and algorithmic outputs, should be available and publicly accessible. Protocols should clearly articulate what is permissible for input into the algorithm, when the algorithm should be applied, and how the results should be interpreted and accounted for in an overall report. Limitations related to the application of the algorithm and interpretation of the results should also be available and publicly accessible.

The fourth pillar for a responsible implementation is to ensure the algorithm has been subject to an appropriate *validation* to demonstrate its key performance characteristics and "fit for purpose" in a given application. This is a broad term that applies to the foundational validation of the algorithm in terms of its conceptual design, software implementation, and representativeness of casework applications. Prior to validation, the purpose of the algorithm and how it is intended to be applied should be clearly defined to enable a determination of whether the key performance characteristics are acceptable for the intended purposes. The conceptual design of the system should address how the algorithm works, such as what features are taken into account, how they are accounted for, and how the output is calculated and the extent to which outputs might change as inputs vary. The software implementation relates to the accurate coding of the algorithm into a software code for execution. Validation of correct implementation can be done by testing the execution of the software under controlled conditions for which a specific output is expected given the inputs or by a review of the source-code. To enable this, both the software executable *and* the source-code should be made publicly accessible for independent review and testing, including the datasets that have been used in the validation effort. If the source-code is not able to be made publicly available, then it is even more critical that, at a minimum, the software executable is

available. The key performance characteristics of the algorithm (e.g., sensitivity, specificity, repeatability, reproducibility) and the parameters for which the key performance characteristics are calculated (e.g., decision thresholds, etc.) should be derived through empirical testing of the algorithm using samples for which ground truth are known and which are representative in type, quality, and condition of those for which the algorithm will be applied in casework, as applicable (i.e., input samples should vary in type, quality, and condition to the extent that differences in these attributes are accounted for by the algorithm and will impact the output). If the algorithm requires training data (e.g., AI/ML systems), the samples used for testing should be distinct from those used during training. Uncertainty in the calculated performance characteristics should be accounted for and available in the validation documentation. Meuwly et al. provide a detailed guideline for approaching validation for evidence evaluation methods which we consider to be a reasonable framework for addressing these issues [184]. Although the focus is specific to those methods which produce a likelihood ratio, the concepts are applicable to the development and validation of many algorithmic methods designed to assist with evidence evaluation and forensic interpretation [184]. In particular, they address key questions such as "what to measure?" (i.e., performance characteristics), "how to measure?" (i.e., performance metrics), and "what should be observed or deemed satisfactory?" (i.e., validation criterion) [184].

The fifth pillar for a responsible implementation is to ensure the algorithm has been subject to an appropriate *verification* to demonstrate the validity of the system when applied by specific end-users in accordance with a specific set of protocols and in a specific operating environment. Verification (often referred to as internal validation) is not intended to be a repeat of the foundational validation as described above. Rather, it is intended to demonstrate that the system is robust to applications in a specific context (e.g., people, training, protocols) and the key performance characteristics derived during validation are applicable to the conditions and circumstances for which it is applied operationally.

The sixth pillar for a responsible implementation is to ensure the individuals using the algorithm have demonstrated *competency* related to the algorithm, its application, and interpretation of results. Collectively, the pillars of education and training form the foundation for practitioners' knowledge related to the algorithm. Competency testing provides a means of evaluating whether an individual has acquired and demonstrated the requisite knowledge related to the algorithm as well as the ability to apply the algorithm operationally in accordance with applicable protocols and within the limits of its validation. Competency testing should be measured against a minimum standard for acceptable performance and be conducted prior to operational deployment of an algorithm by a specific individual.

The seventh pillar for a responsible implementation is to ensure the algorithm and its application operationally is subject to on-going *monitoring* through proficiency testing and audits of casework applications. The on-going monitoring should account for both the algorithm and its application by practitioners. This monitoring should include (i) routine verification of the software implementation of the algorithm to ensure software or hardware changes do not impact its execution, (ii) the relevance and appropriateness of protocols governing the application of the algorithm, and (iii) practitioners' knowledge and abilities to correctly apply the algorithm. This monitoring should be robust enough to detect vulnerabilities or problems with the algorithm or its

application necessitating preventive or corrective action. Finally, the quality assurance program should be agile enough to improve when preventive or corrective actions are warranted.

In addition to ensuring the necessary foundations are in place from a quality assurance perspective to enable the implementation of algorithms, the next task is to identify a *practical* implementation scheme addressing how the algorithm will be deployed operationally. This should include where in the examination scheme the algorithm will be implemented and the manner in which the outcome of the evaluation will be reported to criminal justice stakeholders. As indicated earlier, the deployment of an algorithm may not necessarily need to be a binary choice of "all or nothing" (*either* the human *or* the algorithm). Instead, implementation can take many different forms with varying degrees to which the algorithm impacts the overall outcome of the evaluation. Decisions related to how the algorithm will be deployed will depend on the intended purpose the algorithm (i.e., is the output of the algorithm intended as the *sole-basis* for the evidential information or is it intended to be used as a *supplemental basis* for the information?), the performance characteristics of the algorithm (i.e., is the algorithm appropriate or "fit" for the intended purpose), and consideration of the tradeoff between the potential benefits of the algorithm and perceived risks for a given deployment scheme. Consideration of these issues will be discipline and context dependent. For purposes of exploring this issue further, we will do so against the backdrop of friction ridge examination. We recognize, however, that this discussion is likely applicable across several other pattern evidence domains.

For context, friction ridge examination is traditionally carried out by human experts and interpretations are based solely on their subjective judgment. Empirical measurements are often not taken and detailed standards for conclusions are non-existent leaving the ultimate determination up to the opinion of the expert. Consequently, assessments made during friction ridge examinations are susceptible to variation from one analyst to another (inter-analyst) as well as by the same analyst from one examination to another (intra-analyst). When considering borderline impressions with marginal quality, these variations might result in differences in the overall conclusion. In the broad spectrum, however, while the lack of empirical measurements and standards do not necessarily mean the practice as a whole is unreliable or fraught with error, it does raise questions as to how reliable the evidence is for the case at hand. Thus, there is a need for the friction ridge community to move towards integrating methods to quantitatively assess the quality and strength of friction ridge impression evidence to enable standardization and provide empirical substantiation to analysts' claims. As it relates to the implementation of algorithms, fortunately, this can be accomplished in several different ways and does not require algorithms to completely supplant the role of the expert. Precisely *how* algorithms should be integrated into standard operating procedures and the implications to practice, however, is an open question. Dror and Mnookin [185] briefly touch on this in the context of Automated Fingerprint Identification Systems (AFIS) databases that have become ubiquitous tools for practitioners over the last several decades to enable more efficient searching, storage, and retrieval of friction ridge impressions and known exemplars. As it relates to algorithms for decision making, however, we propose there are three key issues that ultimately govern how algorithms can be applied in practice: (i) whether algorithms are applied before or after the expert has conducted a traditional examination and formed a subjective opinion, (ii) the extent to which the reported result was dependent upon the output of the algorithm, and (iii) the manner in which conflicting outcomes between the expert's judgement and the algorithm's output are resolved. We recognize that an additional point of debate

is how forensic conclusions and statistical information (e.g. the output of an algorithm) should be articulated to fact-finders and other criminal justice stakeholders (e.g., see [107, 108, 186]). While we view that as important, we consider it beyond the scope of the current issue. This is because irrespective of how algorithms are implemented into practice, the manner in which results are articulated to fact-finders can be quantitative or qualitative, each having benefits and limitations, and deserving of a separate discussion.

Taking into account the factors outlined above, we propose to approach the issue of algorithm implementation, and the different ways algorithms can be implemented, similar to how the automotive industry approached the issue of autonomous driving: describing a formal taxonomy of six different levels of automation (i.e., algorithm influence) ranging from Level 0 (no algorithm influence) to Level 5 (complete algorithm influence). Each level represents a gradual transition from human to machine as the basis for forensic conclusions. For pattern evidence disciplines (including friction ridge examination), we propose the six different levels are: Level 0 (No Algorithm), Level 1 (Algorithm Assistance), Level 2 (Algorithm Quality Control), Level 3 (Algorithm Informed Evaluation), Level 4 (Algorithm Dominated Evaluation), Level 5 (Algorithm only). In levels 0 through 2, the human serves as the predominant basis for the evaluation and conclusion with increasing influence of the algorithm as a supplemental factor for quality control (used *after* the expert opinion has been formed). In Levels 3 through 5, the algorithm serves as the predominant basis for the evaluation and conclusion with decreasing influence from the human. The relationship between human and algorithm as well as the basis for conflict resolution and reported conclusions for each level is summarized in Table 7-1 and described in the discussion that follows.

Level	Name	Narrative Definition	Human Role	Algorithm Role	Conflict Resolution	Basis for Conclusion
0	No algorithm	The human is responsible for forming an expert opinion based on subjective observations without any use of the algorithm.	Evaluation	N/A	N/A	Expert Opinion
1	Algorithm Assistance	The human is responsible for forming an expert opinion based on subjective observations. The algorithm *may* be used *after* an initial opinion has been formed. The algorithm serves as an optional assistance tool supplemental to the expert opinion that may be used at the discretion of the examiner.	Evaluation	Supplemental Assistance (optional)	Expert Discretion	Expert Opinion

2	Algorithm Quality Control	The human is responsible for forming an expert opinion based on subjective observations. The algorithm *shall* be used *after* the opinion has been formed. The algorithm serves as a required quality control supplemental to the expert opinion to ensure the evidence conforms to specified criteria supporting a conclusion.	Evaluation	Supplemental Quality Control	Standard Operating Procedures	Expert Opinion (Algorithm Supported)
3	Algorithm Informed Evaluation	The human is responsible for forming an expert opinion based on the output of the algorithm. The algorithm *shall* be used *before* the opinion has been formed. The algorithm serves as an integrated factor informing the opinion.	Human-Algorithm Integrated Evaluation	Human-Algorithm Integrated Evaluation	Standard Operating Procedures	Algorithm Output (Human Supported)
4	Algorithm Dominated Evaluation	The algorithm is used as the basis for the conclusion. The human serves in an oversight capacity to ensure the algorithm is applied appropriately.	Procedural Oversight	Evaluation	Standard Operating Procedures	Algorithm Output
5	Algorithm Only	The algorithm is used as the basis for the conclusion without any human evaluation or oversight.	N/A	Evaluation	N/A	Algorithm Output

Table 7-1: Levels of algorithm implementation describing the relationship between human and algorithm as well as the basis for conflict resolution and reported conclusions for each level.

Level 0 characterizes traditional practices in the majority of forensic science disciplines. Besides the use of automated tools, such as AFIS to augment searching, storage, and retrieval tasks, algorithms do not have any substantive role in the evaluation of the evidence. The conclusion is based solely on the subjective opinion of the expert. This level is deeply rooted in tradition and represents the vast majority of forensic practitioners today (with the exception of

DNA). Although practitioners are most comfortable with this approach, it has been the focus of increasing criticism from scientific and legal actors for the lack of statistical support.

Level 1 represents the lowest level of algorithm implementation. The human relies on traditional practices for the evaluation of the evidence and is responsible for forming an opinion independent of the algorithm. The expert may then use the algorithm *after* the initial opinion has been formed as an optional quality control. The human is considered the ultimate authority on the overall conclusion and is given complete discretion when to run the algorithm and how the output of the algorithm is considered. Conflicts between the expert's judgment and the algorithm output are not required to be formally adjudicated by standard operating procedures. At most, conflicts between the expert's judgment and the algorithm output may cause the expert to seek a second opinion through formal procedures of consultation or verification; however, the output of the algorithm is not a formal component of the examination scheme and therefore the results are not part of the reported conclusion. Since the algorithm is applied after the expert has formed their opinion and the algorithm is not part of the basis for interpretation or the reported conclusion, courts are unlikely to be concerned with the algorithm. This level may be appropriate when practitioners have not had any prior experience with algorithms. The key benefit of this level is that it provides flexibility for when and how the algorithm is used and a means for practitioners to slowly gain comfort with the algorithm and trust in the output. The key limitation to this level is that there is no formal mechanism to ensure experts are not improperly discounting the algorithm when it might conflict with their subjective assessment.

Level 2 represents a level of implementation in which the algorithm is used as a quality control for the ultimate conclusion reported. The human relies on traditional practices for the evaluation of the evidence and is responsible for forming an opinion independent of the algorithm. The expert then uses the algorithm *after* the opinion has been formed as a *required* quality control supplemental to the expert opinion to ensure the evidence conforms to specified criteria supporting a conclusion. In this scenario, the ultimate authority on the reported conclusion is governed by the standard operating procedures. In order for a particular conclusion to be warranted, the expert opinion must be supported by the algorithm output and conforming to criteria specified by the standard operating procedures (e.g., minimum threshold for a quantitative algorithm output). If the expert opinion is not supported by the algorithm, then the conflict is formally adjudicated in accordance with the standard operating procedures. The protocols proposed by Montani et al. [187] to provide a reasonable framework for addressing conflicts between human and algorithm within standard operating procedures. Although the algorithm has a material impact on the overall conclusion reported from a quality control standpoint, since it was applied after the expert has formed their opinion and therefore is not part of the basis for interpretation, courts are less likely to be concerned with the algorithm. The admissibility of the algorithm may be challenged; however, even if the algorithm was found to be inadmissible (e.g., novel algorithms that are not widely adopted and therefore are not yet "generally accepted"), it is unlikely to materially affect the admissibility of the evidence overall since the ultimate conclusion reported is still based on the expert opinion. This level is appropriate when practitioners have gained some experience with the algorithms and have established reasonable trust in the output. The key benefit of this level is that the algorithm is implemented in a way that provides empirical support for the expert opinion, but does not alter traditional interpretation practices related to the expert opinion. The key limitation

to this level is that the expert does not have the opportunity to leverage the output of the algorithm as a factor when evaluating the overall value of the evidence.

Level 3 represents a key transition point between human and algorithm. In Level 2, the algorithm was used supplemental to the expert opinion (*after* the expert formed the opinion). At this level, the algorithm is used *before* the opinion has been formed. In this scenario, the algorithm serves as factor informing the expert opinion; thus, the expert has the benefit of being able to incorporate the output of the algorithm along with their subjective judgment. The ultimate authority on the reported conclusion is governed by the standard operating procedures. In order for a particular conclusion to be warranted, the algorithm output must conform to criteria specified by the standard operating procedures and must be supported by the expert opinion. If the algorithm output is not supported by the expert opinion, then the conflict is formally adjudicated in accordance with the standard operating procedures (e.g., see Montani et al. [187]). Since the algorithm is applied before the expert has formed their opinion and therefore serves as a basis for interpretation, courts are more likely to be concerned with the algorithm at this level than in lower levels. The admissibility of the algorithm may be challenged since it was a factor taken into consideration when forming the expert opinion; however, similar to lower levels, if the algorithm were found to be inadmissible, it is less likely to materially affect the admissibility of the evidence overall since the algorithm output was one of many factors taken into account when forming the expert opinion. This level is appropriate when practitioners have gained considerable experience with the algorithm and have established trust in the output. The key benefit of this level is that the algorithm is implemented in a way enables the expert to leverage the output of the algorithm as a factor when evaluating the overall value of the evidence. The key limitation to this level is that the interpretation remains dependent on subjective elements from the expert.

Level 4 represents a level of implementation in which the algorithm is used as the basis for the ultimate conclusion reported. In this scenario, the human does not form an expert opinion; rather, the expert determines whether the circumstances of the evidence are appropriate for the application of the algorithm and ensures it is applied correctly and in accordance with standard operating procedures. The ultimate authority on the reported conclusion is governed by the standard operating procedures. Since the algorithm serves as the basis for the conclusion, courts are more likely to be concerned with the algorithm at this level than in lower levels. The admissibility of the algorithm may be challenged since it served as the basis for the reported conclusion. Experts will need to have in-depth knowledgeable about the algorithm and be able to be responsive to questions and challenges to the weight of the evidence during testimony. At this level, if the algorithm were found to be inadmissible, it is likely to materially affect the admissibility of the evidence overall since the algorithm output was the basis for the ultimate conclusion. This level is appropriate when the technology is capable of this type of autonomy and practitioners have gained expert knowledge and experience with the algorithm and have established trust in the output. The key benefit of this level is that the algorithm is implemented in a way enables the expert to oversee the process and ensure appropriate application of the algorithm while the reported results are based on the algorithm output rendering them less susceptible to variations caused by human interpretation. The key limitation to this level is that courts may be *less* receptive to algorithms that operate or manifest as a “black-box” and are difficult to explain how the algorithm generated a particular result.

Level 5 represents the highest level of algorithm implementation for which the algorithm operates in a "lights-out" mode without any human involvement or oversight. In this scenario, the algorithm is fully autonomous and reported results are automatically generated by the machine. The admissibility and weight of the results of the algorithm may be challenged since it operates fully autonomously. At this level, if the algorithm were found to be inadmissible, it is almost certain to materially affect the admissibility of the evidence overall since the algorithm was the sole basis for the ultimate conclusion. This level is appropriate for high-performance algorithms and high-throughput operations where this level of automation is necessary and stakeholders have been informed, understood and accepted the benefits and risks associated with such deployment. The key benefit of this level is that the reported results are based entirely on the algorithm output and completely objective. The key limitation to this level is that practitioners are completely supplanted by the algorithm and without an expert able to testify to the application and overall process, courts are unlikely to be receptive to algorithms as a basis for substantive evidence that lack transparency and explainability to how the algorithm generated a particular result.

The levels of algorithm implementation summarized in Table 7-1 and described above illustrate different ways in which algorithms can be implemented—each with different implications to practice. On the one hand, from a scientific perspective, practitioners should swiftly move toward implementing algorithms at higher levels such that the algorithms provide the predominant basis for conclusions. Doing so would promote improved objectivity and consistency in the reported results. On the other hand, practitioners and courts are unlikely to be receptive to such a swift transition and become almost entirely dependent on algorithms without having the opportunity to gain comfort with the systems and establish trust in the outcome. Further, at higher levels of implementation, practitioners are likely to be expected to have a greater depth and breadth of knowledge about the algorithm and be responsive to questions and challenges during testimony. This may be concerning for practitioners that have traditionally required very little to no need for formal education in algorithms and statistical principles. We propose that the optimal approach is for practitioners to identify a target level of implementation that is practical given the current state of available technology and consideration of the tradeoff between the potential benefits and perceived risks for a given deployment scheme, then establish a plan for implementation that begins with Level 1 as a pilot phase and progresses sequentially through the various levels toward the target. Doing so will allow practitioners to gradually gain comfort with the systems, trust in the outcome, and time to increase their depth and breadth of knowledge that will be expected of them during testimony.

Using friction ridge examination as an example, given the current state of technology available for implementation, target levels of implementation might include Levels 2 or 3—either of which are achievable [50, 96] and least impactful to traditional practices. Level 1 should be short-lived as a pilot phase and first step to gain initial comfort with the system and evaluate the performance of the algorithm when applied operationally. Level 4 is possible as a target given current technology, but likely unsettling to many practitioners and some stakeholders since available algorithms do not fully account for all the types of features and distinguishing characteristics practitioners are able to take into consideration during their subjective assessments. Aside from high-quality samples, such as known-to-known comparisons (i.e., "ten-prints"), Level 5 implementation is likely not practical given the current state of available technology for latent print impressions involving partial and degraded samples. In addition to the technology

considerations yielding Levels 2 and 3 as optimal targets, they offer several other benefits that appear to balance the interests and needs of the various stakeholders. Some of these benefits include: (i) practitioners are more likely to adopt algorithms since they remain empowered to express their expert opinion, (ii) fact-finders would no longer be required to rely on testimony *ipse dixit* as the algorithm would provide a means of ensuring analysts' opinions to be empirically supported, (iii) the resource burden on forensic laboratories that would be necessary for educating and training practitioners related to the underpinnings of the algorithm and statistical concepts is minimal compared to what would be necessary to ensure a depth of knowledge necessary to support Level 4 or 5 implementation, (iv) courts are less likely to be faced with resource-intensive admissibility challenges or concerns of Constitutional infringements since the algorithms are merely supplemental to the evaluation of the evidence (versus the predominant basis thereof), and (v) the admissibility of the algorithm can be considered distinct from the admissibility of the expert opinion. As the technology advances in coming years and practitioners (and other criminal justice stakeholders) become more acclimated with the use of algorithms in forensic science, the expectation is that implementation schemes will continue to progress toward higher levels. Irrespective of the level of implementation, however, the expert will remain critical as a steward to the overall process and necessary for the admissibility of the evidence under the Sixth Amendment [179].

7.8 Conclusion

The implementation of algorithms (e.g., statistical methods) in forensic science is complicated. Although scientific and legal scholars have raised concern that many forensic conclusions lack empirical support and researchers have proposed various statistical or algorithmic approaches, the practitioner community has been reluctant to apply them operationally and their implications to litigation have yet to be fully demonstrated. Reactions from practitioners to statistical interventions have ranged from passive skepticism to outright opposition, often in favor of traditional experience and expertise as a sufficient basis for conclusions. In this paper, we explored *why* practitioners are generally in opposition to algorithms and *how* their concerns might be overcome. We accomplished this by considering issues concerning human-algorithm interactions in both real world domains and laboratory studies as well as issues concerning the litigation of algorithms in the American legal system. Ultimately, recognizing the need to heed the calls for algorithms is inevitable, we propose *how*, not *if*, algorithms could be implemented into operational practice that is both *responsible* and *practical*, such that forensic practitioners and other legal and scientific stakeholders are likely to accept.

Following our exploration of the different issues, we made several observations that enabled us to characterize key challenges to implementation. On the topic of human-algorithm interactions, we found that people tend to exhibit a general aversion to algorithms and prefer to rely on their own judgment—often despite knowledge that their own judgment is typically inferior to that of algorithms. This phenomenon is exacerbated when people possess domain expertise, are faced with high-stakes decisions, and are presented with an algorithm that is susceptible to err. Indeed, forensic science has the conditions for which algorithm aversion is most pronounced. From both anecdotal observations of human-algorithm interactions in different domains and recent research, we found that people tend to be more receptive to algorithms if they are integrated as a

factor that *supplements* as opposed to *supplants* human decision making and the human retains some amount of influence on the ultimate outcome. On the topic of litigating algorithms in the American legal system, we found that this can be particularly challenging given the "black-box" nature of many algorithms and, in some cases, the countervailing legal protections against disclosure of the actual algorithm and source-code. The opacity of algorithms will often trigger admissibility challenges as well as raise concerns over infringements to Defendants' Constitutional rights provided by the Fifth and Sixth Amendments.

In our view, despite these issues, algorithms will ultimately be inevitable to ensure the enduring validity and admissibility of forensic evidence for decades to come. In recent years many forensic science disciplines have been put on notice by formal bodies expressing concern from scientific and legal perspectives that expert opinions need to be empirically supported with statistical data (e.g., see [3, 7-9]). An abrupt shift requiring immediate implementation of statistical and algorithmic methods as a condition for admissibility would be impractical and unrealistic; however, we believe it will only be a matter of time until patience wears and courts limit deference to experts and accept opinions *ipse dixit*. Recognizing the inevitable need for algorithms in forensic science and taking into consideration the issues concerning human-algorithm interaction and litigation of algorithms, we propose a strategy for approaching the implementation of algorithms in a *responsible* and *practical* manner by: (i) outlining the foundations that need to be in place from a quality assurance perspective before algorithms should be implemented, such as education, training, protocols, validation, verification, competency, and on-going monitoring schemes; and (ii) proposing a formal taxonomy of six different levels of algorithm implementation ranging from Level 0 (no algorithm influence) to Level 5 (complete algorithm influence) describing various ways in which algorithms can be implemented, similar to how the automotive industry approached the issue of autonomous driving. Each level represents a gradual transition from human to machine as the basis for forensic conclusions and include: Level 0 (No Algorithm), Level 1 (Algorithm Assistance), Level 2 (Algorithm Quality Control), Level 3 (Algorithm Informed Evaluation), Level 4 (Algorithm Dominated Evaluation), Level 5 (Algorithm only). In levels 0 through 2, the human serves as the predominant basis for the evaluation and conclusion with increasing influence of the algorithm as a supplemental factor for quality control (used *after* the expert opinion has been formed). In Levels 3 through 5, the algorithm serves as the predominant basis for the evaluation and conclusion with decreasing influence from the human. We propose the optimal approach is for practitioners to identify a target level of implementation that is practical given the current state of available technology and consideration of the tradeoff between the potential benefits and perceived risks for a given deployment scheme, then establish a plan for implementation that begins with Level 1 as a pilot phase and progresses sequentially through the various levels toward the target. Proceeding in this fashion will increase the likelihood for adoption across all stakeholders and lead to an overall stronger foundation and improvement to the quality and consistency of forensic science.

8 Operationalization of Algorithms: Personal Reflections and Observations

This chapter discusses the implementation of algorithms into operational practice through anecdotal reflections and observations from my own experiences—both as a laboratory manager and as a private analyst. From those distinct experiences, I reflect on my perspective and discuss strategies for implementation, considerations for policies and procedures, and use of the algorithm in litigation.

8.1 Background

My career as a friction ridge examiner formally began in 2008 with the United States Army Criminal Investigation Laboratory (USACIL). Although my time with the USACIL marked the beginning of my experience as a friction ridge examiner, it was not my first introduction to friction ridge impressions. Prior to joining the USACIL, I spent five years with the Georgia Bureau of Investigation Division of Forensic Sciences as a laboratory technician where I was actively working toward friction ridge examination as a long-term career goal. My training at the USACIL consisted of nearly two years of academic study, practical exercises, competency examinations, moot-court testimony evaluations, and on-the-job training under the direction of a supervising expert. In all respects, the training was well structured and highly resourced. Performance was of utmost importance and heavily scrutinized.

Although the emphasis on training and available resources distinguished the USACIL from many other laboratories, the content of the training and overall culture was generally consistent throughout the discipline. Broadly speaking, we were taught that friction ridge skin is unique and permanent, and that a *competent* examiner can accurately individualize an impression to a single source with 100% certainty and to the exclusion of all others in the world. Over time, these fundamental premises became so strongly ingrained that it was blasphemy to suggest an error could be the result of a methodological issue. Instead, human error was to blame, thus calling into question the competency of the examiner and their long-term employment potential [188]. Ultimately, this led to a culture where errors were not discussed and laboratories handled the issues as private personnel matters. Further, we were taught that during testimony opposing counsel might try to challenge our findings and it was incumbent, and to some extent an expectation of employment, that we were able to "defend[…] the analytical approach and results against rigorous cross examination" [189]. Consequently, challenges to these fundamental premises were often regarded as adversarial litigation tactics to create doubt in the minds of a lay judiciary or stemming from ill-informed academic commentators lacking practical experience and perspective. Simply put, the widespread belief was that the science was well established and that our methodology was impeccable. When faced with challenges, rather than reflecting internally on potential systemic vulnerabilities of the methodology and opportunities to strengthen the foundations of the discipline, our responses were often defensive and protective of traditional practices—myself included [190].

Between 2009 and 2013, challenges related to traditional examination methods and reporting continued to intensify—often referencing the variabilities of human judgment and lack of empirical standards and validation as key concerns. In 2012, I was afforded an opportunity to explore emerging research concerning principles of forensic interpretation and applications of

statistical modeling for friction ridge impressions. Although the statistical modeling was initially represented as providing the capability for the community to leverage more value from friction ridge impressions that would traditionally be considered "no value" through human judgment, there was the additional benefit of providing empirical validation to friction ridge impressions overall. Through this experience, I understood the concerns that have been raised over the years from a different perspective and recognized the need for the discipline to move toward measurement-based practices. In my view, however, the most important utility of the models was their ability to provide empirical substantiation to subjective opinions. I was interested, but less concerned about their potential to leverage more value from friction ridge impressions that would traditionally be considered "no value" through human judgment. From my perspective, we needed to provide an empirical foundation to our existing practices before we tried to expand our capabilities. Up to this point, however, many of the tools capable of providing statistical measures were the product of academic exercises and were never made widely available for operational adoption. From a practitioner perspective, I became concerned we were empty-handed in the face of increasing challenges. Frustrated at the lack of operationally accessible tools, in 2013 we began internal research and development efforts that ultimately led to the development of our own tools—DFIQI and FRStat (described in chapters 2 and 3, respectively).

From the outset, when DFIQI and FRStat were first contemplated, our intent was to develop tools that were able to provide quantitative representations of the findings. Through these quantitative representations, we can provide empirical substantiation to examiners' subjective opinions and construct more robust quality assurance programs. The intent was not to automate the examination process or supplant the role of the examiner. Instead, the intent was to develop a system that was operationally useful and worked *with* the examiner—a system that required little computational complexity, relied on the same features as examiners, and was conceptually consistent with traditional practices. Our goal was to provide a tool that was broadly accessible to the practitioner community, conformed to existing and familiar examination practices, and was easily articulable to lay audiences. With this in mind, both DFIQI and FRStat were designed to rely solely on friction ridge minutiae—ridge endings, bifurcations, and dots—as they are the most robust features available for comparison and the most often relied upon by the practitioner community. For DFIQI, the intent was to develop a means of measuring the quality and quantity of features available. As described in chapter 2, this was accomplished by measuring the clarity of ridge detail immediately surrounding each feature along with the quantity of features available to produce overall quality scores for the impression related to predictions of value, complexity, and difficulty based on datasets representing examiner consensus. For FRStat, the intent was to develop a means of measuring the significance of an association between two impressions. As described in chapter 3, this was accomplished by measuring the similarity of feature configurations (in terms of the locations and angles of corresponding minutiae) and evaluating the extent to which the measured similarity would be expected based on datasets of results from impressions known to have been made by same sources and impressions known to have been made by different sources. As the similarity between the configurations increase *and* the number of features in the configurations increase, the values produced by the FRStat also increase representing a stronger association between the impressions. From a technological standpoint, neither DFIQI nor FRStat is particularly novel or ground-breaking. Both are (intentionally) relatively simplistic systems that rely on familiar principles of friction ridge examination and apply established statistical methods

to produce the numerical results. The novelty comes from the ability to use the numerical results in practice.

In December 2014, I assumed a laboratory management role responsible for overseeing the latent print unit of the USACIL, which including forensic fingermark examinations, forensic footwear examinations, forensic tire track examinations, and the Automated Fingerprint Identification System [AFIS] program in support of the criminal investigative mission of the U.S. Department of Defense (DOD) world-wide. In that capacity, I held myself personally accountable for the reliability of the examination results and responsible for ensuring the unit had the proper tools and resources to accomplish this. My top priorities included reframing our reporting language to be more epistemologically defensible, expanding the examiners' exposure to basic statistical concepts and core scientific principles, and strengthening our quality assurance program through the implementation of algorithmic tools (e.g., such as DFIQI and/or FRStat). As a demonstration of my commitment and to stimulate further dialogue on these issues, I published a commentary reflecting on the evolution of my own perspectives over the last few years and vision for the future [191]. Although we were ultimately successful in accomplishing these objectives, as we discussed in chapter 7, the implementation of algorithmic tools in practice is not straight forward. As for the specific algorithms implemented, our priority was on the FRStat algorithm. The DFIQI algorithm had not yet been completed and had taken a lower priority to the FRStat. In the sections that follow, my discussion will focus on reflections related to the implementation of the FRStat algorithm.

8.2 Implementation

The implementation of an algorithmic tool—FRStat—at the USACIL was a two-year effort. The first step toward implementation began in February 2015 with the internal decision to revise our reporting language to be more epistemologically defensible. This involved moving away from a categorical framework using claims that a specific individual is *the* source of an impression or using terms that imply certainty for a single-source attribution. Instead, it involved moving toward a probabilistic framework in which the conclusions were specific to the observations (i.e., probabilities of the observations given propositions versus probabilities of propositions given observations). The use of calculated probabilities was not part of this initial transition. This initial transition focused on changing the framework in which conclusions were conveyed and involved using verbal expressions of probabilistic concepts as opposed to numerically calculated probabilities. Likewise, changes to thresholds for conclusions (whether explicit or implicit) were also not part of this transition. Formal references to the term "identification" were replaced with the term "association" and used as a general descriptor of inclusive results; however, this was merely semantics—the same expectations for conclusions of "association" applied as though they were expressed as a categorical statement of "identification" before the transition. Simply put, the intent was not to expand the scale of conclusions to include those comparisons which would traditionally be reported as "inconclusive." In the new framework, although examiners' *personal* opinions were that two impressions originated from the same source, their *professional conclusion* was to be expressed in a way that conformed to epistemic limits of what could be supported by available research and would ultimately be compatible with probabilistic outputs from an algorithmic tool.

When the decision was first announced internally in February 2015 that the reporting language was expected to change, reactions from the examiners were reminiscent of those discussed in chapter 5. A few were on board, but the majority had not fully understood the need for the change, were unfamiliar and uncomfortable with probabilistic concepts, and were concerned that stakeholders (notably investigators, attorneys, judges, and jurors) would not understand what was being conveyed and how it should be interpreted. As a result, we held off on executing the policy decision, and over the course of the next nine-months we invested heavily in examiners' training and education. Our goal was to strengthen their understanding of the issues concerning reporting and increase their familiarity and comfort with using probabilistic concepts. We assembled numerous workshops using instructors both internal and external to the USACIL and facilitated open discussions concerning the major criticisms facing the discipline, introduction of core scientific principles and concepts related to probability theory and statistics, and strategies for articulating findings in a probabilistic framework during testimony. As examiners' understanding of the issues increased and their familiarity and comfort with probabilistic and statistical concepts improved, we applied a consensus approach to developing the actual reporting language that would be used moving forward. In November 2015, we formally announced the revised reporting framework and new language [95]. Although we were prepared for questions from investigators or attorneys during litigation following this transition, surprisingly, we experienced very few.

Once the revised reporting language was implemented in November 2015, our next objective was to move toward implementing the FRStat as a means of providing a measure of empirical support to examiners' opinions of "association." Between November 2015 and December 2016, we finalized the development and validation of the FRStat (described in chapter 3) and began to lay the rest of the foundation to facilitate implementation. Briefly summarized, throughout 2016, we continued to expose the examiners to more advanced concepts related to probability theory and statistical inference and specific technical details concerning the FRStat. We continued to engage several members of the stakeholder community, including research scientists and statisticians from government and academia, attorneys, and other practitioners, to provide broad exposure of our approach and solicit feedback through conference presentations [192-197], hosting in-depth technical reviews and round-table discussions, and independent reviews of the underlying technical details and validation materials. With input from both practitioners and external stakeholders, including defense attorneys, we established draft policies and procedures related to the use of the algorithm—specifically *who* should use the algorithm, *what* the algorithm should be used on (e.g., eligible inputs), *when* to use the algorithm, *how* to use the algorithm, *how* the results should be reported, and *how* conflicts will be managed and adjudicated between the examiner and the algorithm (discussed in more detail below). Once the protocols were fleshed out and as we neared implementation, examiners were subject to formal training modules, practical exercises, and oral boards to demonstrate their competency with the application before being authorized to use the algorithm in casework.

Finally, in January 2017, the FRStat was implemented into operations at the USACIL as a "soft-launch" evaluation period. Drawing from the taxonomy proposed in chapter 7, this is best described as Level 1 implementation. We remained at Level 1 for nearly three months during which time the examiners gained comfort using the system operationally and we monitored the

robustness of the algorithm in practice, operating protocols, and feedback from the examiners. In March 2017, we transitioned to Level 2 implementation and formally announced the additional revision to our reporting language to include statistical calculations [96]. The only distinction between our implementation at the USACIL and the Level 2 implementation described by the formal taxonomy in chapter 7 is that the numerical results from the FRStat were included on the formal report. Otherwise, our implementation was generally consistent with that in the formal description of Level 2.

Although we were ultimately successful with implementing the FRStat operationally, it would be misleading to imply it was a smooth or straight forward process. On the one hand, as we moved toward operationalizing it, examiners were much more receptive to the decision to implement FRStat compared to when the initial decision was made to revise our reporting framework from categorical to probabilistic (without calculated probabilities). By this point, the examiners had been exposed to probabilistic and statistical concepts for nearly two years, were generally familiar with the FRStat algorithm, and had contributed to the validation. Further, the examiners had experienced testifying in court using the revised reporting language from 2015 and had not met resistance or confusion from other stakeholders. Interestingly, the resistance they faced was actually from their own peers from other laboratories—it was as if the USACIL had broken a sacred code and conceded to the "critics." On the other hand, however, although the examiners were generally receptive to the idea of implementing the FRStat algorithm, the examiners did express concern with respect to the tactical details regarding how the algorithm would be operationalized and the policies and procedures governing its use and reporting of results. Navigating the operating protocols and balancing various perspectives on how the algorithm should be used operationally proved to be the least straightforward and, to some extent, one of the most challenging aspects of the implementation. At the time we were working through these issues, the taxonomy described in chapter 7 was non-existent and we were less familiar with the existing behavioral sciences research regarding human-algorithm interactions. Consequently, we deliberated over these issues with little precedence available, both as a management team and in collaboration with the examiners, and often solicited feedback on specific issues from external stakeholders to ensure we were addressing these issues with as much balance to the various perspectives as we could. Admittedly, we found there are various perspectives on these issues and multiple reasonable approaches that could be applied to address them. In section 8.3 (Policies and Procedures), I briefly summarize the key elements of specific questions that we grappled with in the context of FRStat implementation and some of our rationale for the decisions made concerning the policies and procedures governing its use.

Aside from the specific issues we had to navigate relating to the protocols, readers may wonder how well we measured up to the pillars proposed in chapter 7 concerning a "responsible" implementation. Overall, I would argue we addressed each of the pillars in a satisfactory manner and were thoughtful and deliberate in our implementation scheme. However, I fully admit that there are some areas where improvements could have been made or where we were unable for one reason or another to be as thorough as we might have hoped. In the discussion that follows, I'll briefly contrast specific elements of our implementation of the FRStat algorithm with those seven key pillars described in chapter 7:

Education

This pillar concerns the foundational education related to principles and theory underpinning the algorithm and quantification of the findings, such as probability, statistics, uncertainty, and logic and reasoning. Ideally, this pillar would be addressed through formal education in post-secondary or graduate level coursework prior to employment. However, these requirements have not yet been codified in national training standards, included as pre-requisites to employment, or widely adopted by forensic science programs in the United States. Until this is a formal requirement and widely available, laboratories will need to rely on in-house, *ad-hoc*, or other continuing education opportunities to expose examiners to these concepts. In our situation, we were fortunate to have resources available to address this internally over the course of approximately two years. The first year consisted of a series of workshops using instructors both internal and external to the USACIL. In the beginning, the workshops were given weekly by internal instructors, which included concepts such as logic and reasoning, probability theory, and descriptive and inferential statistics. After several months, the frequency decreased and the topics shifted toward introductory principles related to statistical modeling, quantification of the findings (e.g., random match probabilities, likelihood ratios, posterior probabilities), decision rules and classification, and uncertainty. External instructors were brought in, and examiners were given opportunities to attend statistically-oriented workshops externally. Although participation was expected among all examiners, formal testing of these concepts was not administered.

In addition to our internal focus, we made considerable effort to educate the practitioners from other laboratories and legal stakeholders through presentations and workshops at professional meetings and conferences (e.g., between 2016 and 2018, we provided over twenty presentations and workshops [50, 192-215]). One area where we could have improved our educational outreach efforts is for the public at large. Although we recognized the importance of this, we were primarily focused on practitioners and legal stakeholders, as they were the most likely to be immediately impacted by our use of the algorithm.

Training

This pillar concerns the proper training on the algorithm, including appropriate applications of the algorithm (i.e., how the algorithm works, what features are taken into account, how the features are accounted for, how the output is calculated and the extent to which outputs might vary as inputs vary), key performance characteristics of the algorithm, and strengths and limitations of the system. In our situation, we addressed this pillar through a formalized training module administered internally. The training included both oral instruction and relevant literature, including the user manual, technical validation report, underlying algorithms and calculations, and general applicability of the algorithm along with its limitations. Admittedly, we struggled with the level of training the examiners should be expected to have as it relates to the underlying algorithm itself (e.g., should the examiners have a general familiarity of the algorithm to understand and be able to articulate how it works conceptually and its limitations? Or, should the examiners have in-depth technical knowledge of the underlying calculations as might be expected by the developer or engineer?). At the beginning, we erred on the side of too much information and included specific details related to the underlying calculations. As we began administering

the training materials to examiners, we monitored feedback from examiners regarding their comfort and receptivity. Ultimately, we decided that requiring knowledge of in-depth technical details concerning the underlying algorithmic calculations might be overkill in that the examiners are unlikely to be expected to have that level of technical knowledge that would otherwise be expected of the developer or engineer. If a situation were to manifest where that level of technical knowledge was necessary, we had other sources and reference materials to accommodate. As a result, the training materials for the examiners were modified to ensure the they had a general conceptual familiarity of the algorithm design, development, and underlying operation (rather than in-depth technical details of the algorithmic calculations), and emphasized more in-depth understanding of issues related to operating characteristics (e.g., user manual), key performance characteristics (e.g., validation report), interpretation of the results, and general strengths and limitations of the system.

Protocols

This pillar concerns the existence of written protocols to ensure the algorithm is applied correctly, consistently, and appropriately to the impressions in a given case. Protocols include standard operations of the algorithm, interpretation guidelines, reporting standards, technical review, and adjudication of conflicts between the examiner and the algorithm. As indicated earlier, on the surface this pillar might seem to be one of the easier pillars to address; however, as we began unpacking the various questions and issues that needed to be addressed, we realized it is actually quite complex—so much so that it is deserving of its own section (see section 8.3 Policies and Procedures below). Once we had our policies and procedures codified, however, despite our best efforts, we were unable to make them publicly accessible (e.g., accessible on a public web-server, etc.) since doing so would have violated controls imposed at higher levels in the command structure. However, the protocols were able to be shared with relevant stakeholders as part of a formal discovery request during litigation or with other law enforcement entities in the United States upon formal request.

Validation

This pillar concerns the foundational validation of the algorithm as it relates to its key performance characteristics and "fit for purpose" in a given application. In our situation, I believe we sufficiently addressed this pillar. We ultimately published the entire validation materials in a peer-review journal, which allowed it to be publicly accessible. The published article is reproduced in full in chapter 3. Although the technical report was not published until 2018, prior to implementation we consulted research scientists and statisticians from government and academia, hosted in-depth technical reviews and round-table discussions, and solicited independent reviews of the underlying technical details and validation materials. As detailed in chapter 3, we discuss the key performance characteristics of the algorithm (i.e., sensitivity, specificity, repeatability, and reproducibility), the decision thresholds for which those characteristics apply, limitations of the algorithm, and considerations for policy and procedures governing the application of the algorithm. The datasets used for training and modeling the statistical parameters were distinct from those used for testing, reasonably sized, varied in terms

of their quality, representative to casework, and derived from both controlled and uncontrolled conditions for which ground truth was known and from actual casework.

Verification

This pillar concerns whether the algorithm has been subject to appropriate verification to demonstrate the validity of the system when applied by specific end-users in accordance with a set of protocols and in a specific operating environment. In our situation, we addressed this pillar during the Level 1 implementation (i.e., "soft-launch" evaluation) between January 2017 and March 2017. By this time, the algorithm had been validated, examiners had been trained and competency tested, and draft protocols had been established. The Level 1 implementation allowed us to monitor how well the system worked when applied by the examiners themselves in actual casework conditions. For example, we were able to monitor examiners' use of the system and interpretation of the results, how often examiners' opinions conflicted with the algorithm output (and if so, what was the cause for the conflict), and whether the protocols addressed all the circumstances that could arise when using the system in practice. During this time, we observed that the majority of circumstances in which examiners' opinions conflicted with the output of the algorithm were often due to examiners either not properly annotating the features in a way the algorithm could detect or not thoroughly documenting all the features available as input to the algorithm. Instead, examiners were still accustomed to only documenting a subset of features that were available and which satisfied their personal thresholds to support their opinion. In some situations, this was adequate. In others, the algorithm produced a value that fell below the quality threshold we had imposed to qualify as an "association" versus "inconclusive" conclusion. Over time, as the examiners became more accustomed to the algorithm, these instances declined considerably.

Competency

This pillar concerns whether the examiners using the algorithm have demonstrated competency related to the algorithm, its application, and interpretation of results. In our situation, examiners were administered a formal training module and competency tested prior to being authorized for use of the algorithm in casework. The competency test included oral examination and practical exercise in which examiners had to demonstrate requisite knowledge and understanding of the proper use of the algorithm, interpretation of results, and limitations of the method as well as the ability to apply the algorithm to a specific case and generate accurate results.

Monitoring

This pillar concerns the algorithm and its application in casework is subject to on-going monitoring through proficiency testing and audits of casework. In our situation, I believe we sufficiently addressed this pillar. On-going monitoring was already a normal part of the quality assurance program at the laboratory, so it did not require considerable changes to include FRStat in that existing monitoring scheme. Our routine monitoring was accomplished in several ways:

(a) The software was subject to annual performance checks to ensure software and hardware changes on the computers had not impacted its ability to execute. This was accomplished by running the algorithm on mated print pairs and evaluating the results against ground truth and expected output values. At a minimum, this was done on an annual basis, but anytime the program was executed in normal casework, examiners monitored whether the results were inconsistent with their expected output or appeared atypical. (b) Examiners' on-going proficiency was evaluated by monitoring their use on annual proficiency tests and during normal casework through technical reviews of all case materials prior to issuing a report to ensure accurate annotation of features, application of the software, and conformance with applicable policies and procedures. (c) The appropriateness of the protocols was reviewed and evaluated annually by the Quality Assurance Lead and on an ad-hoc basis when examiners triggered a condition for the Quality Assurance Lead to review a case (such as permission to deviate, conflicts between the examiner's opinion and algorithm output) or through open feedback of circumstances encountered during casework or technical review. (d) The thoroughness and appropriateness of the technical review as a quality control mechanism was evaluated by the Quality Assurance Lead by randomly sampling cases throughout the year and conducting case audits consisting of a re-review of the case and evaluation of the quality of the technical review performed at the time the report was issued. As concerns were noted or issues were identified concerning the use of the algorithm, output of the algorithm, or reporting of the results, preventive and corrective measures were implemented to address the issue (e.g., additional training to analysts, updating policies and procedures).

8.3 Policies and Procedures

8.3.1 Background

As we moved toward implementing FRStat into practice, one of the most challenging aspects to sort out was how the system would be used operationally. First, we had to consider the limitations of the algorithm, the most significant being: (1) the similarity statistic values are dependent upon the subjective detection and annotation of friction ridge skin features by the human expert, (2) the method is only able to consider what the expert annotates and is not able to evaluate the accuracy of feature annotations by the expert, (3) the algorithm requires a minimum of five features and a maximum of fifteen features (the minimum was due to the manner in which the similarity calculations are performed, and the maximum was a decisional cut-off due to computational limitations given the current software implementation), (4) the algorithm accounts for lateral distortions of friction ridge skin impressions on flat surfaces and may not capture all types of extreme distortions which may be encountered in practice, such as substrate, matrix, or photographic effects, and (5) the algorithm is not designed to evaluate all aspects of the impression, such as pattern type, feature type, ridge counts, and other types of features considered by an expert; thus, the quantitative results are artificially attenuated and conservative.

After taking into account the limitations of the algorithm, we then had to take into consideration the operational impacts of the implementation, such as: (1) impacts to workflow, throughput, and resource, (2) examiners' perspectives and preferences to maximize their receptivity to the algorithm, (3) stakeholders' perspectives and preferences to maximize their support of the algorithm, and (4) other second- and third-order effects of implementation related

to the potential for cognitive biases, behavioral science issues related to human-algorithm interactions and altered interpersonal dynamics, and legal implications of the algorithm.

Navigating these various issues was not straight forward. Admittedly, at the time we were largely unaware of the discussion presented in chapter 7, and many of the more recent publications were not available. Although we were fortunate to be able to engage other stakeholders' perspectives related to legal implications and speculate about conditions which might increase or decrease certain sensitivities, feedback was not unanimous. We often received multiple perspectives on issues, none of which had really been codified by historical or legal precedent. General issues we had to work through included: *who* should use the algorithm, *what* the algorithm should be used on (e.g., eligible inputs), *when* to use the algorithm, *how* to use the algorithm, *how* the results should be reported, and *how* conflicts will be managed and adjudicated between the examiner and the algorithm. As we began unpacking these topics, we found ourselves facing several tactical questions that needed to be addressed, such as: Should the algorithm be run before or after the examiner documented their opinion? Would the algorithm be run before or after the examiner's opinion was subject to Verification (as part of the ACE-V methodology)? Would the algorithm be applied to the case examiner's feature annotations, the verifying examiner's feature annotations, both individually, or only those features that both examiners independently observed? Could examiners add, remove, or change their feature annotations and re-run the application? What happens when there is a conflict between the examiner's opinion and the output of the algorithm? What happens if the examiner has impressions that bear conditions which exceed the limits of the algorithm (e.g., too few features, simultaneous impressions, composite impressions, etc.)? How should the results be reported? What happens if a court takes issue with the algorithm and it is not admitted? There were no clear answers to these questions and there are multiple reasonable approaches that could be taken to address them.

In the discussion that follows, I briefly summarize key considerations at the time related to each specific question and some of our rationale for the decisions made concerning the policies and procedures governing the use of the algorithm.

8.3.2 Key Questions and Considerations

Should the algorithm be run before or after the examiner documented their opinion?

In general, algorithms are often thought of as being used to inform a final result, and therefore some perspectives held that the algorithm should be run before the examiner formed their opinion. However, friction ridge examination does not need the algorithm for an analyst to complete their examination and form an opinion. This is how the field has been operating for the last century—through visual observation and human judgment. From a practical standpoint, it would be nearly impossible for an examiner that has been trained and practicing in this way for several years to completely reverse course and resist the temptation to form an opinion prior to running the algorithm. If we accept that the examiner is likely going to form their opinion holistically irrespective of the algorithm, then the issues is whether we require the algorithm to be run prior to the examiner documenting their opinion (thus committing to their opinion in the case record). As we considered this issue, we were very cautious on not disrupting the existing

examination methodology that had become so familiar and widely accepted in the discipline and courts. By requiring the algorithm to be run before the examiner documented their opinion, we could not be sure whether the examiner would have arrived at the same conclusion independently, thus running the risk of disrupting this methodology and opening the door for other unexpected and unintended consequences (e.g., would the examiner become overly reliant on the algorithm or obfuscate responsibility for forming their own opinion?). If there were an unforeseen issue concerning the algorithm, this could jeopardize the reliability of the algorithmic output and expert's opinion. Particularly at the beginning, we believed it was important to keep the processes related to human judgment isolated from potential algorithmic influences. This way, if needed, we could always fall back on the results of the human judgment if there was concern over the algorithm.

Further, the construct of the algorithm is such that it requires the examiner to annotate the features believed to correspond. While documentation of corresponding features is required for impressions opined to be associated with one another, it is not required or commonly practiced when features are clearly different from one another and the impressions are excluded as originating from the same source. From an operational perspective, requiring the algorithm to be run before the examiner documented their opinion would require it to be run on *all* comparisons, including those involving clear exclusions which would further require examiners to document non-corresponding features for the sole purpose of input to the algorithm and satisficing the requirement. This would create a substantial amount of extra work on the laboratory with little technical benefit. With the limited resources available for most laboratories, this would be impractical. For these reasons, we decided to have the FRStat algorithm run *after* the expert documented their original opinion and limited to those impressions which the examiner opined were an association. After all, it is the lack of empirical support for association opinions which are most concerning. Although errors related to exclusion decisions are an important issue the discipline needs to address, the FRStat algorithm is not the appropriate solution to address that.

Would the algorithm be run before or after the examiner's opinion was subject to Verification (as part of the ACE-V methodology)?

Having established that the FRStat algorithm should be run after the initial case examiner documented their opinion, the next issue was whether the examiner should run the algorithm before or after Verification. On the one hand, having the examiner run the algorithm before verification could be more operationally efficient—it could provide immediate feedback to the examiner before the laboratory expends the resources to assign another examiner for Verification (e.g., if there were a cause for concern, the case examiner could address it before requesting verification). On the other hand, Verification serves as an important quality control for monitoring concordance of examiners' opinions. If the examiner ran the algorithm prior to Verification and received an expected result causing them to pause before sending to another examiner for review, it could diminish the value of this important monitoring scheme within the quality assurance program. Further, if the verifying examiner was exposed to the algorithm results, it could unduly bias them thereby disrupting their ability to conduct an independent examination. Ensuring the verifying examiner was not exposed to the verifying results could be a solution; however, if the examination resulted in conflicting opinions between the case examiner and verifying examiner, the results of

the algorithm could then impact the established conflict resolution process. A common first step when conflicting opinions occur in a case is for the two examiners to discuss their differences in their observations and interpretations that ultimately led to the different opinions. From a management perspective, if the results of the FRStat algorithm were available, it could bias these discussions in ways that might be unpredictable from a behavioral sciences standpoint (e.g., would examiners concede to the results of the algorithm without scrupulously working through the issues? Or would dominant personalities use the algorithm as evidence in favor of their position without consideration of the accuracy of the differing feature interpretations or annotations that were inputs to the algorithm?). To be sure the algorithm did not disrupt any existing practices, we decided to have the FRStat algorithm run *after* the Verifying examiner completed their independent examination and both examiners formed the same opinion of association. If conflicting opinions occurred between examiners during Verification, established conflict resolution procedures would apply before the algorithm were run.

Would the algorithm be applied to the case examiner's feature annotations, the verifying examiner's feature annotations, both individually, or only those features that both examiners independently observed?

As described in chapter 3, the FRStat algorithm requires an examiner to document the features believed to correspond before running the algorithm. Consequently, the output results of the algorithm are dependent on the input annotations by the examiner. It is well known that examiners often vary in their feature annotates. Naturally, this will lead to the numerical output of the FRStat to vary between different examiners. Although the numerical results will vary (the extent to which they are expected to vary is described as part of the validation in chapter 3), the overall inference of whether the results support an "association" opinion vs. "exclusion" opinion are not expected to vary. However, this leads to questions concerning which set of features to use when running the algorithm and how the results should be recorded. There are essentially four different options available: (i) only the features documented by the case examiner are used, (ii) only the features documented by the verifying examiner are used, (iii) the algorithm is run twice—once on the features documented by the case examiner and again using the features documented by the verifying examiner, or (iv) the algorithm is run once only using a consensus set of features that both examiners independently documented during their respective examinations. The first and second options are straight forward. Of those two options, the first option seems most reasonable from a legal perspective since the case examiner (not the verifying examiner) is often the one that is signing the report and responsible for testifying to the results. The verifying examiner is often viewed as a quality control rather than a joint-examiner working the case. The third option is sensible on the surface—run the algorithm on both sets of annotations and record the results. However, if the numerical results were included in the report, which one is provided? Would only one be included, both, or some combination of the two (e.g., average or median)? If only one was included, then the case examiner's result seemed most sensible since they would be the one signing the report and testifying, and it is unlikely the case examiner would be permitted to testify to the results of the verifying examiner. If both results were included, the same concerns remained about the case examiner testifying to the results of the verifying examiner as well as the additional concern that including multiple results for the same comparison could be confusing to other stakeholders. Alternatively, a mathematical combination of the two numerical results (e.g.,

the average) could be taken and included on the report. While this makes sense from a scientific perspective, there were concerns that the reported numerical result (as a mathematical combination of the two) could be influenced by features interpreted by the verifying examiner that were not initially interpreted by the case examiner. Consequently, there was concern that averaging the two results could cause the case examiner to indirectly testify to interpretations that were not theirs, which would likely not be permitted by courts. The fourth option would be a viable solution if both sets of annotations are considered—only run the algorithm on the consensus set of features that both the case examiner and the verifying examiner interpreted consistently. Of the four options available, the first and fourth options were the most sensible, justifiable, and practical. Between those two, however, the first option offered the most efficiency and captured all of the original interpretations of the case examiner. The fourth option offered the most robustness to variations in feature annotations, but would require additional steps for the examiners to reconcile the consensus feature sets for every case. Ultimately, preference was given to the first option—only those features documented by the case examiner were run through the algorithm after Verification was completed. This was most preferred by practitioners, seemed acceptable from different stakeholders' perspectives, and was most efficient in terms of operational throughput.

Could examiners add, remove, or change their feature annotations and re-run the application?

Two of the major limitations of the FRStat are that it is dependent upon the personal detection and annotation of friction ridge skin features by the human expert and it is only able to consider what the expert annotates and is not able to evaluate the accuracy of feature annotations by the expert. As a result, the numerical output of the FRStat is a quantitative reflection of the feature annotations input by the examiner. Traditionally, in the absence of the FRStat algorithm, examiners will document their features for illustration purposes rather than measurement purpose. At times, a feature might be documented inaccurately (although for illustration purposes, it is clear what feature the examiner is referring to) or pairs of corresponding features between the impressions might not be annotated very precisely. Although accurate and precise annotations are desired, they are not required (provided a third-party reviewer could understand what features the annotations refer to). With FRStat, however, the algorithm measures the similarity of the annotations, thus accuracy and precision are important and reflected in the numerical output. All else equal, inaccurate or imprecise annotations will often result in lower similarity whereas accurate and precise annotations will result in higher similarity. If, after running the FRStat, examiners realize that their annotations were imprecise or inaccurate, causing the result to be erroneously low, could they change their feature annotations and re-run the system? Likewise, in the absence of the FRStat algorithm, examiners will often document only enough features to support their opinion irrespective if there are more features available between the impressions. Thus, if, after running the FRStat, they realized there were additional features which should have been included, could they annotate the additional features observed and re-run the system?

The short answer is no. This was a major point of discussion during implementation. From an examiner's perspective, it is well known that the examination methodology (i.e., ACE) is often thought of as a circular process. Examiners will document the features observed on the mark during analysis and, if suitable for comparison, proceed to compare it with an exemplar print. During comparison, examiners might change how they initially interpreted a feature, realize they

erroneously interpreted certain features, or realize they erroneously overlooked certain features. Although it is hoped that these situations are not frequent, the reality is that they do occur and are generally permissible in the discipline. Oftentimes, however, such changes are not thought of as being consequential to the overall outcome of the examination. Because these practices were generally considered permissible under traditional examination schemes, some examiners believed it should be permissible with FRStat. Other stakeholders' perspectives were different. Although the modifications to the features might be reasonable and justifiable, there is no way to know whether the examiners' interpretations were biased by exposure to the exemplar print or their preconceived opinion of an association with some magnitude. Consequently, we needed to establish a robust policy around this issue while also recognizing that these situations will occur (i.e., examiners will make mistakes with their annotations) so we needed to decide how we would monitor and address those issues. Clearly, offering a *carte blanche* permission to modify features at the whim of the examiner was not appropriate. Nor was it appropriate to completely disregard that evidence or erroneously under-report the results. In these situations, the critical issue was coming up with a way to ensure the features input to the FRStat were interpreted independent of potentially biasing information and a system was in place to enable management to monitor conditions in which these issues manifested and control when and how exceptions would be permitted. In chapter 4, we discussed the DFIQI (or similar quality metric algorithm) could be a suitable solution to help address this issue by controlling the inputs to the FRStat algorithm in terms of the quality of the features. Theoretically, those features annotated in high quality areas should be annotated both accurately and precisely given the high clarity of ridge detail and less likely to be subject to biased interpretations. Our issue was that the DFIQI was not yet complete and validated. As a result, we had to come up with other solutions.

After deliberating on this issue for some time and seeking feedback from other scientific and legal stakeholders, we ultimately decided to address the competing perspectives and mitigate the potential for cognitive biases by applying procedures that governed which features were eligible to be run by FRStat from the outset. The procedure expanded on concepts of "linear ACE-V," which were already in place prior to the introduction of the FRStat. The concept of "linear ACE-V" requires features to be documented on the mark during Analysis representing what features were observed and how they were interpreted prior to exposure to the reference exemplar, then features observed during comparison are documented separately representing what features were observed and how they were interpreted after exposure to the reference exemplar. Doing so creates a transparent record of any changes to features before and after exposure to the reference exemplar. To control the inputs to FRStat and ensure that the numerical output was not influenced by features which might have been subject to biased interpretations, we instituted a concept of "eligible features" that could be input to the FRStat. An "eligible feature" is one in which the examiner documented as being observed during Analysis (prior to exposure to the reference exemplar), expressed as having high or medium confidence in its presence (e.g., using a color-coded classification scheme such as "GYRO"), and documented as corresponding during Comparison *without* changes in the interpretation between Analysis and Comparison. A change in interpretation between Analysis and Comparison was determined by differences in the documentation of the feature between the two. Generally speaking, feature annotations tend to be approximately the width of a ridge. If the documentation of a feature during Comparison overlapped the documentation of that same feature during Analysis, then the feature was considered to be consistently interpreted between Analysis and Comparison and therefore eligible

for entry into the FRStat. However, if the documentation of the feature during Comparison did not overlap the documentation of that same feature during Analysis, then it was considered to have been subject to a different interpretation after exposure to the reference exemplar and was not eligible for entry into the FRStat. Features that were not eligible for entry into the FRStat could remain documented on the images for illustrative purposes; however, they would not be taken into account by the numerical output of the FRStat (thus resulting in a lower result than otherwise might be expected if those features were included). The accuracy of feature annotations and proper application of this policy was monitored in every case during technical review.

In situations where only a subset of available features was eligible for input to the FRStat and the numerical results were insufficient to warrant a conclusion of "association" from being reported, the examiners were required to notify the Quality Assurance Lead for resolution. The Quality Assurance Lead would arrange for an independent examination of the impression. If features were deemed "eligible" during the independent examination that corresponded to the same features annotated by the original case examiner during Comparison (but were ineligible based on the original case examiner's documentation), then the Quality Assurance Lead would grant permission for those features to be input to the FRStat. The independent examination was considered adequate evidence that the features were able to be interpreted without potential biasing impacts from exposure to the reference exemplar. The case examiner's original documentation would remain, and the circumstances surrounding the situation would be documented in the case file. If the numerical output of the FRStat was sufficient to warrant a conclusion of "association" then the case would proceed as normal. However, if the numerical output of the FRStat was still insufficient to warrant a conclusion of "association," then the case was referred back to the Quality Assurance Lead for resolution to adjudicate the conflict between the examiner's opinion and the algorithm output. Our approach for handling these issues is discussed in the next question below.

What happens when there is a conflict between the examiner's opinion and the output of the algorithm?

Situations in which an examiner opined "association" but the numerical results from the algorithm were insufficient to provide the empirical support, although rare, did occur causing us to consider a strategy for how to adjudicate the conflict. The first time this occurred, the Quality Assurance Lead reviewed the circumstances and provided an administrative resolution to the case. However, to ensure consistency long-term, we needed to identify a more systematic approach to handling these types of situations. This issue raised question as to the superiority of the human opinion versus the algorithm output and vice versa. On the one hand, human judgment was able to take into account many other factors that the algorithm could not (i.e., pattern type, ridge types, ridge counts, third level detail, etc.). When these conflicts occurred, many examiners were in favor of the final conclusion being reported to be consistent with the analyst's opinion. Although this would be most similar to traditional practices prior to the introduction of the algorithm, on the other hand, we could not simply ignore the output of the algorithm when it produced a result the examiner did not agree with. The challenge was to identify a solution which treated each result equally (human opinion and algorithm output) yet ensuring there was still a clear pathway for a reported conclusion, even when they were in conflict, and which did not require an *ad-hoc* administrative resolution by the Quality Assurance Lead. When deliberating over this issue, it is

interesting to note in retrospect that several concerns expressed by the examiners were eerily reminiscent of those discussed in chapter 7 when algorithms were first introduced in the scheme of clinical decision making. Many examiners expressed concern that the algorithm would become the ultimate deciding factor on what would be reported and their opinion would be devalued or worse, masked. The thought that examiners' opinions might be overridden by the algorithm or masked created even more concern and anxiety as we worked through the various ways to approach a reasonable resolution.

Ultimately, we decided the issue of conflicts between examiner opinions and algorithm outputs could be addressed by expanding the conclusion scale to add an additional type of conclusion, a "Limited Association." After taking a step back and looking at the workflow from a systems perspective, we drew a distinction between an examiner's *opinion* and a reported *conclusion*. The examiner's opinion was just that—the outcome of their subjective examination and opinion of "exclusion," "inconclusive," or "inclusion" for the given comparison (to make this distinction clear between an opinion and conclusion in the case documentation, we categorized examiner's opinion as "inclusion" and a reported conclusion as "association"). The opinion belonged to the examiner and would be documented as such no matter what the reported conclusion was. The conclusion, however, was the output of a system controlled by the laboratory's quality assurance program. The conclusion belonged to the laboratory (and by *de facto* an erroneous conclusion would be the fault of the laboratory). While examiners were accountable to their opinion, the laboratory was accountable for the reported conclusion. As such, a conclusion of association was warranted when (i) the case examiner's opinion was inclusion, (ii) the verifying examiner's opinion was inclusion, *and* (iii) the numerical output from the FRStat algorithm exceeded a pre-determined threshold value (in our case, we chose 10 as it offered the greatest balance of sensitivity and specificity based on the validation documentation). When all three conditions were met, the reported result was categorized as an Association conclusion. However, if only the first two of those conditions were met (i.e., the case examiner's and verifying examiner's opinions were inclusion but the output from the FRStat algorithm did not exceed the threshold value), then the examiner's opinion would be documented as an inclusion, but the reported result would be categorized as a Limited Association conclusion. After working through this workflow and the rationale, the examiners seemed content with the approach. Even if the examiner did not agree with the ultimate outcome, there was a sense of procedural justice. From their perspective, they were entitled to their opinion and could have the confidence that nothing, not the management or the algorithm, would require them to change it or risk having it masked. From the management perspective, we could have the confidence that reported conclusions, which we viewed ourselves as being ultimately accountable for, conformed to the quality controls we had put in place. Further, approaching it in this way ensured that these issues would be adjudicated in a consistent and systematic way.

What happens if the examiner has impressions that bear conditions which exceed the limits of the algorithm (e.g., too few features, simultaneous impressions, composite impressions, etc.)?

A major limitation of the FRStat is that it required a minimum of five features to be annotated and required all features to be documented on a single impression represented by a single image. As a result, comparisons that relied heavily on level 3 detail and had fewer than five

features available were not able to be evaluated by the FRStat. Likewise, comparisons involving simultaneous impressions or requiring multiple impressions of the same source of friction skin (i.e., composite impressions) also were not able to be evaluated by the FRStat. In these situations, we had to offer an exception to the use of the algorithm, but we also wanted to ensure that stakeholders understood the algorithm was not used given its limited scope of applicability. Ultimately, we approached this by establishing a sub-category of the Association conclusion. Rather than reporting the conclusion as a "Limited Association" (as would be done if the FRStat was able to be run and resulted in a numerical output that did not achieve the required threshold), we decided to report the conclusion as an "Association" but include language in the report that made clear statistical support was not available (due to limitations of the software application).

How should the results be reported?

There are a number of different ways to approach this issue and there were different perspectives to reconcile. The first issue was whether to include the numerical results in the report or not. If so, then the next question was where in the report should they occur (e.g., body of the report or appendix) and how the numerical results should be conveyed. For the first point, on the one hand, we could keep the existing report wording that we had transitioned to in 2015 (or derivation thereto) and simply document the numerical results in our case notes (not in the report itself). This was a viable solution since the report wording was a verbal expression of probabilistic concepts, and the numerical results simply provided empirical support to the statement. Little, if any, changes would be necessary to the formatting of the reports issued if we did not report the numerical results. The benefits of not reporting the numerical results are that the examiners had gained comfort and familiarity with this revised reporting over the last two years, the wording was defensible and easy to testify to, and the wording seemed to be acceptable by other stakeholders without adding confusion, which numerical results tend to do. Further, a major cause of hesitation from the examiners on including the numerical results upfront in the report was that the values produced by the FRStat underrepresented the "true" strength of the correspondence between the impressions compared to what they believed it should be. Although it was impossible to decipher what the "true" strength of the correspondence between the impressions should be, their concerns were reasonable. Not only was the FRStat algorithm unable to account for all the other discriminating attributes that examiners could, but the dimensionality of the information it was able to take into account was reduced even further to a single summary statistic. This led to concerns from the examiners that the findings would be grossly under-valued by other stakeholders and it would lead to confusion as to whether the results were strong enough to form actionable inferences by investigators and other lay judicial actors. The downside of not reporting the numerical results, however, is that a common criticism of the field is that not all findings from friction ridge impressions are the same—some bear stronger support than others for a given proposition—thus using the same language on the report to account for all impressions when the numerical results are available would mask the significance of the findings. To those stakeholders sharing this view, this seemed to be two steps forward but one step backwards. During our early deliberations, a majority of practitioners shared the viewpoint that the numerical results should be documented in the case notes but not included in the report. A minority of practitioners, as did a majority of other stakeholders, shared the viewpoint that the numerical results should be documented in the report itself so that it is clear to all recipients what the results from the algorithm

were. Ultimately, we decided that our long-term vision was that numerical results *should* be included in reports to ensure stakeholders had a clear understanding of the empirical support for each comparison in the case. With this in mind, however uncomfortable it might feel at the beginning, we felt it was important to proceed with including the numerical results in the report to be consistent with that vision. Concerns related to potential confusion surrounding interpretation of the numerical results would be something we could address in the form by which we conveyed those results.

Having established that the results would be included in the report, our next issue to address was where in the report should they occur and how the numerical results should be conveyed. For the first part, the two options we considered were in the body of the report or an appendix. Including the numerical values in the body of the report was considered to be the most upfront and direct way of representing the results. Perspectives on this point mirrored those discussed earlier related to whether the numerical values should be included in the report or not. Generally speaking, those examiners who were opposed to including the values in the report favored including them in an appendix as an alternative. Others favored including them up front in the body of the report. In their view, this was most consistent with our long-term vision, and including them in the appendix was more appropriate as a short-term transition. Rather than changing again, these proponents thought we ought to go ahead and make the change that we were ultimately headed toward sooner than later. Ultimately, we decided to include the numerical results in the body of the report rather than creating an appendix solely for the purpose of listing the numerical values.

For the second part of this issue, we decided that there are multiple ways in which the numerical values could be conveyed in the body of the report. The four options we considered were: (i) convey only the denominator value of the ratio produced by the FRStat, (ii) convey the numerator and denominator values of the ratio produced by the FRStat separately, (iii) convey the full ratio produced by the FRStat as a single value, or (iv) convey the full ratio produced by the FRStat as a single value and include the threshold value used to qualify the conclusion as an association.

For the first option, this was most similar to the framework of our existing report wording. Our first sentence in our current report wording would simply state the two impressions had corresponding ridge detail. The second statement in our current report wording would provide an indication of the significance of that correspondence in terms of the probability of observing a GSS value greater than the value observed in the case at hand. The benefit of this approach is that it was most consistent with our existing reporting format and examiners felt it would be relatively easy to interpret since it mirrored the format of a random match probability that had been common in DNA. The downside to this approach, however, was that the numerator value was masked. Stakeholders expressed the view that it is important to understand *both* the value of the numerator and value of the denominator to properly interpret the significance of the findings—for example, a higher numerator value indicates stronger results than a lower numerator value when paired with the same denominator.

The second option considered was to convey the numerator and denominator values of the ratio produced by the FRStat separately. On the one hand, this would accommodate the concern

that both values needed to be reported. On the other hand, it was difficult to come up with proposed wording to convey these two values separately while still being succinct and to the point. Further, having two values was thought to be more confusing than providing a single numerical value. Based on these concerns, the second option was not viewed as a favorable path forward.

The third option considered seemed to address the concerns from the first two. For the third option, the results could be conveyed as the full ratio produced by the FRStat as a single value. This would provide a single quantitative value for each comparison result, could be easily included in succinct report wording, and accounted for both the numerator and denominator values. On the surface, this seemed to be the best approach. However, concerns remained over whether stakeholders would be able to properly understand the meaning behind the numerical result. Recipients of the report would clearly see that higher values indicated stronger support, but there were questions as to what a particular value meant. In other words, there were questions around what values were sufficient to be "actionable"? When these questions were raised, it opened the door for another avenue of discussion surrounding the complexities with statistical reporting. Ultimately, there is no clear answer as to whether a particular result is strong enough to warrant "action." This is a decision that must be made by the recipient of the report based on the circumstances, information available, actions contemplated, and potential consequences. Initial reaction by some to this viewpoint was that it was outside the purview of the examiner to decide if a particular action was warranted or not. After several debates on this issue, which, in retrospect were quite philosophical, there was consensus that the responsibility for making a decision to take a particular action rightfully belonged to the recipient of the report and was not within the purview of the examiner. However, there were strong feelings that it was the responsibility of the examiner to provide context for the recipient of the report to make those decisions. This viewpoint was compelling.

The fourth option considered seemed to address all the concerns raised related to the first three. This option expanded on the third by adding a "technical note" to the end of the report which included the threshold to which the ratio produced by the FRStat could be compared to qualify it as a positive "association" according to our protocols. The term "positive" was favored by many and included in the technical note to further clarify that the label "association" was a positive result as opposed to a negative or neutral result, which was also common jargon amongst many lay stakeholders. Overall, with this approach the numerical values were provided up front in the report, but the technical note provided additional context to the underlying meaning of the numerical value that recipients of the reports could use as a reference. Effectively, this approach was a compromise between the competing viewpoints over the broader issue of reporting numerical values. Those who were proponents to statistical reporting were satisfied that the numerical values for each comparison were included in the body of the report. Those who were initially averse to statistical reporting and concerned over the confusion that could arise from reporting in that manner seemed satisfied that the report included some context to the numerical result, thus enabling a dual interpretation scheme. On the one hand, the results could be interpreted on a continuous scale based on the numerical value specific to each comparison. On the other hand, for those who preferred a more traditional interpretation and simplified means of summarizing the results, they could broadly label the result as an "association." Ultimately, we understood that not everyone would be perfectly satisfied with this approach, but decided it was most effective at balancing the different viewpoints.

Having decided that the fourth option for conveying the results on the report was most preferred, our final challenge was to ensure we addressed the concerns raised by the examiners that the values produced by the FRStat underrepresented the "true" strength of the findings compared to what they believed it should be. Admittedly, we struggled with how best to approach this. From a management viewpoint, we were concerned with how to alleviate their concerns without the availability of other tools with more discriminating power resulting in stronger values. Further complicating this was the release of the PCAST report which stated their perspective very clearly: "[s]tatements claiming or implying greater certainty than demonstrated by empirical evidence are scientifically invalid" [7]. Although it could be argued that this statement was not intended to be taken literally in the sense that all forensic findings required statistical support for the result to be considered valid, we wanted to avoid taking too many liberties in our interpretation given the explicit nature of the statement. When deliberating on this issue, we found that including the threshold value in the technical note for the FRStat result to be compared against to qualify as a "positive association" did help alleviate some of the examiners' concerns. However, there was interest in taking it a step further. Ultimately, we decided to expand the technical note to be explicit about the limitations to the software and the fact that it was unable to account for all of the discriminating attributes that are taken into account by an expert; thus, the reported results indicate a conservative estimate for the statistical strength of the association. This approach seemed to accommodate the examiners' concerns about potential under-valuations of the findings based on the numerical result while at the same time reinforcing the fact that they too had completed their subjective examination—the FRStat was a means of providing a numerical output and empirical support to the examination; it was not the sole basis of the examination.

What happens if a court takes issue with the algorithm and it is not admitted?

This was a concern shared by many examiners at the outset—if the algorithm were to be found inadmissible, would it affect the admissibility of the evidence overall? In short, the answer is "it depends." As we see from the discussion in chapter 7, it is likely to depend on how the algorithm is implemented and the extent to which the algorithm might have impacted the overall interpretation of the evidence. Given the significance of the initial reactions, we felt it was necessary to take a step back and first address the concerns over whether we believed the algorithm would be admissible or not. Clearly, our intent was not to introduce a method without regard for admissibility requirements, so we considered this to be an unlikely, but possible risk. As we reviewed in chapter 7, there are two primary issues to consider from the legal perspective: (1) admissibility under the Constitution, and (2) admissibility under existing standards.

For the first issue (admissibility under the Constitution), first and foremost, we were welcoming to any legal party seeking to exercise their rights to Due Process and Confrontation and were prepared to support such requests however we could. For example, details about the development, design, and conceptual operations of the algorithm, along with its validation and limitations were published and available [50], and the software application containing the algorithm itself was accessible, if needed. For the second issue (admissibility under existing standards), we were primarily focused on addressing admissibility requirements under Federal Rule 702 [177] and the factors outlined by the Daubert standard [173] as those captured the

requirements we would most likely face in both military courts, federal courts, and, in one way or another, most state and local jurisdictions as well. In our view, we would have no issue meeting the majority, if not all, of the requirements under Rule 702 and Daubert. Our documentation related to the development and conceptual design, as well as validation methods, and results were thorough and representative to casework; the validation materials were subject to peer review (and published shortly after implementation); and our protocols governing its use allowed us to easily demonstrate proper application of the algorithm. The only area where there was ambiguity, from our viewpoint, was related to the general acceptance prong of Daubert—whether the theory or technique "has attracted widespread acceptance within a relevant scientific community" [173]. The ambiguity was centered around how a court might characterize "widespread acceptance" and "relevant scientific community." If the "relevant scientific community" is characterized as the general scientific and statistical communities, we were confident the method employed by the algorithm would suffice. The underlying statistical techniques employed and methods for conducting the calculations were based on well-established methods. There was nothing groundbreaking from a scientific perspective in terms of how the algorithm operates. However, if the "relevant scientific community" is characterized as the friction ridge practitioner community, then we were less certain how the court might rule. In our view, it would then depend on how the court would characterize "widespread acceptance." At the time, we were the only laboratory in the United States that was moving forward with operationalizing the FRStat algorithm. Would "widespread acceptance" of a method require it be widely operationalized?—Clearly a court could not take such a rigid view, otherwise courts would never be able to admit new techniques as there is always going to be a first laboratory to operationalize it. Or, would "widespread acceptance" simply require widespread awareness of the technique and acceptance of its validity from a conceptual standpoint? However a court might characterize those phrases, we did not want to be liberal in our assumptions and were sensitive to the mere possibility of the algorithm being subject to admissibility challenges on the basis of general acceptance, particularly if we did not ensure the algorithm and underlying technical documentation were available. Consequently, almost immediately after we implemented the algorithm, we were able to make both the algorithm and technical documentation available to law enforcement and research institutions based within the United States. This provided unfettered access to the algorithm and validation information to both academia and practitioners throughout the United States, allowing them to independently test and evaluate it on their own accord.

Ultimately, despite our best efforts to ensure the algorithm would satisfy admissibility requirements, we recognized that there was always the possibility that a court might rule against admissibility. It would be naïve to think otherwise. For this reason, and consistent with discussions in earlier questions within this section, we were sensitive to implementing the algorithm in a way that would cause the least disruption to our existing examination methodology as well as isolate the process by which examiners form their subjective opinions from algorithmic influences. By keeping those two separate, and ensuring the algorithm were run *after* the expert had formed their own opinion, we were confident that if a court were to rule against admissibility of the algorithm, we could fall back on reporting the results based on the subjective judgment of the examiner (as it was formed independent and without influence from the algorithm). With this in mind, and drawing from the taxonomy discussed in chapter 7 (i.e., [53]), our target level of implementation was Level 2. Implementing at this level ensured we were able to achieve our objective of providing empirical substantiation to analysts' opinions while at the same time

keeping traditional examination practices completely intact in the event that we needed to fall back on them if the results from the algorithm were not admitted. After working through this rationale with the examiners, we found that their initial concerns subsided related to the admissibility of the algorithm and its potential impacts to the admissibility of the friction ridge evidence overall.

8.3.3 Example Workflow with FRStat

In the section above, we discussed several key questions and considerations that affected how to approach implementation and development of the policies and procedures governing its use operationally. In this section, I provide a short summary and generalized description of an example workflow related to the use of FRStat which is conceptually similar to what we instituted in casework. It is important to note that this is a simplified description to illustrate the workflow. Several important elements related to the policies and procedures are succinctly stated and those that are less specific to the FRStat are omitted altogether for brevity).

1. The case examiner will conduct an Analysis of a mark, detecting and annotating all features available (specifically, minutiae—ridge endings, bifurcations, and dots) in terms of their location and angles. The case examiner will make a determination of whether the impression is suitable for Comparison. If so, the examiner will proceed to the next step. If not, the examiner will document the result and cease further examination.

2. The case examiner will conduct a Comparison between the features observed in the mark and features observed in an exemplar print. The examiner will document similarities observed between the two impressions using a separate layer in a digital processing software (to ensure the feature annotations documented during their Analysis in step 1 are distinct from those documented during Comparison in this step).

3. The case examiner will Evaluate the similarities and differences observed between the mark and the exemplar print and render one of the following three opinions "exclusion," "inconclusive," or "inclusion."

4. The case examiner will submit the impressions for Verification. Verification is completed independently by a separate qualified examiner. This is completed by repeating steps 1 through 3 by the verifying examiner. The documentation and opinion of the verifying examiner is saved in the case file and returned to the case examiner.

5. The case examiner will compare their opinion with that of the verifying examiner. If they are in agreement, the case examiner will proceed to step 6. If they are discordant, then the case examiner will notify the Quality Assurance Lead of a conflict and proceed with formal conflict resolution procedures (for purposes of this example workflow, formal conflict resolution procedures are not discussed). Once the conflict between the two examiners is resolved and both examiners have matching opinions, the case examiner will document the circumstances in the case file and proceed to step 6.

6. If case examiner's and verifying examiner's *opinions* are both "inclusion," the case examiner will proceed to step 7. Otherwise, the case examiner will report the results using standard reporting language for "exclusion" or "inconclusive" results.

7. The case examiner will document between 5 and 15 features eligible for entry into FRStat. Eligible features are those that the examiner annotated during Analysis (step 1) and Comparison (step 2) *and* for which the annotations overlap one another on the digital image. Features annotated that do not overlap with one another or were first annotated after exposure to the exemplar print are not eligible for input to the FRStat. Features annotated for entry into FRStat will be done on a separate layer in a digital processing software (to ensure features annotated for FRStat entry are distinct from those annotated in during Analysis (step 1) and Comparison (step 2)).

8. The case examiner will run the FRStat and evaluate the results. If the ratio value produced by the FRStat is equal to or greater than 10, the examiner will proceed to report a *conclusion* of "Association." If the ratio value produced by the FRStat is between a value of 1 and 10, the examiner will proceed to Step 8a. If the ratio produced by the FRStat is less than 1, the examiner shall notify the Quality Assurance Lead of the conflict between their opinion and the FRStat output and proceed to step 8b.

    a. In situations when the ratio produced by the FRStat is between 1 and 10, the case examiner can either: (i) proceed to report a "Limited Association" conclusion, or (ii) request permission to input additional features into FRStat that were originally detected by the case examiner but deemed "ineligible." If option (ii) is selected, the Quality Assurance Lead will review the results of the Verification or coordinate for an independent examination to be performed. Features deemed "eligible" for entry to the FRStat by the independent examination which correspond to features originally detected by the case examiner can then be permitted for entry into FRStat (the case examiner will document this permission in the case file). If the ratio produced by the FRStat is still between 1 and 10, the case examiner shall proceed to report a "Limited Association" conclusion.

    b. In situations when the ratio produced by the FRStat is less than 1, the Quality Assurance Lead will review the circumstances and provide an administrative resolution to the issue. If it is determined that there are faulty annotations on the input images or the images exceeded the limitations of the software such that the output results are invalid, the issue will be corrected and re-run (if possible) or a conclusion will be reported as an "Association—Lacking Statistical Support." In circumstances where the input images did not appear to exceed the limitations of the software, then the conclusion will be reported as "inconclusive."

8.3.4 Example Report Phrasing with FRStat Results

Results of Examination

1. Analysis of Exhibit 1 revealed one friction ridge impression suitable for comparison.

2. The standards bearing the name [NAME] submitted with this request were used for comparison.

3. The friction ridge impression on Exhibit 1 and the standards bearing the name [NAME] have corresponding ridge detail. The probability of observing this amount of correspondence is approximately ## times greater when impressions are made by the same source rather than by different sources.

Technical Note

The statistical calculations in this report were generated using FRStat software. Results equal to or greater than 10 indicate a positive association between two impressions. Correspondence is measured with respect to the spatial relationships and angles of the annotated ridge details and reflected as a similarity statistic. Uncertainty of measurement is calculated using an iterative random sampling scheme for the annotated details. The reported result is the lower bound of the 99% confidence interval. FRStat is not designed to evaluate all aspects of the impressions, such as pattern type, feature type, intervening ridge counts, and other details considered by an examiner; thus, the reported results indicate a conservative estimate for the statistical strength of an association between two impressions.

8.4 Litigation (case study)

The first case involving the use of the FRStat operationally that went to trial was *United States v. SSG Luis Burgo-Castro* in 2018. The case was a military court-martial at Fort Huachuca, Arizona, involving drug charges. When the friction ridge findings were presented at trial, the numerical results of the FRStat were presented. The parties accepted the results from the FRStat algorithm without substantive discussion or challenge and the FRStat was admitted.

The first case involving the use of the FRStat in civilian court was *Ex Parte Areli Escobar* in 2019. This case involved a post-conviction writ of Habeas Corpus from Cause No. D-1-DC-09-301250 in the 167th District Court Travis County, Texas filed pursuant to the provisions of Texas Code of Criminal Procedure Article 11.071, involving a murder conviction and death penalty sentence from 2011. In short, under Article 11.071, the Applicant claimed that new scientific evidence was now available that was not available at the time of the original trial, which indicates the friction ridge evidence admitted at the trial was unreliable and, had that new scientific evidence been presented at trial, the Applicant would not have been convicted. The claim was predicated, in part, on concerns related to the categorical expression of friction ridge examination results and the need to express them probabilistically. In November 2018, I was asked by the State in this case to examine the fingermark impressions and apply the FRStat algorithm to allow a

probabilistic expression of the findings. The results from the FRStat algorithm were then subsequently challenged. The nature of the challenge is reminiscent of the discussion presented in chapter 7; however, the context of the challenge was peculiar. In the sections that follow, I present a brief summary of the case background, details related to the friction ridge evidence generally, and description of the arguments related to the application of the FRStat and the specific concerns raised by the use of the algorithm. The details presented in this section are not exhaustive and intended for illustrative purposes. Although this case represents the first known instance in which a court has had to grapple with the legal issues surrounding the use of an algorithm to provide a probabilistic result related to friction ridge evidence in the United States, the circumstances that led to the use of the FRStat were idiosyncratic.

8.4.1 Case Background

On May 13, 2011, Areli Carbajal Escobar was found guilty of the May 31, 2009, aggravated sexual assault and murder of a 17-year old female, Bianca Maldonado, in her apartment in Austin, Texas. In the early morning hours of May 31, 2009, Maldonado was found deceased in her apartment next to her one-year-old son, who was found unresponsive, but alive. Maldonado suffered from several injuries related to aggravated sexual assault and over 40 stab wounds throughout her body, including her face and chest. Following an initial investigation, Escobar, who lived in the same apartment complex, was arrested and charged with the crime. According to court records, there was no information presented suggesting Escobar and Maldonado knew each other prior to the incident.

At trial, the two items of physical evidence that were presented which most directly linked Escobar to the crime included DNA and fingerprint evidence. The DNA analyst testified that Maldonado's DNA could not be excluded as the source of the DNA recovered from Escobar's shoes, with a random match probability of approximately 1 in 10 quadrillion. The fingerprint analyst testified that Escobar's fingerprint was found on a lotion bottle recovered next to Maldonado's body inside the apartment (latent no. 132.09). On May 20, 2011, Escobar was sentenced to death for the crime. Upon subsequent appeal, the court affirmed the applicant's conviction and sentence [216].

In May 2013, Escobar filed his initial post-conviction application for writ of habeas corpus. The Court denied relief [217]. On February 10, 2017, Escobar filed a subsequent habeas application [218]. In this subsequent application [218], Escobar alleged that the factual or legal basis for his claims was unavailable on the date he filed the previous application and argued that he is entitled to relief under Article 11.073 [219]. Article 11.073 provides a new legal basis for habeas relief in cases where an applicant can show by a preponderance of the evidence that he would not have been convicted if the newly available scientific evidence had been presented at trial. Article 11.073 applies to relevant scientific evidence that was not available to be offered by the defendant at trial, or that contradicts scientific evidence relied on by the State at trial. Among the concerns noted, Escobar specifically challenged the reliability of the DNA evidence and fingerprint evidence, the two items of physical evidence most directly linked him to the crime. The concerns over the DNA evidence relate to problems discovered at the Austin Police Department crime laboratory during an audit by the Texas Forensic Science Commission [220].

The audit discovered that some staff members were not properly trained and that incorrect methods were used to examine DNA samples, resulting in the closure of the DNA unit in June 2016 [220]. The concerns over the fingermark evidence relate to claims that the examiner relied on scientifically invalid methods to link the partial fingermark found on the lotion bottle two feet away from Maldonado's body (latent no. 132.09). On October 18, 2017, the Court accepted the claims in the application *prima facie* and remanded the case to the convicting court for consideration on the merits [221]. In the next section, specific details related to the fingermark evidence are summarized.

8.4.2 Fingermark Evidence

One of the key pieces of physical evidence linking Escobar to the crime was a partial bloody fingermark found on a lotion bottle approximately two feet away from Maldonado's body in her apartment (latent no. 132.09). On his subsequent habeas application, Escobar alleged that "[n]ew scientific evidence not available at the time of the trial demonstrates that the State relied on scientifically invalid fingerprint comparison evidence purporting to link Mr. Escobar to the inside of the apartment where Bianca Maldonado's body was found. When coupled with the other scientifically invalid evidence presented at Mr. Escobar's trial, this new scientific evidence would have undermined the reliability of Mr. Escobar's conviction and death sentence by a preponderance of all remaining valid and available evidence" [218]. In his claim, Escobar alleged that the original fingerprint examination and reported results did not conform to newly published standards and guidelines demonstrating that the fingerprint evidence was either highly impeachable or should not have been admitted at all. To support his arguments, Escobar relied heavily on the affidavit of his expert, Dr. Simon Cole, and findings from a report by the Expert Working Group on Human Factors in Latent Print Analysis, jointly convened by the National Institute of Standards on Science and Technology (NIST) and the National Institute of Justice (NIJ), entitled *Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach*; the announcement by the USACIL that they will no longer offer reports of "individualization" or "identification," but rather offer reports framed in probabilistic terms; and the publication of a report by the U.S. President's Council of Advisors on Science and Technology (PCAST) entitled *Forensic Science in Criminal Courts: Ensuring the Scientific Validity of Feature-Comparison Methods*, which included friction ridge analysis. By way of summary, Escobar claimed: (i) the low quality of the fingermark raises concern over whether it should have been identified at all or, at best, only identified if the examination was done in accordance with protocols pertaining to "complex" examinations by the Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST), e.g., enhanced documentation and blind verification—none of which were followed; (ii) the fingermark examination was impacted by implicit and explicit biases and proper protocols were not followed to mitigate its influence on the examination causing question over the reliability of the results, and (iii) the examiner's conclusion of a "match" exceeds the limits of what can be scientifically supported and is not a valid means of presenting the evidence to the trier of fact [218].

To expand upon the second issue, Escobar's claim was predicated on the fact that the original friction ridge examiner initially reported on September 29, 2009 that Escobar had been excluded as a possible source of the fingermark found on the lotion bottle (latent no. 132.09). On

Thursday May 5, 2011, just as the trial was beginning, the Assistant District Attorney (ADA) contacted the examiner and asked her to compare the unidentified prints in the case to the known prints of the victim's mother and younger sister. On Monday, May 9, 2011, the friction ridge examiner notified the ADA that she had matched the fingermark on the lotion bottle (latent no. 132.09) to the left ring finger of Escobar. During testimony, the examiner stated she had missed an "orientation clue" in her initial review of the fingermark leading her to believe the mark came from the right hand. When questioned, the examiner stated she was not prompted by the prosecutor to re-look at Escobar's prints; rather, she said she did it on her own accord. Escobar noted, however, that in an email exchange between the ADA and the examiner, the ADA expressed being "worried" about the outstanding mark. Consequently, Escobar claimed "there is significant evidence of implicit and express bias on the part of the testifying print examiner, [REDACTED], which went both undisclosed to the jury and uncorrected by the examination protocols in place at the APD crime lab where she worked" [218]. Escobar argued that "[the ADA] was highly motivated to produce a positive identification on the lotion bottle print. As a result, she urged [the examiner] to conduct further comparisons while the trial was already underway. To the extent that urgency was communicated, of course, the bias was explicit. ... [And,] the fact that [the examiner] allegedly took it upon herself to re-examine the latent print and affirmatively compare it again to the actual defendant in the case, Mr. Escobar, is, as Dr. Cole points out, precisely the type of implicit bias that current scientific protocols affirmatively disallow" [218].

On the third issue, Escobar relied heavily on the findings from the PCAST report [7], claiming that the results of the friction ridge examination should instead be expressed in probabilistic terms, e.g., expressed as the relative probabilities of the evidence if the two impressions derive from the same, or different, sources. Escobar goes further and references the decision by the USACIL that they will no longer offer reports of "individualization" or "identification," but rather offer reports framed in probabilistic terms as an example of the field's recognition of those concerns [218].

In response to these claims raised by Escobar, the District Attorney's office responsible for representing the State in this case contacted the undersigned in November 2018. I was requested to conduct an independent examination of the fingermark impression, in accordance with current scientific protocols, including the use of the FRStat, if applicable, as was currently operationalized at the USACIL. Shortly after receiving the materials, I completed my examination which resulted in my opinion of an association between the fingermark impression on the lotion bottle (latent no. 132.09) and the prints bearing the name Escobar. The results were verified under blind conditions by another certified fingermark examiner. Upon notification of the results from the verifying examiner, eligible features from my examination were encoded and the images were input to the FRStat. The evaluation by the FRStat produced a ratio value of approximately 150,000, which provided a measure of empirical support to the association between the fingermark impression on the lotion bottle (latent no. 132.09) and the exemplars bearing the name Escobar. The results of my examination, including the numerical values produced by the FRStat, were delivered in a report dated December 5, 2018 [222] (disclosed by the State to Escobar on December 7, 2018).

On December 10, 2018, shortly after receiving the fingermark examination report and results of the FRStat evaluation, Escobar was granted a continuance to delay the hearing on the fingermark evidence originally scheduled for December 17 and 18, 2018. Escobar retained Prof.

Cedric Neumann to challenge the results of the FRStat. Following subsequent requests for discovery and continuance, the Court rescheduled the hearing to March 18 and 19, 2019. On March 5, 2019, just before the hearing was scheduled to begin, Escobar requested an additional continuance to provide more time to prepare a challenge to the FRStat. This additional request was denied. On March 12, 2019, Escobar filed a *Motion to Exclude the Expert Testimony of Henry Swofford Related to the FRStat Evidence Pursuant to Daubert v. Merrill Dow Pharmaceuticals* [223]. The hearing concerning the fingermark evidence took place on March 18 and 19, 2019; however, the Court did not have sufficient time to address the issues related to the FRStat, and therefore a subsequent hearing occurred on June 17 and 18, 2019. In the next section, specific details related to the litigation of the FRStat are summarized.

8.4.3 FRStat

The use of the FRStat in this case immediately became a point of contention between the parties, resulting in a number of subsequent court filings related to the admissibility of the algorithm and which party was to bear the burden of demonstrating its admissibility. Given the peculiar nature of this case and circumstances that led to the use of the FRStat, and to enable readers to compare the issues raised related to the FRStat algorithm to the discussion in chapter 7 (specifically, part IV "Algorithms and the American Legal System"), lengthy quotations from court filings are provided in this section to ensure each party's arguments and positions related to the FRStat are not distorted.

Immediately after the results of the fingermark examination and numerical values produced by the FRStat were provided in the fingermark examination report dated December 5, 2018 [222], Escobar sought disclosure of details related to the algorithm, source-code, and validation materials and information, claiming "the FRStat program is essentially a 'black box'" and "secret method" for which the State bears the burden of establishing admissibility under the Daubert standard [224, 225]. On January 1, 2019, in response to their December 18, 2018, motion for discovery, and again on March 3, 2019, in response to their February 25, 2019, motion for discovery, the State disclosed much of the information requested and noted that the information concerning validation had already been published in a peer reviewed journal (e.g., disclosure included details related to the design, development, and underlying operation of the algorithm, including operating characteristics (e.g., user manual), key performance characteristics (e.g., validation report), procedures related to interpretation of the results, and general strengths and limitations of the system, as well as raw data related to the validation of the algorithm and data used as inputs for the specific case at hand). The algorithm itself had already been provided to Prof. Neumann in June 2016, stemming from his direct request through the University of South Dakota[31]. The Court chose not to compel discovery of the source-code, however, as it had not been made public by the U.S. Army. In response, on March 5, 2019, Escobar filed a *Renewed Motion to Continue Hearing Regarding Admissibility of FRStat Evidence* [226]. In this motion, Escobar alleged that, based on initial review of the information provided and through his expert, Prof. Neumann, "the scientific principle underlying the FRStat model is not valid or verifiable," and "the FRStat method is not

[31] Prof. Neumann assisted M. Ausdemore, a doctoral candidate at South Dakota University, with her contributions to the development and validation of the FRStat while she was employed as a research associate at the USACIL as part of an academic partnership between the USACIL and South Dakota University.

commonly accepted by the relevant scientific community" [226]. Escobar argued that "[w]ithout adequate time to [prepare] a *Daubert* challenge, Mr. Escobar will be deprived of his right to due process in this post-conviction challenge to forensic evidence used to secure his conviction and death sentence" [226]. In response, on March 6, 2019, the State argued that under article 11.073 of the Texas Code of Criminal Procedure, the State does not shoulder the burden for the admissibility of the FRStat. Instead, the sole reason FRStat was used was in response to Escobar's own claims that the fingermark evidence should be presented probabilistically [227]. Specifically, in their *Response in Opposition to Applicant's Renewed Motion to Continue Hearing Regarding Admissibility of FRStat Evidence* filed on March 6, 2019, the State argues [227]:

> *Applicant claims, pursuant to article 11.073 of the Texas Code of Criminal Procedure, that, had new scientific evidence related to friction ridge analysis that is currently available but not available at the time of his trial been presented, he would not have been convicted.*
>
> *The new scientific evidence upon which Applicant relies includes evidence that the Defense Forensic Science Center (formerly known as the U.S. Army Criminal Investigation Laboratory Latent Print Branch) reports its examination conclusions in probabilistic terms. Through his expert, Dr. Simon Cole, Applicant asserts that, "because it is scientifically and logically implausible that an examiner could know the source of an impression[,] [e]xaminers should more properly address the relative probabilities of the evidence if the two impressions derive from the same, or different, source." Applicant, through Dr. Cole, specifically notes that this scientific methodology is currently used by the U.S. Army Criminal Investigation Laboratory. The practical application of this methodology is known as FRStat.*
>
> *The State, in an effort to respond to Applicant's "new scientific evidence" argument, hired latent fingerprint expert Henry Swofford, who, as the former chief of the Latent Print Branch of the U.S. Army Criminal Investigation Laboratory, developed the FRStat application advocated by the Applicant in his Application. The State asked Swofford to examine the latent print at issue in this case and, using FRStat, to report on the relative probability that the latent print and Escobar's known exemplar are derived from the same source. On December 7, 2018, the State forwarded Mr. Swofford's completed report to Applicant's attorneys.*
>
> *Applicant thereafter requested and was granted a continuance of the December 17 and 18 hearing on this matter. He again moved for continuance of the hearing that was reset for February 18 and 19, 2019. That motion was granted[,] and the matter was reset for March [18] and [19], 2019. Applicant now claims that he needs yet another continuance to properly and thoroughly address FRStat. In essence, he asserts that his counsel and his expert are still unprepared to address, at the March 18 and 19, 2019[,] hearing, the new scientific evidence that he himself initially relied upon to challenge the trial evidence in this case. Applicant incorrectly asserts, ... that his first introduction to the possibility that FRStat could be used in this case was on December 7, 2018. As outlined [above], it was actually Applicant, in his Application filed on February 10, 2017, who first championed FRStat as the method by which a comparison between a known print and a latent print can*

be expressed in probabilistic terms. Simply put, Applicant's [claim] is necessarily predicated upon familiarity with FRStat. Applicant has not had, as he claims in his motion, mere weeks to prepare to address FRStat. He has had over two years.

In a bizarre turn of events, likely brought on by the fact that the FRStat results reported by Mr. Swofford are extremely damning to him, Applicant is now engaged in a full-on battle against his own science. Applicant has retained an expert, Dr. Cedric Neumann, to challenge the only application he once touted as capable of putting out a probabilistic result from a comparison between a latent print and a known print. Applicant has seemingly developed amnesia regarding the fact that it was he, and not the State, who championed the FRStat model to begin with and is now, instead, asserting that FRStat is scientifically invalid and unreliable. Applicant claims he needs an additional month-long continuance to mount this challenge against his own science. This, despite the fact that he himself first proposed FRSat as the tool by which to practically apply his science at the time he filed his application in February of 2017. If FRStat is, indeed, scientifically unreliable and invalid, shouldn't Applicant have been fully aware of this before advocating it to the Court as the tool by which to practically apply its new science? Unless the Applicant is explicitly abandoning that portion of his claim where he asserts that the results of a comparison between a latent and a known print should be expressed in terms of probabilities, then Applicant's objection to FRStat should be overruled and his continuance denied.

In his final analysis, the Applicant has conflated the issues here and, in doing so, is trying to place, on the State, a burden that the State does not have. It is Applicant's burden, under article 11.073, to demonstrate the existence of new relevant scientific evidence that is currently available and that would be admissible at trial. Moreover, it is Applicant's burden to show that had any new, relevant, currently available scientific evidence been presented at trial, the outcome would have been different and he would not have been convicted. That new scientific evidence, according to the Application filed on February 10, 2017, is, in part, the notion that results of comparisons between latent and known prints are best expressed in probabilistic terms and the only tool explicitly offered in the Application as the way to practically apply this science is FRStat. Just because the State chose to do what Applicant did not—by applying the FRStat to the evidence in this case to see whether a result would be generated that "contradicts" the scientific evidence at trial—does not mean that the burden is now shifted to the State to defend the new science Applicant purports is available, admissible and would have changed the outcome of Applicant's trial.

It is critical to note that, in offering the FRStat results obtained by Mr. Swofford, the State is not conceding nor taking the position that results of comparisons between latent and known prints are best expressed in probabilistic terms. Likewise, the State is not taking the position that FRStat or any other model or tool meant as a practical application of Applicant's new science would satisfy a Daubert challenge. It is not the State's burden to satisfy a Daubert challenge as to either FRStat or the underlying methodology that gave rise to the need for FRStat. FRStat is Applicant's science. In offering the results from the application of the evidence in this case to FRStat, the State is merely demonstrating that,

upon practical application of Applicant's new science to the evidence in this case using the sole tool to which Applicant has pointed, the outcome of his trial would have been no different.

Again, the new scientific evidence, according to the Application filed on February 10, 2017, is, in part, that results of comparisons between latent and known prints are best expressed in probabilistic terms. The only tool implicitly offered in the Application as the way to practically apply this science is FRStat. If FRStat is flawed and unreliable as Applicant claims, then Applicant, who has offered only FRStat as the model or tool by which his new science can be practically applied, has failed to satisfy even the first prong of article 11.073, which requires the Applicant to demonstrate there is relevant new science that is currently available. Absent a practical application of Applicant's new science, the new science cannot be said to be currently available.

Moreover, it should not escape the Court's notice that Applicant's request for a continuance to determine the validity of FRStat is accompanied by a supporting affidavit from their expert, Dr. Cedric Neumann, in which he has clearly already formed his conclusions regarding FRStat's validity and has provided page after page of testimony justifying said conclusions. He appears to need no extra time to address the validity of the method. In fact, the State notes that Dr. Neumann's two primary areas of research at South Dakota State University, as stated in his CV, is "the quantification of the weight of fingerprint evidence (development of a statistical model and reporting standards)." Neumann is clearly already well-versed regarding FRStat and models/tools similar to it.

(…)

There is no more legitimate way for the Court to decide the ultimate question of whether the jury would have convicted using the newly available scientific methodology than to actually apply the methodology that the Applicant espouses as reliable to the evidence in this case. The State's sole motivation in retaining Mr. Swofford was to apply to the evidence the science that Applicant claimed would entitle him to relief. It should come as no surprise to the Applicant that the State would seek to have the methodology that Applicant himself touts as "best practices" applied to the evidence in this case, especially in light of the fact that Applicant apparently has, as of this date and for no apparent reason, not sought to do so himself. In the final analysis, Applicant should need no continuance to address the scientific methodology upon which he has so firmly staked his claim.

It was less than a week later, on March 12, 2019, when Escobar filed the *Motion to Exclude the Expert Testimony of Henry Swofford Related to FRStat Evidence Pursuant to Daubert v. Merrill Dow Pharmaceuticals* [223]. In his motion, Escobar requested for the testimony related to the FRStat be excluded or grant a *Daubert* hearing on the admissibility of the proposed testimony. In this motion, Escobar argued [223]:

The State asserts it bears no burden to establish the admissibility under Daubert of the scientific evidence it seeks to introduce because, according to the State, FRStat is the science Mr. Escobar relied on to bring a new science claim under 11.073. This claim is

belied both by a review of Mr. Escobar's 11.073 claim regarding fingerprint evidence in his Subsequent Application and a review of the chronology of when FRStat first became available.

As noted [earlier], in his Subsequent Application, Mr. Escobar argued that the new science forming the basis of his 11.073 claim is the development in understanding of the unreliability of latent print analysis, specifically the "match" testimony offered at Mr. Escobar's trial. ... Mr. Escobar supported his claim with an affidavit from Dr. Simon Cole explaining that the "match/no-match" testimony offered at Mr. Escobar's trial is scientifically unsound and improperly purports to identify the true source of a latent print to the exclusion of all others. Indeed, the [Court of Criminal Appeals] authorized the claim based on Dr. Cole's expert affidavit and remanded the case for further factual development as set forth in Mr. Escobar's Subsequent Application. At no point did Mr. Escobar argue that FRStat, or any other probabilistic model, should have been used at his trial or in post-conviction proceedings.

As evidence that Mr. Escobar, not the State, is the proponent of FRStat, the State cites the discussion of probabilistic evidence in Dr. Cole's affidavit. To the extent that the Subsequent Application included discussion of the developing movement toward probabilistic evidence, it was to illustrate the growing understanding in the field that the type of latent print evidence introduced in Mr. Escobar's trial is scientifically invalid and misleading, and that probabilistic models have the potential for generating more reliable evidence than subjective match/no-match determinations. It in no way constitutes an endorsement of any specific probabilistic model—in particular an untested and unvetted one such as FRStat—as scientifically reliable implementation of probabilistic theory.

As further evidence that Mr. Escobar's new science claim is premised upon FRStat, the State asse[r]ts that Mr. Escobar, "through Dr. Cole, specifically notes that this scientific methodology is currently used by the U.S. Army Criminal Investigation Laboratory" and that "[t]he practical application of this scientific methodology is known as FRStat." This assertion misrepresents both the arguments in Mr. Escobar's Subsequent Application and the timeline of when FRStat became available. In the Subsequent Application, Dr. Cole listed examples of "important new developments [that] have occurred in the field of friction ridge analysis since [Mr. Esobar's trial] May 2011"; within this context, Dr. Cole stated that the "Defense Forensic Science Center [announced] that they will no longer offer reports of "individualization" or "identification," but rather offer reports framed in probabilistic terms." This is a reference to the U.S. Army's announcement in 2015 that, in response to the growing recognition in the scientific community that forensic laboratories should refrain from reporting latent print analysis using conclusive "identification" language that purports to attribute a single source latent print to the exclusion of all others, it would cease employing such language in its reports. It was not until March 9, 2017, after Escobar had filed his Subsequent Application, that the U.S. Army announced the implementation of FRStat in order to report friction ridge skin comparisons in terms of statistical probabilities.

Thus, the State, not Mr. Escobar, is the proponent of Mr. Swofford's opinion testimony related to FRStat. As such, the State bears the burden of establishing the admissibility of the evidence by clear and convincing evidence. Moreover, by stating that it has not taken the position that FRStat would satisfy a Daubert challenge, the State effectively concedes that it has not met its burden of establishing the scientific validity and reliability of the evidence it proffers.

Escobar continued, through his expert, Prof. Neumann, alleging the "State's proposed testimony concerning probabilistic evidence is based upon novel and unvetted scientific evidence that does not satisfy the *Daubert* requirements for admissibility" [223]. Specifically, Escobar claimed: (1) "the ratio generated by the FRStat is not based upon sound scientific principles," (2) "FRStat inflates its ultimate result by failing to use relevant hypotheses," (3) FRStat does not account for rarity of features," (4) the FRStat misapplies statistical principles to overstate the uniqueness of characteristics in fingerprints" [223]. He argued further, that (5) "Henry Swofford is not sufficiently qualified in the fields of statistics or computer programming to offer an expert opinion regarding statistics or computer programming," (6) "FRStat has not been subjected to adequate validation studies or proper peer review and is not generally accepted in the scientific community," (7) "Discrepancy between the theoretical and observed error rates for FRStat indicates that the model and software have additional significant flaws," and (8) "It cannot be established that FRStat is properly applied to the finger joint evidence in this case" [223].

After challenging the admissibility under *Daubert* related to the validity of the FRStat methodology and means by which the results were expressed, Escobar continued and challenged the admissibility under the Constitution. Specifically, Escobar claimed [223]:

The State is seeking to introduce evidence that is based on information which the U.S. Army refuses to share with the public—specifically, source code data and validation reports—because of the U.S. Army's proprietary interest in the information. Mr. Swofford has stated that he is prohibited from disclosing such information until such time the U.S. Army makes the information public. It cannot be the case that proprietary interests trump the due process rights of a condemned man to challenge evidence which has demonstrated errors in its code. This is especially true where, as here, the U.S. Army is already making the program freely available to labs that want to test it out while they decide how best to distribute it, and on option apparently under consideration is "to release the program for free, with the code open and available."

To fully confront and put evidence derived from FRStat to the adversarial test, access to its source code and validation reports is necessary. The algorithm's underlying model reflects the theory and intended processes behind the probabilistic analysis, while the source code shows how that intended process has been put into practice. Source code could reveal that concepts not included in the underlying model are somehow being expressed in the program, or vice versa. There can be accidental mistakes, or "bugs," in the program. There can be errors in the programming code itself. Therefore, the source code is essential to properly access all potentially exculpatory evidence within the State's proffered evidence relevant to the State's claims.

(…)

Algorithms, developed by humans, are fallible. Yet they are profoundly persuasive, even when improperly designed. Source code would reveal the variables and assumptions made by the algorithm; adversarial testing of the code is critical to determine the degree of accuracy and ensure that ultimate decisions are made based upon reliable information.

As noted, the U.S. Army is reportedly in the process of determining the best way to release and distribute the software. Mr. Swofford suggests that an option "would be to release the program for free, with the code open and available." Under such circumstances, the source code should be released to Mr. Escobar so that he may have a fair and full opportunity to test the accuracy and reliability of the software. Alternatively, the State should be barred from utilizing the software to bring evidence against Mr. Escobar until such time that the U.S. Army decides to make the code publicly available (and then, of course, only if the Court were to determine that it meets the reliability criteria set forth in Daubert as noted above).

Finally, Escobar concluded his motion by challenging the admissibility on the basis of relevance, claiming "FRStat is not relevant to the issues presented in Mr. Escobar's challenge to the scientific reliability of the latent print evidence presented at trial." [223]. Specifically, Escobar claimed:

While the State's proffered FRStat evidence fails the Daubert standard that methodologies be properly applied, the evidence fails on the more primary premise of relevance, which is the first element required for the State to establish admissibility under Daubert. ... The FRStat evidence proffered by the State is not relevant to the issue presented in this case: whether Mr. Escobar's due process rights were violated when the State presented testimony that misled the jury by stating Mr. Escobar was the source of the latent print at issue, when new standards have established this "match/no-match" testimony to be scientifically invalid. FRStat is irrelevant to this analysis.

The State argues that the FRStat results show that Mr. Escobar was probably the source of the latent print, but as established by the affidavit of Cedric Neumann, the principles underlying this proposition are incorrectly applied and statistically meaningless. Thus, the State's proposed evidence fails to provide the requisite "valid scientific connection to the pertinent inquiry." As such, the evidence "frustrates rather than promotes intelligent evaluation of the facts" and, thus, would not have assisted the jury in this case in any meaningful way.

Further, even if FRStat were capable of generating scientifically valid and reliable evidence, it would not be relevant to Mr. Escobar's claim pursuant to Article 11.073 which challenges the latent print evidence at trial. As discussed above, the State mistakenly argues that Mr. Escobar's 11.073 claim is dependent upon FRStat and must fail without it. ... [T]his argument misconstrues Mr. Escobar's 11.073 claim which is, in fact, premised on the developments in understanding of the unreliability of latent print analysis. Nonetheless, the State implicitly posits the untenable proposition that if Mr. Escobar has

established that scientific developments demonstrate the unreliability of the old paradigm for latent print analysis, Mr. Escobar can only prevail on his 11.073 claim if he is able to point to a new scientific method that replaces the old, unreliable one. This would be akin to asserting, for example, that a claim based on scientific developments showing that bite mark evidence is junk science could only succeed if the applicant is able to demonstrate that there is a new scientific method to replace the debunked bite mark evidence. Clearly, this is not the standard contemplated by 11.073. Rather, as stated, Mr. Escobar's claim depends upon demonstrating that the now-known scientific unreliability of the latent print comparison evidence heard by the jury undermines confidence in the reliability of his conviction. Any new, post-trial evidence produced by FRStat, be it scientifically valid or not, is simply not relevant to this inquiry.

On March 18 and 19, 2019, the Court heard testimony related to the fingermark evidence and claims outlined by Escobar in his Subsequent Application; however, the Court did not hear testimony related to the subsequent examinations and FRStat results. A second hearing was scheduled and on June 17 and 18, 2019, the Court admitted the testimony of both Prof. Cedric Neumann and Henry Swofford related to the subsequent fingermark examination on the evidence at issue (latent no. 132.09) and results produced by the FRStat[32]. Following the hearing, the State supplemented the court record with an affidavit from Henry Swofford addressing each of Prof. Neumann's technical concerns raised concerning the FRStat [230] as well as affidavits from Dr. Simone Gittelson [231] and Dr. Karen Kafadar [232] related to their reviews of the FRStat algorithm and opinion regarding its validity and reliability. In closing, the State argued, among other things, "[t]he Court should consider the FRStat analysis and results in evaluating the Applicant's claim" [233]. Specifically [233]:

In his original Application and in his own Exhibit #87, the Applicant referenced the FRStat software developed by the U.S. Army crime lab as an example of the application of statistical principles to reporting friction ridge identifications. The Applicant then filed a challenge to the admissibility of the FRStat results. It is important to recognize that FRStat is not a substitute for the opinion of a certified examiner using the ACE-V process. FRStat is simply a means for a fact-finder to assign some sort of evidentiary weight to the examiner's opinion. The State asserts that the FRStat results are pertinent to show that, even if the most cutting-edge methods are used, the result is the same. The undisputed testimony is that FRStat is the only method of expressing results with a numeric value that is currently being used in live casework. The FRStat results are significant because they demonstrate that the Applicant cannot satisfy his burden under 11.073 of showing that, had the newly available relevant scientific evidence been presented at trial, he would not have been convicted.

FRStat, at its most fundamental level, is a mathematical expression of the standard for making a source identification decision. ... Prior to FRStat, jurors or judges did not have

[32] During testimony in June 2019, Prof. Neumann was asked about why he had not expressed any concerns regarding FRStat in scientific literature prior to this litigation despite having access to the software application and materials since June 2017. Following the June 2019 hearing, Prof. Neumann published his concerns regarding FRStat in *Law, Probability & Risk* (see [228]). A follow-up *Letter to the Editors* was published shortly thereafter in response to his claims (see [229]).

any means by which to assign any degree of confidence to the analyst's opinion. FRStat is a method by which to judge the strength of the association that forms the basis for the analyst's opinion.

FRStat is based on a method that expresses the degree of correspondence between features identified by an examiner in terms of a global similarity statistic (GSS). The program generates a GSS for the prints compared by an analyst and calculates a GSS based on the specific features used by the analyst. The program compares the similarity statistic from the evidence (GSS_t) to the similarity statistics from two distinct groups: mated prints (known to originate from the same source) and non-mated prints (known to originate from different sources). It calculates what proportion of the mated (same source) prints have a statistic less than the GSS_t and what proportion of the non-mated (different source) prints have a statistic greater than the GSS_t. It reports these scores in the form of a ratio where a result greater than one indicates stronger support and a result less than one indicates less support. A score of one would suggest that the GSS_t is equally likely in both groups of prints.

The Applicant's criticism of FRStat, presented by Dr. Neumann, is grounded in an incorrect interpretation of FRStat's purpose. It does not generate a likelihood ratio or measure the "weight of the evidence" in the strict statistical sense. It does not supplant the ACE-V methodology or tell an end user whether or not a specific individual made a specific latent print. It is used only after an analyst has made an identification and had it verified in accordance with the ACE-V method. It offers a means of showing quantitative support for the analyst's opinion to help finders of fact assess the strength of the opinion. The attached affidavits from Henry Swofford, Dr. Karen Kafadar, and Dr. Simone Gittelson demonstrate that Dr. Neumann's criticisms are not dispositive as to the theoretical underpinnings and statistical methodologies of the FRStat software. The Court should consider FRStat and the results obtained as valid and relevant to the Applicant's claims, especially given the limited purpose for which the evidence is offered.

It is undisputed that FRStat is the only method of providing statistical support for an analyst's identification of a friction ridge impression that is currently being used in live casework. Given the Applicant's assertion that friction ridge analysis results should be reported in probabilistic terms, the testimony of Henry Swofford and the results from the FRStat are relevant because they demonstrate that, even if the results are reported in probabilistic terms as the Applicant suggests, the result of the original identification of Item 132.9 to the Applicant is unchanged.

On December 31, 2020, the Court issued its *Findings of Fact and Conclusions of Law* detailing the Court's ruling on the case and recommendations to the Texas Court of Criminal Appeals [234]. In its ruling, the Court did not address the FRStat specifically. Instead, the Court addressed the ultimate issue concerning the fingermark evidence overall that was facing the Court—whether new scientific evidence was currently available that was not available at the time of the original trial (in 2011), which indicates the fingermark evidence admitted at the original trial was unreliable and, had that new scientific evidence been presented at trial, he would not have been convicted. Although the Court recognized Escobar's specific concerns raised, the Court

denied relief on the basis of the fingerprint evidence[33] concluding (quite succinctly): "The Court finds that the applicant did not demonstrate that new scientific evidence concerning friction ridge analysis significantly undermines the reliability of the latent print evidence presented at trial" [234].

Looking at this case retrospectively, there are a few key observations we can make. First, although the Court ultimately upheld the reliability of the fingermark evidence presented at trial, it noted that "there is certainly room for technological advances in the [latent print] comparison process" [234]. As such, the friction ridge community is implicitly encouraged to continue advancing toward offering a statistical basis to reported conclusions. Second, the legal challenges that were raised concerning the use of the FRStat algorithm reinforce the issues that were discussed in chapter 7 and illustrate what laboratories will need to be prepared to address when using a statistical tool. Third, probabilistic reporting is desired by litigators and, in principle, considered more scientifically defensible compared to categoric conclusions, but it is not a panacea to issues concerning forensic science reporting. We need to be careful about how statistical tools are used and how the results are reported to be clear about what the numerical values represent and do not represent.

To elaborate further on the third point, the chief concern raised by Prof. Neumann in this case was the compatibility of the results output by FRStat and the language used to report the conclusions of the fingerprint examination [228]. The language used to report the results output by the FRStat in this case [222] was consistent with the example report phrasing presented in section 8.3.4 above. While Prof. Neumann acknowledges that the FRStat was not designed to calculate a likelihood ratio based on the validation materials [50], he points out the similarity in which the results were reported with that of a likelihood ratio thereby creating ambiguity in what the numerical results actually represent and how they should be interpreted. Leveraging this ambiguity, Prof. Neumann offered several challenges to the mathematical properties of the ratio produced by the FRStat compared to those of a likelihood ratio, ultimately claiming that the language used to report the results output by the FRStat "is technically meaningless" [228]. Although we caution that the language used to report the results from a statistical assessment and the validity of the statistical assessment itself are distinct and should not be confused [229], his point is well taken—the forensic science community must be clear about the how the tool was used (i.e., as a tool for quality assurance versus a measure of the "weight of evidence") and what the numerical values represent (and do not represent) as different and mathematical constructs can have distinct properties that will impact how they should be interpreted. As such, this case

[33] Although the Court denied relief based on the latent print evidence, the Court did rule in favor of granting relief on the basis of the DNA evidence, recommending to the Texas Court of Criminal Appeals that the conviction be reversed and the applicant be provided a new trial. The Court concluded: "In light of the significant quality issues uncovered at the [Austin Police Department (APD] DNA lab, including the failure of the lab's entire quality assurance system, there can be no confidence that the lab produced valid and accurate results. ... [And,] ... [t]he Court finds that, after removing the DNA evidence presented at trial, the remaining evidence relied on by the State was questionable and circumstantial. The Court finds by a preponderance of the evidence that the outcome would have been different, especially in light of the testimony of a sitting juror that he was 'on the fence' until the DNA evidence was submitted" [234]. The Court's recommendation was then reviewed by the Texas Court of Criminal Appeals. On January 26, 2022, the Texas Court of Criminal Appeals issued its final ruling denying relief. In that ruling, the Texas Court of Criminal Appeals disagreed with the trial court's recommendation that relief be granted given the incriminating nature of the other evidence in the case and because the Applicant failed to demonstrate "on the preponderance of the evidence [he] would not have been convicted" [235].

illustrates it is not enough that these technical distinctions are only addressed in the validation materials. Instead, it illustrates the importance of also including an explicit caveat in the forensic examination report itself. Looking forward, this can be accomplished through a modification to the Technical Note presented in section 8.3.4 above, clarifying that the FRStat was used as a measure of the quality of the association (for quality assurance purposes) and that the results produced by the FRStat are not a measure of the "weight of evidence" or a likelihood ratio that can be used to inform the posterior probability of a proposition.

9 Looking Forward: Impact and Recommendations

This chapter summarizes key developments and findings from this thesis and discusses the impact and implications of the work, provides recommendations for friction ridge examination practices, and proposes areas for future research.

9.1 Impact

The work presented in this thesis has broad impact and implications—both theoretical and practical, ranging from statistics and evidence quantification to social psychology and human behavior—affecting policy, procedure, training, quality assurance, research, reporting and testimony, and litigation pertaining to the implementation and use of algorithms operationally in friction ridge examination. First, the work presented in Part I (chapters 2 through 4) provides the tools necessary to practically apply statistical measures to friction ridge impression findings and explore more objective interpretation schemes. This alone is significant as it provides a means for the friction ridge community to apply practical solutions and respond to the growing concerns from scientific and legal stakeholders regarding the lack of empirical substantiation to their opinions and quantitative measures of the quality and significance of the findings in a given case. Although other algorithms have been proposed over the years, few, if any, have been developed with implementation in mind and packaged in an accessible and user-friendly software application that can be easily integrated into typical examination workflows. However, the mere availability of the tools is not enough. Next, we must understand how to implement those tools operationally to realize their benefit. The work presented in Part II (chapters 5 through 8) provides important insights into the perspectives, sources of concern, and other sociopsychological factors affecting the implementation and use of algorithms operationally. These insights, gleaned across diverse stakeholders and domains, enable us to explore key challenges to the adoption of algorithmic tools and offer strategic approaches to implementation that will lower the barriers to adoption and increase stakeholder acceptance. The anecdotal reflections discussed in chapter 8 offer first-hand accounts of the journey that was undertaken to implement the FRStat algorithm operationally in a forensic laboratory in the United States and how it was handled by judicial actors during litigation—the first time this has ever occurred related to friction ridge impression evidence in the United States. While the development of the algorithms and software tools in Part I are significant, issues concerning operational implementation in Part II are most impactful, as they address foundational issues that have largely been unaccounted for and remain unexplored from prior works. As we see from this thesis, it is the implementation challenges—not the technology development—that present the greatest barriers to strengthening the foundations of the friction ridge discipline. Taken together, this thesis provides insights that have the potential for impacting the development, implementation, and regulation of algorithms in forensic science, and the roadmap proposed in chapter 7 provides a foundation for establishing international standards related to the use of algorithms operationally for forensic interpretation. In the discussion that follows, I briefly summarize the salient points from each chapter and the associated impact and implications to policy, procedure, and practice.

Chapter 2 – Quality Assessment Software (DFIQI)

The DFIQI algorithm presented in chapter 2 is a novel method, developed as a stand-alone software application, capable of measuring the clarity of friction ridge features (locally) and evaluating the quality of impressions (globally) across three different scales: value, complexity, and difficulty. The evaluation and validation results show performance characteristics that are generally in agreement with experts' personal determinations. The most significant impacts of this algorithm are that it provides fingerprint experts: (1) a more rigorous approach to friction ridge examination by providing an empirical foundation to support their determinations from the Analysis phase of the examination methodology, (2) a framework for organizations to establish transparent, measurable, and demonstrable criteria for Value determinations, (3) a means of flagging impressions that are vulnerable to erroneous outcomes or inconsistency between experts (e.g., higher complexity and difficulty), and (4) a method for quantitatively summarizing the overall quality of impressions for ensuring representative distributions for samples used in research designs, proficiency testing and error rate testing.

The DFIQI algorithm is important in that it provides a means for the friction ridge community to respond to the growing concerns from scientific and legal stakeholders (e.g., [3, 7]). The quantitative measures provided by the DFIQI algorithm enables greater transparency to which features were considered and how they were evaluated, improves consistency between examiners, and reduces impacts of cognitive biases related to the interpretation of feature clarity and assessment of overall quality of the impression. In addition to serving as a tool to be responsive to the most pressing concerns from the scientific and legal communities [3, 7-9], the DFIQI algorithm builds on prior works by Langenburg [13] and Hicklin [15], among others, to enable a practical application of some of their recommendations. Specifically:

a. The DFIQI algorithm provides an objective basis to conveying the clarity of features that are selected, thereby augmenting the subjective assessment of the GYRO method.

b. The DFIQI algorithm can be used to assess the level of quality of impressions (value, complexity, and difficulty) early in the examination process and trigger enhanced quality assurance practices, where relevant.

c. Decision thresholds related to the Analysis phase of the examination methodology can be formalized using quantitative measures based on empirical research evidence and documented transparently *a priori* in organizational policies and procedures to ensure consistency between examiners and minimum standards for quality assurance.

d. Competency tests, proficiency tests, and error rate tests can be based on measurable attributes of the impression providing greater transparency and specificity as to the extent they are representative of the quality of impressions encountered in casework and vice versa.

In addition to the theoretical impacts of the DFIQI algorithm, importantly, it has the potential to be applied practically. The algorithm has been implemented into an easy-to-use and freely available software application (available at: https://doi.org/10.5281/zenodo.4426344) that

does not require advanced training, computational resources, or financial overhead. The application can be downloaded and installed on most stand-alone computers without requiring any advanced technical support, additional software coding or dependencies.

While the impacts and implications of the DFIQI algorithm are far reaching, there are technical limitations to consider. The most significant is that the algorithm relies on the features annotated by the expert and does not take into account all aspects of the friction ridge detail. Consequently, the system should not be considered as a means of supplanting expert interpretation and judgment when analyzing friction ridge detail. Rather, the DFIQI should be considered a tool to support experts' judgments or detect potentially problematic impressions necessitating further quality assurance review.

Chapter 3 – Statistical Interpretation Software (FRStat)

The FRStat algorithm presented in chapter 3 is a novel method, developed as a stand-alone software application, capable of measuring the similarity of feature configurations between two impressions and providing a statistical assessment of the significance of an association. The evaluation and validation results show strong performance characteristics, often with values supporting specificity rates greater than 99%. The most significant impacts of this algorithm are that it provides fingerprint experts: (1) a more rigorous approach to friction ridge examination by providing an empirical foundation to support their subjective opinions during the Evaluation phase of the examination methodology that two impressions potentially share a common source, (2) a framework for organizations to establish transparent, measurable, and demonstrable criteria for Association conclusions, and (3) a means of flagging comparisons that are vulnerable to erroneous outcomes or lack sufficient empirical support for the association.

Like the DFIQI, the FRStat algorithm is important as it provides a means for the friction ridge community to respond to the growing concerns from scientific and legal stakeholders (e.g., [3, 7-9]). The quantitative measures provided by the FRStat algorithm enables greater transparency to which features were considered and the overall support they provide to the opinion, improves consistency between examiners, and reduces impacts of cognitive biases related to the interpretation of similarity and assessment of overall strength of an association. In addition to serving as a tool to be responsive to the most pressing concerns from the scientific and legal communities [3, 7-9], the FRStat algorithm builds on prior works by Langenburg [13] and Hicklin [15], among others, to enable a practical application of some of their recommendations. Specifically:

a. Decision thresholds related to the Evaluation phase of the examination methodology can be formalized using quantitative measures based on empirical data and documented transparently *a priori* in organizational policies and procedures to ensure consistency between examiners and minimum standards for quality assurance.

b. Analysts' opinions that two impressions share a common source can be augmented by a demonstrative means of representing the strength of the corresponding features between two impressions.

c. Examination results can be conveyed along a continuum, providing greater transparency as to the significance of a given comparison and avoiding discretization error that could otherwise result in borderline cases when conclusions are restricted to arbitrarily defined categories

d. Examination results can be expressed probabilistically with explicit recognition of the inherent uncertainties and within appropriate epistemic limits, rather than categorically with implications of absolute certainty that risk exaggerating the weight of the evidence and unduly biasing judicial fact-finders.

To further illustrate its significance, in September 2019, Dr. Karen Kafadar, the President of the American Statistical Association, testified before the United States Congress and referenced the FRStat algorithm as one of three key examples of progress for the forensic science community since the 2009 NAS report [121].

In addition to the theoretical impacts of the FRStat algorithm, like the DFIQI, it has the potential to be applied practically. The algorithm has been implemented into an easy-to-use and freely available software application (available at: https://doi.org/10.5281/zenodo.4426484) that does not require advanced training, computational resources, or financial overhead. The application can be downloaded and installed on most stand-alone computers without requiring any advanced technical support, additional software coding or dependencies.

While the impacts and implications of the FRStat algorithm are far reaching, there are technical limitations to consider. The most significant are that (1) the similarity statistic values are dependent upon the subjective detection and annotation of friction ridge skin features by the human expert, (2) the algorithm is only able to consider what the expert annotates and is not able to evaluate the accuracy of feature annotations by the expert, (3) the algorithm requires a minimum of five features and a maximum of fifteen features, (4) the algorithm accounts for lateral distortions of friction ridge skin impressions on flat surfaces and may not capture all types of extreme distortions which may be encountered in practice, such as substrate, matrix, or photographic effects, and (5) the algorithm is not designed to evaluate all aspects of impressions, such as pattern type, feature type, ridge counts, and other types of features and information considered by an expert; thus, the quantitative results are artificially attenuated and conservative. Consequently, the system should not be considered as a means of supplanting expert interpretation and judgment when comparing and evaluating friction ridge detail. Rather, the FRStat should be considered a tool to support experts' judgments or detect potentially problematic comparisons necessitating further quality assurance review.

Chapter 4 – Toward Objectivity: Integrating Algorithmic Outputs

The integration of algorithmic outputs discussed in chapter 4 allows us to explore more objective interpretation schemes by linking the output of one algorithm as the input to another thereby reducing human influence as the intermediary between the Analysis and Evaluation phases. The discussion provided in this chapter demonstrates that algorithmic integration is

possible and could be a valuable tool for improving the objectivity, transparency, and standardization across different phases of the examination process. The results show a marginal increase in cases that would have been flagged for quality assurance review if the algorithms had been integrated together and implemented operationally compared to just using the FRStat as a final check before issuing a report. From a quality assurance perspective, these results suggest the integration of the algorithms is an effective tool to triage cases and suggest when and where to strategically focus resources, as compared to applying arbitrary sampling schemes, which are unlikely to detect potential issues as efficiently as the algorithms.

This chapter also allows us to implicitly explore the possibilities of semi-automated workflows and the impact this approach could have operationally. Although our exploration here in this thesis is limited to integrating the DFIQI and the FRStat algorithms to augment one another (e.g., the output of the DFIQI related to feature clarity is used as a control for which features are reliable enough to be utilized by the FRStat), we can imagine further integration with automated feature detection algorithms and automated comparison software, such as Automated Fingerprint Identification Systems (AFIS). AFIS algorithms have long been a valuable tool for the friction ridge community to provide efficient means of searching, storing, and retrieving friction ridge impressions; however, they stop short of providing statistical measures that may be used to articulate the significance of the correspondence. Through further integration of the DFIQI and FRStat algorithms (or other algorithms that may be developed bearing similar capabilities) with automated feature detection algorithms and AFIS algorithms, we can envision semi- or fully-automated workflows where algorithms are responsible for the full range of the examination to be within reach. In these circumstances, the human expert would no longer bear the burden of manual and laborious examination tasks. Instead, the expert could transition to an oversight role managing the performance of the algorithms, yielding overall greater operational throughput along with improved consistency and accuracy of results (e.g., level 4 implementation as described in chapter 7).

As this chapter describes, the DFIQI and FRStat algorithms can be practically integrated through a bridge between the algorithms or enforced through policy and procedure; thus, the practical benefits of algorithmic integration of these systems are feasible today. However, the greater benefit of this exploration is to demonstrate the feasibility more broadly and illustrate the real possibility that friction ridge examination could plausibly transform from a subjective method to an objective method in the near-to-mid-term, as aspired and encouraged by members of the scientific community (e.g., see [7]).

Chapter 5 – Evaluation of Practitioners' Perspectives

The survey presented in chapter 5 allows us to explore reporting practices across the friction ridge discipline and practitioners' perspectives related to probabilistic reporting. Prior efforts to introduce statistical thinking and probabilistic reasoning into practice have been met with mixed reactions—a common one being skepticism, or downright hostility, toward this objective. Recognizing that the outputs of algorithms are almost always based on statistical principles and probability theory and that methods of reporting their results are likely to take some probabilistic form, it is critical to understand why practitioners have responded with such aversion. This study

enables us to consider, from a social science perspective, sources of concern and barriers to implementation among the current practitioner workforce pertaining to the introduction of statistics and probabilistic reporting methods in friction ridge examination. Key findings from this survey include: (1) despite the apparent receptivity to probabilistic reporting by a small number of practitioners, we found that the vast majority—approximately 98%— of respondents continue to report categorically with explicit or implicit statements of certainty; and (2) two-thirds of respondents perceive probabilistic reporting as "inappropriate"—their most common concern, shared by approximately 80% of respondents—being that defense attorneys would take advantage of uncertainty or that probabilistic reports would mislead, or be misunderstood by, other criminal justice system actors. Free text responses provided by practitioners were diverse and not limited to issues of whether probabilistic reporting is scientifically more appropriate. In fact, some respondents acknowledged probabilistic approaches were more scientifically appropriate yet continued to defend traditional categorical reporting practices. Overall, attitudes toward probabilistic reporting appear to be influenced by educational, philosophical, psychological, and complex judicial implications and longstanding cultural and institutional norms.

The findings from this survey are insightful. Although practitioners' general resistance to probabilistic concepts is anecdotally unsurprising, the depth and breadth of those reactions illustrate the complexity surrounding this very narrow issue. The candidness of the respondents has allowed a glimpse into the raw and unpolished perspectives shared by many practitioners. This insight is critical to understanding how to move forward. If practitioners remain averse to the mere idea of introducing probabilistic concepts in their reporting schemes, it is almost incomprehensible to think they would be willing to embrace more complicated algorithms that are built on those very principles. In that sense, the first step toward implementing algorithms requires consideration of practitioners' perspectives related to their receptivity to introducing probabilistic reasoning into their reporting practices. The most significant conclusions we can draw from this survey are (1) there is a dire need for establishing foundational educational curricula within the pattern evidence disciplines related to statistics, probability, and uncertainty as pillars to understanding how to properly interpret forensic findings and ensure practitioners are competent and comfortable in doing so, and (2) there is a need to develop a common understanding of what it means to be a forensic expert and the role they play within the broader criminal justice system as it relates to the expression of forensic results. Currently, these preconditions remain largely unaccounted for, which has stifled progress toward responding to the issues raised by scientific and legal commentators. Addressing these issues will require long-term commitment and coordinated investments from all stakeholders, including forensic science administrators, educators, trainers, practitioners, policy makers, and the judiciary to establish and deliver the curricula, develop and enforce scientifically appropriate reporting standards, and provide operational environments that are conducive to transitioning from categorical reporting to probabilistic reporting.

Chapter 6 – Evaluation of Stakeholders' Perspectives

The semi-structured interviews presented in chapter 6 allows us to explore various perspectives from key stakeholders within the criminal justice system (forensic science managers, prosecuting attorneys, defense attorneys, judges, and other academic scientists and scholars) on issues related to: (i) interpretation and reporting practices (with or without algorithmic tools) and

(ii) the implications of the use of computational algorithms as a means of calculating the probabilistic values assigned to forensic science evidence in the American legal system. While forensic practitioners will ultimately be responsible for implementing the proposed solutions, it would be incomplete to focus *solely* on perspectives of forensic practitioners. To fully understand the issues and more effectively facilitate improvements to traditional practices, we must also account for the perspectives of *all* stakeholders within the criminal justice system—not just practitioners. This study yielded rich information illustrating stakeholders' diverse viewpoints on various issues and provided valuable insights into the different perspectives affecting the current discourse in forensic science. More importantly, the construct of this study enabled us to consider these perspectives in greater breadth and depth based on unstructured responses from participants' *own words*.

This study resulted in several key findings. On issues related to interpretation and reporting practices, generally speaking, we found prosecutors' perspectives often represented one extreme end of a spectrum and defense attorneys' perspectives represented the other extreme. Perspectives from laboratory managers tended to align more closely with prosecutors in the sense that they maintained perspectives that traditional practices were acceptable (although maybe not ideal); however, judges and academic scholars tended to align more closely with defense attorneys in the sense that they were more critical and expressive of their concerns as it relates to traditional practices. Further, although stakeholders generally agreed on the roles and responsibilities of experts and the importance of ensuring that opinions expressed during testimony are accompanied by the underpinnings or statistical data to support those opinions, they differed in their views related to whether forensic practitioners are adequately fulfilling those roles and responsibilities and whether disclosing the underpinnings is necessary from scientific and legal perspectives. On issues related to the use of computational algorithms in court, stakeholders generally agreed with one another and pointed to several benefits algorithms could provide to improving the scientific rigor and efficiency of examination practices; however, they often differed in how they viewed the limitations of those algorithms. Prosecutors were most concerned that algorithms would unduly complicate reporting and testimony, making it more difficult for lay fact-finders to understand the testimony. Defense attorneys, judges, academic scholars, and laboratory managers, on the other hand, were most concerned about the transparency surrounding these systems, how to ensure the underlying validity of the systems, and the risks of analysts and lay fact-finders blindly relying on the output of algorithmic systems without fully understanding and accounting for their limitations. These concerns highlight the need to carefully consider the more nuanced details around their development and implementation—the central issue being how algorithmic systems can be trusted for court purposes that can directly impact human liberty decisions.

Ultimately, all stakeholders expressed a shared desire to ensure that forensic conclusions are both scientifically defensible and easily interpretable and that computational algorithms are developed and implemented in a responsible and practical manner so as to uphold the values of fairness and equal justice under the law. How this can be accomplished, however, requires careful consideration of the implications and second and third order effects of various proposed options. Overall, the insights gained from these interviews provide a critical foundation for us to ensure we have a balanced view of the various perspectives on these complicated issues so that we can consider a strategy moving forward that is both cognizant and respectful of the different views and generally amenable across all stakeholder groups.

Chapter 7 – Implementation of Algorithms: A Responsible and Practical Roadmap

The literature review and discussion presented in chapter 7 addresses key challenges and considerations and presents a path forward for the implementation of algorithms in pattern and impression evidence disciplines within forensic science. The discussion allows us to first explore human-algorithm interactions outside the narrow context of forensic science and understand *why* practitioners (in general) tend to oppose algorithmic interventions and *how* their concerns might be overcome. Drawing on the wealth of literature addressing this issue within the domains of medicine and autonomous driving, we are able to gain an understanding of the diverse sociopsychological factors affecting the implementation of algorithms operationally and identify subtle strategies that have demonstrated success with lowering the barriers to adoption and increasing practitioners' receptivity to algorithms. Then, within the context of forensic science, we are able to explore judicial consequences resulting from the use of algorithms in legal settings as it relates to admissibility and Constitutional provisions within the context of the American legal system. Finally, with these considerations in mind, we are able to outline both a roadmap for implementing algorithms operationally and a taxonomy for describing the various ways algorithms can be implemented in a *responsible* and *practical* manner that is responsive to the needs and perspectives of all criminal justice stakeholders.

These findings from the literature related to human-algorithm interactions are enlightening. Most surprising was the consistency of those results over time as the medical community, once reliant solely on human judgment for clinical decision making, began to embrace algorithmic developments and transition to more objective foundations over a period of half a century. Equally interesting was the consistency of those findings across diverse domains unrelated to medicine, such as the transition toward autonomous driving and the public's willingness to rely on algorithms for dynamic driving tasks. The key finding is that the implementation of algorithms into domains traditionally dominated by human judgment is often fraught with resistance. People tend to exhibit a general aversion to algorithms and prefer to rely on their own judgment—often despite knowledge that their own judgment is typically inferior to that of algorithms. This phenomenon is exacerbated when people possess domain expertise, are faced with high-stakes decisions, and are presented with an algorithm that is susceptible to err. Although the actual source of these reactions has not yet been fully understood, some researchers have pointed to various sociopsychological factors, overconfidence bias, and a general lack of trust in algorithms' abilities to account for idiosyncratic factors as possible explanations for the behaviors. Anecdotal observations of human-algorithm interactions in both of these different domains as well as recent research have suggested that people tend to be more receptive to algorithms if the algorithms are integrated as a factor that *supplements* as opposed to *supplants* human decision making so that the human retains some amount of influence on the ultimate outcome.

These details provide important context when considering the implementation of algorithms into forensic science. Indeed, forensic science has the major conditions for which algorithm aversion is most pronounced: (i) forensic examination results (in the pattern evidence domains particularly) are traditionally based entirely on subjective judgment, (ii) forensic examiners possess expertise, and (iii) forensic conclusions involve high-stakes decision-making—

inaccurate conclusions could deprive individuals of life and liberty, not to mention risking the professional reputation and career of the experts themselves. Thus, we have no reason to expect the reactions and behaviors of forensic practitioners to be substantially different than what has been observed in research and other domains explored. In fact, to some extent we have already observed similar behaviors manifest: practitioners' reactions to the mere notion of implementing statistical approaches have been met with criticism and opposition, and when given the opportunity to incorporate algorithms into their decision-making, practitioners tend to disregard them in favor of their own judgments.

In addition to characterizing the concerns from practitioners as it relates to their willingness to embrace algorithms, we also have to maintain perspective of the sensitive (and unique) needs of the criminal justice system. Ultimately, the legal system is concerned with ensuring defendants receive fair and equitable justice under the law. Accordingly, courts will need to consider the admissibility of algorithms against existing legal standards and ensure they are used in a way that does not infringe on defendants' Constitutional rights. From our review, we found that this can be particularly challenging given the "black-box" nature of many algorithms and defendants' Constitutional rights to confront and challenge the evidence against them. Successful implementation will require consideration of not only practitioners' willingness to adopt the algorithms, but also sensitivity to the needs of the legal system. The implications of this are wide-ranging and impactful to how statisticians and researchers might propose various algorithms, how computer scientists might encode the algorithms into computer source-code, how vendors will distribute the algorithms and whether—and to what extent—they will exert trade-secret and other intellectual property related protections, how the algorithms will be validated, and how practitioners will use the algorithms operationally.

Collectively, these details are important as they allow us to predict how a transition toward more objective foundations using algorithms is likely to play out in forensic science and identify a roadmap *a priori* to navigate the transition more efficiently and effectively with consideration of the diverse perspectives and sensitive needs of all stakeholders. First, the seven pillars we describe for a responsible implementation outline key elements of a quality assurance system that should be in place before an algorithm is implemented operationally. Second, the formal taxonomy we propose to support a practical implementation provides a common language and framework for describing the various ways algorithms can be implemented operationally and the implications of such implementation at each level. Taken together, this roadmap has the potential to impact (1) how researchers approach the development of algorithms, (2) how practitioners will approach the implementation of algorithms, and (3) how policy makers will regulate the use of algorithms across all forensic science disciplines. As such, it provides a foundation for establishing international standards related to the implementation and use of algorithms operationally for forensic interpretation and reporting.

Chapter 8 – Operationalization of Algorithms: Anecdotal Reflections and Observations

The discussion presented in chapter 8 describes the journey leading up to, and including, the implementation of the FRStat algorithm into operational practice at a federal laboratory in the United States. The discussion is presented retrospectively and based on my *own* experiences—

both as the laboratory manager responsible for the implementation and as a private analyst in which the algorithm became a point of contention during litigation. From a retrospective view, I am able to provide greater context to the issues we faced and special insights into the factors we deliberated over that were influential in our ultimate decisions concerning the policies and procedures governing the use of the algorithm operationally as well as how well our approach to implementation compared to the roadmap outlined in chapter 7. Additionally, I was able to review a case involving the use of the FRStat algorithm in litigation—providing insights related to how the algorithm was received from a legal perspective during actual litigation and the various arguments raised by the parties during the case. These insights provide a level of context to the issues that is not typically available in other theoretical discussions. In that light, the anecdotal reflections presented in chapter 8 offer impactful details that are missing from traditional literature sources which will undoubtedly shape others' perspectives and inform downstream approaches and decisions related to the implementation and use of algorithms operationally. I would argue that it is difficult to understate the value of this practical perspective.

In addition to the practical insights described above, the acts related to the implementation of the algorithm operationally had immense impact—not only in terms of the policies, procedures, and practices at our own laboratory, but also around the country. In the journey leading up to the implementation of the algorithm, examiners were confronted with our decision to move away from categorical reporting in lieu of probabilistic approaches. While it was not our intention that others in the community would have to answer to our decisions, it was an unavoidable consequence that helped keep the issue of probabilistic reporting at the forefront of the discipline. It was no longer possible to "write off" the challenges to traditional reporting schemes as partisan litigation tactics. In my view, the impact of this as it relates to disrupting long-standing practices and culture is unquestionable. It represented a concession within the discipline that we needed to evolve considering the concerns that have been raised over the years by scientific and legal commentators. Further, and more importantly, it demonstrated that it was possible, and practical, to evolve without significant interruption to existing workflows and practices. When the FRStat algorithm was ultimately implemented, the implications grew even stronger as practitioners were no longer able to argue that statistical applications were impossible. The mere availability of the algorithm and the decision by some *not* to explore it revealed that many in the friction ridge community were not evolving not just because of technology limitations; rather, it highlighted a deeper issue at hand—some were steadfastly resistant to the idea of introducing statistical methods, and progress toward improving the scientific foundations of the friction ridge discipline were stifled due to a widespread cultural hesitation exacerbated by lack of foundational education, among other more complex sociopsychological concerns.

Finally, the case review describing the use of the FRStat algorithm in litigation has significant historical implications as it is the first case in a civilian (non-military) court in the United States in which a judge had to grapple with issues concerning the algorithm applied to friction ridge evidence. Thus, through this case review, we are able to explore first-hand how the judicial system reacted to the use of the algorithm and the arguments raised by both parties compared to the theoretical discussions of how the legal system might react presented in chapter 7 (i.e., [53]). While the first-hand review is illuminating, the context of the case was peculiar—it was not a traditional trial for which the prosecutors sought to introduce the algorithm proactively. Instead, the need for probabilistic reporting of findings from friction ridge examinations was raised

in a post-conviction appeal for relief and retrial, claiming, among other things, the categorical statements presented at trial related to the friction ridge findings were exaggerated and unsupportable; instead, the applicant under appeal pointed to our laboratory and suggested probabilistic methods should have been used. In response, and for purposes of accommodating the applicant's request, prosecutors requested the FRStat to be applied to the friction ridge impressions. Immediately thereafter, however, the applicant began to challenge the results of the FRStat algorithm and attack its credibility and admissibility of the evidence—evidence to which the prosecutors argued were the result of the applicant's own request. Ultimately, the Court did not issue any ruling on the admissibility of the FRStat algorithm specifically. Instead, given the specific question before the court and burden on the applicant to raise new evidence that demonstrated probabilistic methods would have resulted in a different outcome at trial if they were to be used, the judge accommodated the applicant's later request not to consider the FRStat results but in doing so concluded that the applicant failed to demonstrate probabilistic methods would have resulted in a different outcome at trial. The court did, however, recognize the utility in leveraging probabilistic models and statistical tools to provide "technological advances" to the friction ridge examination process. Although the context of the litigation and conditions which led to the use of the FRStat are atypical and limit its impact to future legal precedence, the significance of this case is clear, and the insights related to the arguments raised concerning the use of the algorithm and how the results from the algorithm are expressed are invaluable as we consider our approach moving forward.

9.2 Recommendations

The work completed as part of this thesis has lasting impacts and wide-ranging implications for the friction ridge discipline. The availability of the tools presented in Part I and the results of the discussions presented in Part II support several recommendations to strengthen the foundations of friction ridge examination and improve our understanding of the reliability of the findings used as evidence that our nations' legal system depends on. Many recommendations are directed toward the friction ridge community and encourage changes to existing policies, procedures and practices. Other recommendations are directed toward institutions and other stakeholders that play a role in forensic science more broadly, including those responsible for establishing standards and enacting legislation. Collectively, it is the responsibility of all stakeholders to consider their parts in moving the friction ridge discipline forward. As a general caveat, when making the recommendations below, I do so without reservation of resource requirements or procedural requirements that would be necessary to enact such changes. I also offer these recommendations without consideration for peculiar circumstances or downstream implications of such recommendations that will undoubtedly need to be addressed as a result. As such, I recognize that these recommendations are not exhaustive and, while many of these recommendations are achievable today, some might be aspirational and require multi-stakeholder engagement. For those that are not immediately actionable, my hope is that these recommendations will stimulate further dialogue among the relevant stakeholders with the intent of exploring and debating the issues in further detail to identify a path forward that is sensitive to the needs and concerns of all stakeholders that make up the criminal justice system. Finally, I offer these recommendations as my own personal views and opinions, and they should not be construed as official or reflecting those views or positions of any

people, entities, or organizations for which I am affiliated. The order for which these recommendations are provided is not reflective of their priority or significance.

9.2.1 Better algorithms should be developed and made accessible to all stakeholders

The algorithms presented in chapters 2 and 3 provide a means to empirically substantiate examiners' opinions; however, they are far from optimal. Additional, more powerful algorithms should be developed with improvements to both the performance and usability. The algorithms should take into account more discriminating attributes than just minutiae and be more robust to extreme distortions, resulting in increased sensitivity and specificity. The algorithms should have higher computational efficiency, enabling measurements to be made against large databases containing millions of other impressions. The algorithms should be designed such that they are easily able to be integrated with other algorithms and software applications, such as automated feature extraction software and automated comparison software, such as AFIS to minimize subjective influences. The algorithms should be designed with user interfaces that are intuitive to practitioners, require little-to-no advanced training to operate, and are compatible with existing workflows and examination methodologies.

In addition to improved performance and usability, the algorithms should be accessible to all stakeholders in the criminal justice system. The algorithms, and their underlying source-code, should be available for independent review. The algorithms should be conceptually comprehensible and explainable to lay fact-finders. The algorithms should be developed and supported by government, educational, or non-profit institutions, and they should be publicly accessible to all practitioners and stakeholders and void of potential conflicts or other financial or proprietary interests. For algorithms that require training data, care should be taken to ensure the data were not collected in ways to create biased samples, are adequately sized and appropriately representative to the issues, are accurately labeled, and are publicly available to permit independent review and testing. Finally, all validation materials should be made publicly available and accessible to permit independent review and testing.

9.2.2 Algorithms should be regulated by an independent authority

Algorithms used for criminal justice purposes should be regulated by an independent authority, such as by a government, educational, or non-profit institution. The independent authority should consist of a balance of scientific, legal, and forensic science practitioner stakeholders and address the foundational validation of the method as well as factors related to its design, development, and operational use. The validity of the algorithm should be assessed following independent review of the materials proposed to support the validation and/or independent testing of the algorithm and source-code in a controlled and systematic manner. Approvals should be based on empirical evidence that the algorithm does what it purports to do, is capable of achieving acceptable performance appropriate for the intended application and level of implementation proposed, and offers accessibility of the relevant materials related to the validation to permit additional review and scrutiny as provided by Constitutional provisions. To ensure equality of arms (i.e., resources) [236] and unbiased assessments, courts and members of

the judiciary should be able to defer to the independent authority on issues related to admissibility—the courtroom is not the appropriate forum for complex scientific issues to be debated nor are judges and fact-finders adequately equipped to make informed decisions on issues regarding scientific validity of complex algorithmic systems.

9.2.3 Standards specifying minimum requirements for implementing algorithms operationally should be established.

The discussion in chapter 7 proposed seven pillars that should be accounted for to ensure an appropriate foundation has been established to support the implementation of algorithms in a responsible manner. An appropriate foundation requires a formalized quality management system to be in place to ensure conformance with the requisite requirements related to the following: education, training, protocols, validation, verification, competency, and on-going monitoring schemes. Just as laboratory accreditation ensures that laboratories conform to specified requirements to be considered competent for testing (e.g., conformance to ISO/IEC 17025 – General requirements for the competence of testing and calibration laboratories [237]), standards should be in place to ensure laboratories conform to specified requirements related to these seven pillars to demonstrate their competence to apply algorithms operationally.

Further, the discussion highlighted that algorithms can be implemented in many different ways with different implications to the examination methodology and subsequent litigation. A formal taxonomy should be adopted by a standards setting body—much like the standard J3016 Levels of Driving Automation [157] adopted by the automotive industry—to standardize the terms and definitions related to the implementation of algorithms according to six different levels, with each level representing a gradual transition from human to machine as the basis for forensic conclusions, ranging from Level 0 (no algorithm influence) to Level 5 (complete algorithm influence). A standardized taxonomy will be critical to promote a common language and framework for describing the various ways algorithms can be implemented operationally and the implications of such implementation at each level as well as establishing the extent to which algorithms are validated for operational use (e.g., an algorithm might be considered valid for Levels 2 or 3 implementation, but not valid for Levels 4 or 5 implementation). The taxonomy must be standardized internationally to ensure all stakeholders can assess the validity of the algorithm for its intended purpose and have a clear understanding of how the algorithm is implemented and the associated implications.

9.2.4 Standards specifying minimum educational requirements for forensic practitioners should be expanded to include a more rigorous emphasis on scientific interpretation.

The friction ridge community lacks a fundamental understanding of principles of scientific interpretation and basic concepts related to probability and statistics, uncertainty, logic, reasoning, and decision making. Consequently, many are woefully unaware and naively resistant to issues concerning traditional reporting practices and the limitations of forensic conclusions. This lack of understanding has led many to espouse irrational arguments and appeals to personal opinion or experiences, which has severely stifled the community's ability to respond and evolve in light of

the growing concerns raised by scientific and legal communities. Requirements related to minimum qualifications and standards of competency for friction ridge practitioners, particularly within the United States, have emphasized tactical skills over scientific knowledge. While this has resulted in more examiners entering the field with baseline technical capabilities, they lack a critical perspective that is necessary to properly interpret forensic findings. Standards setting bodies should expand their requirements related to minimum qualifications and standards for competency to include a more rigorous emphasis on principles of forensic interpretation, requiring all candidates to undergo basic and advanced coursework related to probability, statistics, uncertainty and logic and reasoning. Educational institutions should evaluate their current curricula and require these courses in their forensic science programs, as well as make these courses available for existing practitioners to strengthen the knowledge of the current workforce. Finally, forensic science managers should invest in, and prioritize, topics related to forensic interpretation for all practitioners as part of their continuing education cycles as quickly as possible.

9.2.5 Reported results should be scientifically defensible and expressed with clear characterizations of their limitations

Friction ridge examiners should cease reporting associative conclusions in a categorical framework using terms or phrases which imply absolute certainty or the capacity to individualize an impression to a single source. Such claims have been shown to be scientifically indefensible. Instead, practitioners should report their findings in a probabilistic framework and explicitly account for the uncertainties inherent in the examination and interpretation (i.e., the theoretical possibility of coincidental correspondence between non-mated impressions and error rates associated with the performance of examiners through blind testing). Reporting results probabilistically ensures conclusions are more defensible and expressed in a way that conform to epistemic limits of what can be supported by available data and research. Policy makers and practitioners should not wait until theoretically ideal algorithmic tools are available before transitioning to a probabilistic framework. Rather, this transition should occur immediately to avoid the perpetuation of indefensible claims. In situations where statistical data or algorithmic tools are available, the results should be expressed quantitatively and within the limits set by validation. In situations where sufficient statistical data does not exist to permit empirically grounded quantitative claims, the results should be expressed qualitatively such that the existence and potential sources of uncertainties are explicitly acknowledged. In all situations, however, the basis for those probabilistic expressions should be made clear (e.g., human judgment vs. algorithmic tools) and those expressions should not use numeric references unless they are based on validated statistical data and methods. As research continues to evaluate an optimal means of expressing probabilistic information to lay audiences, practitioners should use a combination of different probabilistic approaches and descriptors to communicate the results. Doing so will leverage the strengths of different approaches and increase the likelihood that lay stakeholders and fact-finders appropriately comprehend the strengths and limitations of the findings.

9.2.6 Analysts' opinions should be distinguished from reported conclusions

Quality assurance managers should create clear distinctions between analysts' *opinions* and reported *conclusions*. The analysts' opinions are just that—the outcome of their subjective assessment of the findings based on their knowledge, skills, training, and experience. The reported conclusion, however, is the outcome of a system controlled by a laboratory's quality assurance program, which would naturally encompass the analyst's opinion, among other inputs and criteria. The individual analysts are responsible and accountable for their opinions while the laboratory quality managers are responsible and accountable for the reported conclusions. This is an important distinction, as it holds the laboratory *and* the analyst accountable to the veracity of the result (as opposed to just the analyst, which has typically been the case when reported conclusions are based entirely on the analyst's subjective assessment). Although we would expect and strive for ensuring analysts' opinions coincide with reported conclusions, it might not always be the case. When these situations occur, ground truth is not available to decide which outcome is "correct," and laboratories should not arbitrarily disregard one outcome over the other; thus, quality managers will need to have a process for adjudicating the different outcomes in a balanced and transparent manner. Creating this distinction will promote a sense of procedural justice in the adjudication process and increase examiners' receptivity to algorithmic tools. It will provide reassurance to examiners and stakeholders that their subjective expertise is valued and accounted for in the overall result while at the same time ensuring appropriate safeguards are in place if all criteria (e.g., algorithmic assessments and decision thresholds) are not met to justify the conclusion.

9.2.7 Algorithms should be implemented operationally

Friction ridge examiners should implement algorithms in an effort to be responsive to the growing concerns from scientific and legal stakeholders regarding the lack of empirical substantiation to their opinions and quantitative measures of the quality and significance of the findings in a given case. Although the algorithms need not be those specific ones presented in chapters 2 and 3 if there are other algorithms available and preferred, the DFIQI and FRStat are freely available, publicly accessible, and foundationally validated. The implementation plan should take into consideration the various stakeholders' perspectives presented in chapters 5, 6, and 7 as well as the nuanced issues and anecdotal reflections discussed in chapter 8. Implementation need not be immediate or done in haste. Rather, it should be done deliberately and in accordance with the roadmap outlined in chapter 7. First, the foundation must be laid out by thoroughly addressing and accounting for each of the seven pillars outlined as it relates to a responsible implementation (education, training, protocols, validation, verification, competency, and on-going monitoring schemes). Then, once the foundation has been established, the algorithm(s) should be implemented starting at Level 1 and then progressing sequentially toward the intended target (e.g., Level 2 or 3).

9.2.8 Examination, interpretation, and reporting practices should be governed by centralized policy and oversight

Over the years, considerable investments have been made to improving forensic science practices. Many traditional practices have been criticized for lacking rigorous scientific foundations, and recommendations for improvement have been made by scientific committees (e.g., [3, 7-9]) and academic scholars (e.g., [13, 15, 16]); however, in the absence of catastrophic events (e.g., the erroneous identification of Brandon Mayfield by the Federal Bureau of Investigation [238]) the practice of friction ridge examination, including many pattern evidence disciplines more broadly, have not demonstrated the capacity to change on their own. This is not necessarily without lack of effort by some practitioners, however. While compelling change is logistically and politically complicated and often a last resort, for forensic science it is now becoming clear as a foundation for a path forward. Not only would establishing enforceable requirements ensure consistency across jurisdictions and provide a mechanism for accountability, it would also provide the basis for forensic service providers, particularly those that are publicly funded, to obtain the resources that they have been deprived of for so long that are necessary to support such changes. The unfortunate reality is that many forensic service providers are woefully underfunded, understaffed, undertrained, and operating in subpar conditions that are conducive to error. Under these conditions, maintaining status quo becomes the primary objective, and acting on recommended improvements are largely an unattainable aspiration. Without enforceable requirements to drive the investment of resources, existing practices will remain permissible and the same vulnerabilities that have been identified over the last decade (or more) will continue. How such policy should be established and the oversight be executed are more complicated and would require significant consideration of political, legislative, and logistical issues. Nevertheless, the need for enforcement and accountability has been realized. For example, although the United Kingdom (UK) has long held the office of the Forensic Science Regulator (FSR) to provide centralized guidance related to forensic practices, the lack of enforcement has stifled its effectiveness to effect change. The Forensic Science Regulator Act of 2021 seeks to change that and provides the UK FSR with the powers of investigation and enforcement [239]. In the United States, the political and legislative divisions create additional layers of complexity, but efforts have been made toward this objective—particularly at the state level (e.g., Texas Forensic Science Commission [240], New York State Commission on Forensic Science [241], and the newly enacted legislation to establish the Illinois Forensic Science Commission [242]). At the national level, the OSAC [243] is a step in the right direction for establishing consensus-based recommendations and standards, but it falls short of possessing any investigation or enforcement capabilities. Looking forward, researchers, policy makers, legislators, educators, laboratory managers, practitioners, litigators, judges, academia, and other relevant stakeholders need to come together to debate how such policy and oversight ought to be structured, resourced, and implemented given the political, legislative, and logistical complexities involved.

10 Conclusion

Friction ridge examination has become one of the most widely practiced forensic science disciplines since it was first introduced into law enforcement operations in the early 1900s. Over the years, friction ridge findings were often regarded as incontrovertible evidence that a particular individual touched an item or was present at the scene of a crime. In recent years, however, examination methods relying on visual comparison and subjective judgment have become a focal point of criticism by the broader scientific and legal communities, challenging the validity and reliability of traditional interpretation and reporting practices. The chief concern was the variability in examination results and the lack of empirical standards and quantitative and statistical measurements to substantiate or document reported conclusions. Without defined measurements or an empirical basis for which the significance of the dactyloscopic observations is evaluated, it is unclear *what* contributed to the overall assessment of the findings and *how* it was evaluated. While the lack of empirical standards and measurements do not necessarily suggest the practice as a whole is unreliable or fraught with error, it does raise questions as to how reliable the assessment is for a specific case at hand.

Over the years, there have been several notable efforts by researchers in which quantitative and statistical tools have been introduced through computational algorithms to strengthen the rigor of friction ridge examination methods; however, none have successfully made it into the hands of practitioners and implemented into routine casework operations. While there are a number of different reasons for this, the single greatest challenge that is often underestimated is the longstanding cultural hesitation and the paradigm shift that would be required to facilitate such a transition. Attention must be directed toward how to most effectively navigate the implementation of these tools in a field that has largely been dominated for so long by human interpretation and experience-based judgment. In the forensic sciences, little effort has been given to such a critical issue. The research conducted in this thesis sought to change that.

The objectives of this thesis were twofold: (1) to develop, validate, and make publicly accessible algorithms and software applications for friction ridge examination capable of (a) assessing the clarity of friction ridge skin features and overall quality of impressions and (b) evaluating the statistical strength of correspondence between two impressions, and (2) to develop strategies for practical application and implementation of these (and similar) tools in an operational forensic science laboratory in a manner that maximizes practitioner receptivity and acceptance across all stakeholders. Successful accomplishment of these objectives required careful consideration of the needs and expectations of the adversarial legal environment, laboratory operational workflows and throughput requirements, practitioner knowledge and skills, and appropriate balancing of human intuition and judgment with quantitative and empirical standards relating to procedures governing the use of the tools and reporting and testimony of the results. Over the course of eight years, between 2013 and 2021, these objectives were achieved through a series of structured empirical studies and discussions related to the development, validation, and operationalization of computational algorithms in forensic science practice.

This thesis is separated into two parts, reflecting the major objectives of the research. Part I (chapters 2 through 4) describes the design, development, and validation of two different algorithmic tools—DFIQI and FRStat—that enable examiners to practically apply statistical

measures to friction ridge impression evidence and explore more objective interpretation schemes. Part II (chapters 5 through 8) explores practitioner and stakeholder perspectives on issues related to the adoption and implementation of algorithmic tools into practice; discusses salient challenges, considerations, and a path forward related to the implementation of algorithms in domains largely dominated by human judgment; and describes details surrounding the actual implementation of the FRStat algorithm into operational practice at a federal forensic laboratory in the United States along with subsequent litigation.

The DFIQI algorithm is designed to measure the clarity of friction ridge features (locally) and evaluate the quality of friction ridge impressions (globally). The FRStat algorithm is designed to evaluate the statistical strength of the correspondence between two impressions. The DFIQI and FRStat algorithms are both freely available and publicly accessible[34]. Both algorithms were developed as stand-alone software applications within a common infrastructure and with consideration of balancing performance against issues related to computational complexity and transparency. While both applications are available stand-alone, the common infrastructure enables them to be algorithmically combined in future work to promote greater workflow automation. When compared to other, more computationally intensive and complex algorithmic tools that have been proposed, both the DFIQI and FRStat resulted in comparable, and in some conditions, superior performance. In an operational environment, both algorithms are intended to (1) strengthen a laboratory's quality assurance program by providing a framework to establish policies and procedures for examination decisions at different points in the broader examination methodology geared toward flagging impressions that are generally lower quality and more vulnerable to disagreements between experts and (2) to enable results to be reported in a more transparent and standardized fashion with clearly defined criteria for conclusions. While the development of the DFIQI and FRStat are significant as they provide a means for the friction ridge community to apply practical solutions and respond to the growing demands from scientific and legal stakeholders for greater empirical substantiation to their opinions and quantitative measures of the quality and significance of the evidence in a given case, the mere availability of the tools is not enough—we must understand how to navigate issues related to their implementation in an operational environment in order to fully realize their benefit.

Perspectives from friction ridge practitioners discussed in chapter 5 reveal a broad range of attitudes toward probabilistic reporting (with and without algorithms) that appear to be influenced by educational, philosophical, psychological, and complex judicial implications and longstanding cultural and institutional norms. Although a small number of practitioners surveyed expressed receptivity to probabilistic reporting, the vast majority—approximately 98%— of respondents continue to report categorically with explicit or implicit statements of certainty. Two-thirds of respondents perceive probabilistic reporting as "inappropriate"—their most common concern, shared by approximately 80% of respondents—being that defense attorneys would take advantage of uncertainty or that probabilistic reports would mislead, or be misunderstood by, other criminal justice system actors. Free text responses provided by practitioners were diverse and not limited to issues of whether probabilistic reporting is scientifically more appropriate. In fact, some respondents acknowledged probabilistic approaches were more scientifically appropriate yet continued to defend traditional categorical reporting practices.

[34] The DFIQI software application can be accessed at: https://doi.org/10.5281/zenodo.4426344. The FRStat software application can be accessed at: https://doi.org/10.5281/zenodo.4426484.

Perspectives from other criminal justice stakeholders (forensic science managers, prosecuting attorneys, defense attorneys, judges, and other academic scientists and scholars) discussed in chapter 6 on issues related to: (i) interpretation and reporting practices more broadly (with or without algorithmic tools) and (ii) the implications of the use of computational algorithms as a means of calculating the probabilistic values assigned to forensic science evidence in the American legal system were more complex and diverse. On issues related to interpretation and reporting, prosecutors' perspectives often represented one extreme end of a spectrum and defense attorneys' perspectives represented the other extreme end. Perspectives from laboratory managers tended to align more closely with prosecutors in the sense that they maintained perspectives that traditional practices were acceptable (although not ideal); however, judges and academic scholars tended to align more closely with defense attorneys and were much more critical and expressive of their concerns pertaining to traditional practices. On issues related to the use of computational algorithms in court, stakeholders generally agreed with one another in pointing to several benefits algorithms could provide toward improving the scientific rigor and efficiency of examination practices; however, they often differed in how they viewed the limitations of those algorithms. Prosecutors were most concerned that algorithms would unduly complicate reporting and testimony, making it more difficult for lay fact-finders to understand the testimony. Defense attorneys, judges, academic scholars, and laboratory managers, on the other hand, were most concerned about the transparency surrounding these systems, how to ensure the underlying validity of the systems, and the risks of analysts and lay fact-finders blindly relying on the output of algorithmic systems without fully understanding and accounting for their limitations.

Looking outside of the forensic sciences at issues related to human-algorithm interactions, the research described in chapter 7 helps us begin to understand *why* practitioners (in general) tend to oppose algorithmic interventions and *how* their concerns might be overcome from analogous circumstances that have demonstrated success over time. The wealth of literature addressing issues related to human-algorithm interactions, both in general and within the domains of medicine and autonomous driving, which are considered relevant proxies to issues related to forensic science, reveal diverse sociopsychological factors affecting the implementation of algorithms operationally. Ultimately, we find that the implementation of algorithms into domains traditionally dominated by human judgment is often fraught with resistance. People tend to exhibit a general aversion to algorithms and prefer to rely on their own judgment—often despite knowledge that their own judgment is typically inferior to that of algorithms. This phenomenon is exacerbated when people possess domain expertise, are faced with high-stakes decisions, and are presented with an algorithm that is susceptible to err—all conditions which are applicable to forensic science. Although the actual source of these reactions has not yet been fully understood, some researchers have pointed to various sociopsychological factors, overconfidence bias, and a general lack of trust in algorithms' abilities to account for idiosyncratic factors as possible explanations for the behaviors. Both anecdotal observations of human-algorithm interactions in different domains over the years and recent research have suggested, however, that people tend to be more receptive to algorithms if they are integrated as a factor that *supplements* as opposed to *supplants* human decision making and the human retains some amount of influence on the ultimate outcome.

Within the context of forensic science, judicial consequences resulting from the use of algorithms in legal settings create another dimension of complexity. Ultimately, the legal system is concerned with ensuring defendants receive fair and equitable justice under the law. Accordingly, courts will need to consider the admissibility of algorithms against existing legal standards and ensure they are designed and used in a way that does not infringe on defendants' Constitutional rights. Although this is a relatively novel topic, various legal arguments proposed in the literature align with many of the perspectives expressed by legal stakeholders discussed in chapter 6 and suggest this can be particularly challenging given the "black-box" nature of many algorithms and the ability of defendants to exercise their Constitutional rights to confront and challenge the evidence against them. The implications of these findings are wide-ranging and impactful with regard to how statisticians and researchers might propose various algorithms; how computer scientists might encode the algorithms into computer source-code; how vendors will distribute the algorithms and whether, and to what extent, they will exert trade-secret and other intellectual property related protections; how the algorithms will be validated; and how practitioners will use the algorithms operationally.

Taking into account these various issues and perspectives from both practitioners and key criminal justice stakeholders, chapter 7 of this thesis proposes a roadmap for the forensic science community to implement algorithms in a *responsible* and *practical* manner. First, seven pillars are described that outline key elements of a quality assurance system that should be in place before an algorithm is implemented operationally. These pillars address issues related to education, training, protocols, validation, verification, competency, and on-going monitoring schemes. Second, a formal taxonomy is established to provide a common language and framework for describing the various ways algorithms can be implemented operationally and the implications of such implementation schemes. This taxonomy consists of six different levels of algorithm implementation, ranging from Level 0 (no algorithm influence) to Level 5 (complete algorithm influence), similar to how the automotive industry approached the issue of autonomous driving. Each level represents a gradual transition from human to machine as the basis for forensic conclusions and includes: Level 0 (No Algorithm), Level 1 (Algorithm Assistance), Level 2 (Algorithm Quality Control), Level 3 (Algorithm Informed Evaluation), Level 4 (Algorithm Dominated Evaluation), Level 5 (Algorithm only). In levels 0 through 2, the human serves as the predominant basis for the evaluation and conclusion with increasing influence of the algorithm as a supplemental factor for quality control (used *after* the expert opinion has been formed). In Levels 3 through 5, the algorithm serves as the predominant basis for the evaluation and conclusion with decreasing influence from the human.

Having explored the various stakeholders' perspectives and considered the issues in a more structured manner in the preceding chapters, chapter 8 provides anecdotal reflections on the journey leading up to, and including, the implementation of the FRStat algorithm into operational practice at a federal laboratory in the United States and how it was handled by judicial actors during litigation—the first time this has ever occurred related to friction ridge impression evidence in the United States. This discussion, based on first-hand experience, provides specific details related to the nuanced issues that went into the decision to implement the algorithm operationally, as well as initial reactions and counter actions that were taken to address the concerns raised by practitioners and other criminal justice stakeholders at the time, all of which ultimately led to the development of specific policies and procedures governing the use of the algorithm and workflow

operationally. Although the implementation of the FRStat algorithm and subsequent litigation occurred prior to many of these issues being explored in a more structured manner (as described in chapters 5 through 7), the concerns that were raised and the issues that had to be worked through in real-time reinforce the importance of approaching the implementation of algorithms using the proposed framework.

The work presented in this thesis has broad impact and implications—both theoretical and practical, ranging from statistics and evidence quantification to social psychology and human behavior—affecting policy, procedure, training, quality assurance, research, reporting and testimony, and litigation as it relates to the implementation and use of algorithms operationally in friction ridge examination. The availability of the tools presented and the results of the discussions and proposed framework support eight key recommendations described in chapter 9 to strengthen the foundations of friction ridge examination and improve our understanding of the reliability of evidence that our nations' legal system depends on. Many recommendations are directed toward the friction ridge community and encourage changes to existing policies, procedures, and practices. Others are directed toward institutions and other stakeholders that play a role in forensic science. Collectively, it is the responsibility of all stakeholders to consider their parts in moving the friction ridge discipline forward, and we need to prepare for this transition now before it is inevitably consequential to the enduring validity and admissibility of forensic evidence.

Looking forward, I view a bright but challenging future for the friction ridge discipline, and in many respects, for the broader forensic science community. As we work toward establishing stronger empirical foundations and improving the rigor for many pattern comparison disciplines, we will find that computational algorithms will provide an important role in accomplishing this. This will not be easy, however, and will bring to the forefront a completely new set of challenges facing the forensic science and legal communities that will span across several dimensions. Addressing these challenges will require long-term commitment and coordinated investments from all stakeholders, including forensic science administrators, educators, trainers, practitioners, policy makers, and the judiciary. My hope is that this thesis both lays a foundation for a broader conversation related to these issues to continue and provides a basis for this journey to begin and a framework to guide us forward.

11 References

1. Cole, S.A., *The 'Opinionization' of Fingerprint Evidence.* BioSocieties, 2008. **3**(1): p. 105-113.
2. Haber, L. and R.N. Haber, *Scientific Validation of Fingerprint Evidence Under Daubert.* Law, Probability and Risk, 2008. **7**(2): p. 87-109.
3. National Research Council, *Strengthening Forensic Science in the United States: A Path Forward*. 2009, Washington, D.C. USA: The National Academies Press.
4. Koehler, J.J. and M.J. Saks, *Individualization Claims in Forensic Science: Still Unwarranted.* Brook. L. Rev., 2010. **75**(4): p. 1187-1208.
5. Saks, M.J., *Forensic Identification: From a Faith-Based "Science" to a Scientific Science.* Forensic Science International, 2010. **201**(1-3): p. 14-17.
6. Cole, S.A., *Individualization is Dead, Long Live Individualization! Reforms of Reporting Practices for Fingerprint Analysis in the United States.* Law, Probability and Risk, 2014. **13**(2): p. 117-150.
7. *Report to the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-comparison Methods*. 2016, Washington, D.C., USA: Executive Office of the President of the United States, President's Council of Advisors on Science and Technology.
8. AAAS, *Forensic Science Assessments: A Quality and Gap Analysis–Latent Fingerprint Examination*. 2017, The American Association for the Advancement of Science: Washington, D.C. USA.
9. Expert Working Group on Human Factors in Latent Print Analysis, *Latent Print Examination and Human Factors: Improving the Practice Through a Systems Approach*. 2012, National Institute of Standards and Technology and National Institute of Justice.
10. *Standards for Examining Friction Ridge Impressions and Resulting Conclusions*. 2013, Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST).
11. Dror, I.E., C. Champod, G. Langenburg, D. Charlton, H. Hunt, and R. Rosenthal, *Cognitive Issues in Fingerprint Analysis: Inter-and Intra-Expert Consistency and the Effect of a 'Target' Comparison.* Forensic Science International, 2011. **208**(1-3): p. 10-17.
12. Biedermann, A., S. Bozza, and F. Taroni, *The Decisionalization of Individualization.* Forensic Science International, 2016. **266**: p. 29-38.
13. Langenburg, G., *A Critical Analysis and Study of the ACE-V Process*. 2012, PhD thesis, Ecole des Sciences Criminelles, Faculté de droit et des Sciences Criminelles, Université de Lausanne.
14. Kalka, N.D., M. Beachler, and R.A. Hicklin, *LQMetric: A Latent Fingerprint Quality Metric for Predicting AFIS Performance and Assessing the Value of Latent Fingerprints.* Journal of Forensic Identification, 2020. **70**(4): p. 443-463.
15. Hicklin, A.R., *Improving the Rigor of the Latent Print Examination Process*. 2017, PhD thesis, Ecole des Sciences Criminelles, Faculté de droit et des Sciences Criminelles, Université de Lausanne.
16. Eldridge, H., *Understanding, Expanding, and Predicting the Suitability Decision in Friction Ridge Analysis*. 2020, PhD thesis, Ecole des Sciences Criminelles, Faculté de droit, des Sciences Criminelles et d'Administration Publique.
17. Alonso-Fernandez, F., J. Fierrez-Aguilar, and J. Ortega-Garcia. *A Review of Schemes for Fingerprint Image Quality Computation*. in *3rd COST-275 Workshop on Biometrics on the*

Internet, COST-275, Hatfield, United Kingdom, 27-28 October, 2005. 2005. EU Publications Office (OPOCE).
18. Nill, N.B., *IQF (Image Quality of Fingerprint) Software Application*. 2007, MITRE Corp: Bedford, M.A., USA.
19. Fronthaler, H., K. Kollreider, J. Bigun, J. Fierrez, F. Alonso-Fernandez, J. Ortega-Garcia, et al., *Fingerprint Image-Quality Estimation and its Application to Multialgorithm Verification.* IEEE Transactions on Information Forensics and Security, 2008. **3**(2): p. 331-338.
20. Hicklin, R.A., J. Buscaglia, M.A. Roberts, S.B. Meagher, W. Fellner, M.J. Burge, et al., *Latent Fingerprint Quality: A Survey of Examiners.* Journal of Forensic Identification, 2011. **61**(4): p. 385-419.
21. Murch, R.S., A.L. Abbott, E.A. Fox, M.S. Hsiao, and B. Budowle, *Establishing the Quantitative Basis for Sufficiency Thresholds and Metrics for Friction Ridge Pattern Detail and the Foundation for a Standard.* US Department of Justice, Washington DC, 2012.
22. Yoon S, L.E., Jain A, *On Latent Fingerprint Image Quality*, in *Computational Forensics: 5th International Workshop, IWCF 2012 and 2014*, U. Garain and F. Shafait, Editors. 2015, Springer International Publishing: Cham. p. 67-82.
23. Hicklin, R.A., J. Buscaglia, and M.A. Roberts, *Assessing the Clarity of Friction Ridge Impressions.* Forensic Science International, 2013. **226**(1-3): p. 106-117.
24. Bryson, S.J., *American National Standard for Information Systems-Data Format for the Interchange of Fingerprint, Facial & Other Biometric Information.* NIST Special Publication. **500**: p. 290.
25. Sankaran, A., M. Vatsa, and R. Singh. *Automated Clarity and Quality Assessment for Latent Fingerprints*. in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. 2013. Institute of Electrical and Electronics Engineers (IEEE).
26. Pulsifer, D.P., S.A. Muhlberger, S.F. Williams, R.C. Shaler, and A. Lakhtakia, *An Objective Fingerprint Quality-Grading System.* Forensic Science International, 2013. **231**(1-3): p. 204-207.
27. Kellman, P.J., J.L. Mnookin, G. Erlikhman, P. Garrigan, T. Ghose, E. Mettler, et al., *Forensic Comparison and Matching of Fingerprints: Using Quantitative Image Measures for Estimating Error Rates through Understanding and Predicting Difficulty.* PloS one, 2014. **9**(5): p. 1-14.
28. Chugh, T., K. Cao, J. Zhou, E. Tabassi, and A.K. Jain, *Latent Fingerprint Value Prediction: Crowd-Based Learning.* IEEE Transactions on Information Forensics and Security, 2017. **13**(1): p. 20-34.
29. Neumann, C., C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz, D. Meuwly, et al., *Computation of Likelihood Ratios in Fingerprint Identification for Configurations of Three Minutiae.* Journal of Forensic Sciences, 2006. **51**(6): p. 1255-1266.
30. Zhu, Y., S.C. Dass, and A.K. Jain, *Statistical Models for Assessing the Individuality of Fingerprints.* IEEE Transactions on Information Forensics and Security, 2007. **2**(3): p. 391-401.
31. Egli, N.M., C. Champod, and P. Margot, *Evidence Evaluation in Fingerprint Comparison and Automated Fingerprint Identification Systems—Modelling Within Finger Variability.* Forensic Science International, 2007. **167**(2-3): p. 189-195.

32. Neumann, C., C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz, and A. Bromage-Griffiths, *Computation of Likelihood Ratios in Fingerprint Identification for Configurations of Any Number of Minutiae.* Journal of Forensic Sciences, 2007. **52**(1): p. 54-64.
33. Egli Nicole, M., *Interpretation of Partial Fingermarks Using an Automated Fingerprint Identification System*. 2009, PhD thesis, Ecole des Sciences Criminelles, Faculté de droit et des Sciences Criminelles, Université de Lausanne.
34. Su, C. and S. Srihari, *Evaluation of Rarity of Fingerprints in Forensics.* Advances in Neural Information Processing Systems, 2010. **23**: p. 1207-1215.
35. Lim, C.Y. and S.C. Dass, *Assessing Fingerprint Individuality Using EPIC: A Case Study in the Analysis of Spatially Dependent Marked Processes.* Technometrics, 2011. **53**(2): p. 112-124.
36. Choi, H., A. Nagar, and A.K. Jain. *On the Evidential Value of Fingerprints*. in *2011 International Joint Conference on Biometrics (IJCB)*. 2011. Institute of Electrical and Electronics Engineers (IEEE).
37. Neumann, C., I. Evett, and J. Skerrett, *Quantifying the Weight of Evidence from a Forensic Fingerprint Comparison: A New Paradigm.* Journal of the Royal Statistical Society: Series A (Statistics in Society), 2012. **175**(2): p. 371-415.
38. Neumann, C., I.W. Evett, J.E. Skerrett, and I. Mateos-Garcia, *Quantitative Assessment of Evidential Weight for a Fingerprint Comparison. Part II: A Generalisation to Take Account of the General Pattern.* Forensic Science International, 2012. **214**(1-3): p. 195-199.
39. Abraham, J., C. Champod, C. Lennard, and C. Roux, *Spatial Analysis of Corresponding Fingerprint Features from Match and Close Non-Match Populations.* Forensic Science International, 2013. **230**(1-3): p. 87-98.
40. Alberink, I., A. de Jongh, and C. Rodriguez, *Fingermark Evidence Evaluation Based on Automated Fingerprint Identification System Matching Scores: The Effect of Different Types of Conditioning on Likelihood Ratios.* Journal of Forensic Sciences, 2014. **59**(1): p. 70-81.
41. Anthonioz, N.E. and C. Champod, *Evidence Evaluation in Fingerprint Comparison and Automated Fingerprint Identification Systems—Modeling Between Finger Variability.* Forensic Science International, 2014. **235**: p. 86-101.
42. Neumann, C., C. Champod, M. Yoo, T. Genessay, and G. Langenburg, *Quantifying the Weight of Fingerprint Evidence through the Spatial Relationship, Directions and Types of Minutiae Observed on Fingermarks.* Forensic Science International, 2015. **248**: p. 154-171.
43. Leegwater, A.J., D. Meuwly, M. Sjerps, P. Vergeer, and I. Alberink, *Performance Study of a Score-Based Likelihood Ratio System for Forensic Fingermark Comparison.* Journal of Forensic Sciences, 2017. **62**(3): p. 626-640.
44. Yoon, S., K. Cao, E. Liu, and A.K. Jain. *LFIQ: Latent Fingerprint Image Quality*. in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. 2013. Institute of Electrical and Electronics Engineers (IEEE).
45. Champod, C. and I.W. Evett, *A Probabilistic Approach to Fingerprint Evidence.* Journal of Forensic Identification, 2001. **51**(2): p. 101-122.
46. McKasson, S., *I Think Therefore I Probably Am.* Journal of Forensic Identification, 2001. **51**(3): p. 217-221.

47. Bush, L., *In Support of Fingerprint Evidence.* Journal of Forensic Identification, 2001. **51**(5): p. 457-460.
48. Cordle, M. and A.J. Morlan, *Letter to the Editor.* Journal of Forensic Identification, 2001. **51**(3): p. 684-685.
49. Swofford, H., C. Champod, A. Koertner, H. Eldridge, and M. Salyards, *A Method for Measuring the Quality of Friction Skin Impression Evidence: Method Development and Validation.* Forensic Science International, 2021. **320**: p. 110703.
50. Swofford, H.J., A.J. Koertner, F. Zemp, M. Ausdemore, A. Liu, and M.J. Salyards, *A Method for the Statistical Interpretation of Friction Ridge Skin Impression Evidence: Method Development and Validation.* Forensic Science International, 2018. **287**: p. 113-126.
51. Swofford, H.J., S.A. Cole, and V. King, *"Mt. Everest—We are Going to Lose Many": A Survey of Fingerprint Examiners' Attitudes Toward Probabilistic Reporting.* Law, Probability and Risk, 2021. **19**(3-4): p. 255-291.
52. Swofford, H. and C. Champod, *Probabilistic Reporting and Algorithms in Forensic Science: Stakeholder Perspectives within the American Criminal Justice System.* Forensic Science International: Synergy, 2022: p. 100220.
53. Swofford, H. and C. Champod, *Implementation of Algorithms in Pattern & Impression Evidence: A Responsible and Practical Roadmap.* Forensic Science International: Synergy, 2021: p. 100142.
54. Choong, Y., F. Rakebrandt, R.V. North, and J.E. Morgan, *Acutance, An Objective Measure of Retinal Nerve Fibre Image Clarity.* British Journal of Ophthalmology, 2003. **87**(3): p. 322-326.
55. Moore, R., *An Analysis of Ridge-to-Ridge Distance on Fingerprints.* Journal of Forensic Identification, 1989. **39**(4): p. 231-238.
56. Langenburg, G. and C. Champod, *The GYRO System-A Recommended Approach to More Transparent Documentation.* Journal of Forensic Identification, 2011. **61**(4): p. 373-384.
57. Eldridge, H., M. DeDonno, J. Furrer, and C. Champod, *Examining and Expanding the Friction Ridge Value Decision.* Forensic Science International, 2020. **314**: p. 110408.
58. Team, R.C., *R: A Language and Environment for Statistical Computing (Version 3.6.3, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/).* 2018.
59. Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, et al., *pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves.* BMC Bioinformatics, 2011. **12**(1): p. 1-8.
60. John, J. and H. Swofford, *Evaluating the Accuracy and Weight of Confidence in Examiner Minutiae Annotations.* Journal of Forensic Identification, 2020. **70**(3): p. 289-309.
61. Stoney, D.A., M. De Donno, C. Champod, P.A. Wertheim, and P.L. Stoney, *Occurrence and Associative Value of Non-Identifiable Fingermarks.* Forensic Science International, 2020. **309**: p. 110219.
62. Team, R.C., *R: A Language and Environment for Statistical Computing (Version 4.0.2, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/).* 2020.
63. Kuhn, H.W., *The Hungarian Method for the Assignment Problem.* Naval Research Logistics Quarterly, 1955. **2**(1-2): p. 83-97.
64. Fagert, M. and K. Morris, *Quantifying the Limits of Fingerprint Variability.* Forensic Science International, 2015. **254**: p. 87-99.

65. Garris, M.D. and R.M. McCabe, *NIST Special Database 27: Fingerprint Minutiae From Latent and Matching Tenprint Images*. 2000: US Department of Commerce, National Institute of Standards and Technology.
66. Ulery, B.T., R.A. Hicklin, J. Buscaglia, and M.A. Roberts, *Accuracy and Reliability of Forensic Latent Fingerprint Decisions.* Proceedings of the National Academy of Sciences, 2011. **108**(19): p. 7733-7738.
67. Ulery, B.T., R.A. Hicklin, M.A. Roberts, and J. Buscaglia, *Measuring What Latent Fingerprint Examiners Consider Sufficient Information for Individualization Determinations.* PloS one, 2014. **9**(11): p. e110179.
68. Bookstein, F.L., *Principal Warps: Thin-Plate Splines and the Decomposition of Deformations.* IEEE Transactions on pattern analysis and machine intelligence, 1989. **11**(6): p. 567-585.
69. *Ballistics Toolmark Research Database*. 2016, National Institute of Standards and Technology: https://www.nist.gov/programs-projects/nist-ballistics-toolmark-database.
70. CSAFE, *Forensic Science Data Portal. Center for Statistical Applications in Forensic Evidence. https://forensicstats.org/data/*. (n.d.).
71. Bush, L., *The Authority of Fingerprint Experts: Is It Based on Belief or Science?* Journal of Forensic Identification, 2009. **59**(6): p. 599-608.
72. Jayaprakash, P.T., *Practical Relevance of Pattern Uniqueness in Forensic Science.* Forensic Science International, 2013. **231**(1-3): p. 403.e1-403.e16.
73. Lieberman, J.D., C.A. Carrell, T.D. Miethe, and D.A. Krauss, *Gold Versus Platinum: Do Jurors Recognize the Superiority and Limitations of DNA Evidence Compared to Other Types of Forensic Evidence?* Psychology, Public Policy, and Law, 2008. **14**(1): p. 27-62.
74. Garrett, B. and G. Mitchell, *How Jurors Evaluate Fingerprint Evidence: The Relative Importance of Match Language, Method Information, and Error Acknowledgment.* Journal of Empirical Legal Studies, 2013. **10**(3): p. 484-511.
75. Koehler, J.J., *Intuitive Error Rate Estimates for the Forensic Sciences.* Jurimetrics, 2017: p. 153-168.
76. Ribeiro, G., J.M. Tangen, and B.M. McKimmie, *Beliefs About Error Rates and Human Judgment in Forensic Science.* Forensic Science International, 2019. **297**: p. 138-147.
77. van Straalen, E.K., C.J. de Poot, M. Malsch, and H. Elffers, *The Interpretation of Forensic Conclusions by Criminal Justice Professionals: The Same Evidence Interpreted Differently.* Forensic Science International, 2020. **313**: p. 110331.
78. Cole, S.A., *Suspect Identities: A History of Fingerprinting and Criminal Identification*. 2001, Cambridge: Harvard University Press.
79. Ashbaugh, D.R., *Quantitative-Qualitative Friction Ridge Analysis: An Introduction to Basic and Advanced Ridgeology*. 1999, Boca Raton: CRC press.
80. Bali, A.S., G. Edmond, K.N. Ballantyne, R.I. Kemp, and K.A. Martire, *Communicating Forensic Science Opinion: An Examination of Expert Reporting Practices.* Science & Justice, 2020. **60**(3): p. 216-224.
81. Cole, S.A., *Where the Rubber Meets the Road: Thinking About Expert Evidence as Expert Testimony.* Vill. L. Rev., 2007. **52**: p. 803-842.
82. Robertson, B.W., *Fingerprints, Relevance and Admissibility.* NZ Recent L. Rev., 1990: p. 252-258.
83. Stoney, D.A., *What Made Us Ever Think We Could Individualize Using Statistics?* Journal of the Forensic Science Society, 1991. **31**(2): p. 197-199.

84. Broeders, A., *Of Earprints, Fingerprints, Scent Dogs, Cot Deaths and Cognitive Contamination—A Brief Look at the Present State of Play in the Forensic Arena.* Forensic Science International, 2006. **159**(2-3): p. 148-157.
85. Meuwly, D., *Forensic Individualisation From Biometric Data.* Science & Justice, 2006. **46**(4): p. 205-213.
86. Saks, M.J. and J.J. Koehler, *The Individualization Fallacy in Forensic Science Evidence.* Vand. L. Rev., 2008. **61**: p. 199-219.
87. Cole, S.A., *Forensics Without Uniqueness, Conclusions Without Individualization: The New Epistemology of Forensic Identification.* Law, Probability and Risk, 2009. **8**(3): p. 233-255.
88. Page, M., J. Taylor, and M. Blenkin, *Uniqueness in the Forensic Identification Sciences—Fact or Fiction?* Forensic Science International, 2011. **206**(1-3): p. 12-18.
89. Eldridge, H., *The Shifting Landscape of Latent Print Testimony: An American Perspective.* Journal of Forensic Science and Medicine, 2017. **3**(2): p. 72-81.
90. Campbell, A., *The Fingerprint Inquiry Report*. 2011, Edinburgh, UK: APS Group Scotland.
91. Aitken, C., P. Roberts, and G. Jackson, *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings (Practitioner Guide No. 1).* Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses, Royal Statistical Society's Working Group on Statistics and the Law, 2010. **42**.
92. Garrett, R., *IAI Letter RE: NAS Report to The Honorable Patrick J. Leahy, Chairman, Senate Committee on the Judiciary*. 2009.
93. OSAC-FRS, *OSAC FRS Comment on FR Doc # NA*, in *Comment ID: DOJ-OLP-2016-0012-0067*. 2016, Friction Ridge Subcommittee (FRS), Organization of Scientific Area Committees (OSAC) for Forensic Science: https://www.regulations.gov/comment/DOJ-OLP-2016-0012-0067.
94. Cole, S.A., *A Discouraging Omen: A Critical Evaluation of the Approved Uniform Language for Testimony and Reports for the Forensic Latent Print Discipline.* Ga. St. UL Rev., 2018. **34**(4): p. 1103-1128.
95. Department of the Army Defense Forensic Science Center, *INFORMATION PAPER, Subject: Use of the Term "Identification" in Latent Print Technical Reports*. 2015.
96. Department of the Army Defense Forensic Science Center, *INFORMATION PAPER, Subject: Modification of Latent Print Technical Reports to Include Statistical Calculations*. 2017.
97. OSAC-FRS, *Standard for Friction Ridge Examination Conclusions*. 2018, Friction Ridge Subcommittee (FRS), Organization of Scientific Area Committees (OSAC) for Forensic Science.
98. Triplett, M., *Complexity, Level of Association and Strength of Fingerprint Conclusions.* Journal of Cold Case Review, 2015. **1**(2): p. 6-15.
99. Triplett, M., *Fingerprint Examination: A Defined Method (Ver. 3).* Michele Triplett's Fingerprint Information, 2018.
100. Evett, I., *The Logical Foundations of Forensic Science: Towards Reliable Knowledge.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2015. **370**(1674): p. 20140263.
101. Evett, I., *Avoiding the Transposed Conditional.* Science & Justice, 1995. **35**(2): p. 127-132.

102. Morrison, G.S., F.H. Sahito, G. Jardine, D. Djokic, S. Clavet, S. Berghs, et al., *The International Criminal Police Organization (INTERPOL) Survey of the Use of Speaker Identification by Law Enforcement Agencies.* Forensic Science International, 2016. **263**: p. 92-100.
103. Langenburg, G., C. Neumann, S.B. Meagher, C. Funk, and J.P. Avila, *Presenting Probabilities in the Courtroom: A Moot Court Exercise.* Journal of Forensic Identification, 2013. **63**(4): p. 424-488.
104. Thompson, W.C. and E.J. Newman, *Lay Understanding of Forensic Statistics: Evaluation of Random Match Probabilities, Likelihood Ratios, and Verbal Equivalents.* Law and Human Behavior, 2015. **39**(4): p. 332-349.
105. Bayer, D., C. Neumann, and A. Ranadive, *Communication of Statistically Based Conclusions to Jurors–A Pilot Study.* Journal of Forensic Identification, 2016. **66**(5): p. 405-427.
106. Garrett, B., G. Mitchell, and N. Scurich, *Comparing Categorical and Probabilistic Fingerprint Evidence.* Journal of Forensic Sciences, 2018. **63**(6): p. 1712-1717.
107. Thompson, W.C., R.H. Grady, E. Lai, and H.S. Stern, *Perceived Strength of Forensic Scientists' Reporting Statements About Source Conclusions.* Law, Probability and Risk, 2018. **17**(2): p. 133-155.
108. Eldridge, H., *Juror Comprehension of Forensic Expert Testimony: A Literature Review and Gap Analysis.* Forensic Science International: Synergy, 2019. **1**: p. 24-34.
109. Wells, G.L., *Naked Statistical Evidence of Liability: Is Subjective Probability Enough?* Journal of Personality and Social Psychology, 1992. **62**(5): p. 739-752.
110. Willis, S., L. McKenna, S. McDermott, G. O'Donell, A. Barrett, B. Rasmusson, et al., *ENFSI Guideline for Evaluative Reporting in Forensic Science. European Network of Forensic Science Institutes*. 2015.
111. Resolutions & Legislative Committee, *Resolution VII.* International Association for Identification, 1979.
112. Resolutions & Legislative Committee, *Resolution V.* International Association for Identification, 1980.
113. Resolutions & Legislative Committee, *Resolution 2010-18.* International Association for Identification, 2010.
114. Imwinkelried, E.J., *Computer Source Code: A Source of the Growing Controversy Over the Reliability of Automated Forensic Techniques.* DePaul L. Rev., 2016. **66**: p. 97-132.
115. Kwong, K., *The Algorithm Says You Did It: The Use of Black Box Algorithms to Analyze Complex DNA Evidence.* Harv. JL & Tech., 2017. **31**: p. 275-301.
116. Roth, A., *Machine Testimony.* Yale LJ, 2016. **126**: p. 1972-2053.
117. Cino, J.G., *Deploying the Secret Police: The Use of Algorithms in the Criminal Justice System.* Ga. St. UL Rev., 2017. **34**(4): p. 1073-1102.
118. Nutter, P.W., *Machine Learning Evidence: Admissibility and Weight.* U. Pa. J. Const. L., 2019. **21**(3): p. 919-958.
119. Osoba, O.A., B. Boudreaux, J.M. Saunders, J.L. Irwin, P.A. Mueller, and S. Cherney, *Algorithmic Equity: A Framework for Social Applications*. 2019, Santa Monica, CA: RAND.
120. Završnik, A. *Criminal Justice, Artificial Intelligence Systems, and Human Rights*. in *ERA Forum*. 2020. Springer.

121. Kafadar, K., *The Roles of Science and Statistics in Advancing Forensic Science Standards.* Testimony to the House Committee on Science, Space, and Technology, United States Congress, 2019.
122. Kafadar, K., *The Need for Objective Measures in Forensic Evidence.* Significance, 2019. **16**(2): p. 16-20.
123. Committee on Rules of Practice and Procedure, *Preliminary Draft: Proposed Amendments to the Federal Rules of Appellate, Bankruptcy, Civil, and Criminal Procedure, and the Federal Rules of Evidence. Judicial Conference of the United States.* 2021.
124. Reisman, D., J. Schultz, K. Crawford, and M. Whittaker, *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability.* AI Now Institute, 2018: p. 1-22.
125. United States, *Justice in Forensic Algorithms Act of 2019. H.R. 4368.* 2019.
126. Descript, *Machine Transcription Software, San Francisco, CA, USA. https://www.descript.com/.* 2021.
127. Forensic Science Regulator, *Development of Evaluative Opinions*. Codes of Practices and Conduct, 2021. **FSR-C-118**(1).
128. Thompson, W.C., *How Should Forensic Scientists Present Source Conclusions*. The Seton Hall Law Review, 2018. **48**: p. 773-813.
129. *United States v. Llera-Plaza, 188 F. Supp. 2d 549 (E.D. Pa. 2002).*
130. *Johnson v. Commonwealth, 12 S.W.3d 258 (KY. 2000).*
131. Ramos, M., *Reference: Report Entitled "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods".* National District Attorneys Association, 2016.
132. ASCLD, *Statement on September 20, 2016 PCAST Report on Forensic Science.* American Society of Crime Laboratory Directors, 2016.
133. AFTE, *Response to PCAST Report on Forensic Science.* Association of Firearm and Toolmark Examiners, 2016.
134. NACDL, *President's Council of Advisors on Science and Technology (PCAST) Issues Major Forensic Science Report; Calls for Stronger Scientific Standards.* National Association of Criminal Defense Lawyers, 2016.
135. *Innocence Project Applauds President Obama's Science Advisors' Landmark Report Calling for Essential Improvements to Forensic Disciplines.* Innocence Project, 2016.
136. Hill, K., *Wrongfully Accused by an Algorithm*, in *The New York Times*. 2020: https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.
137. Persons, T.M. and M. Mackin, *Technology Readiness Assessment Guide: Best Practices for Evaluating the Readiness of Technology for Use in Acquisition Programs and Projects*. 2020, US Government Accountability Office: Washington, D.C., USA.
138. Meehl, P.E., *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. 1954, Minneapolis: University of Minnesota Press.
139. Kahneman, D., *Thinking, Fast and Slow*. 2011, New York: Macmillan.
140. Grove, W.M. and P.E. Meehl, *Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical–Statistical Controversy.* Psychology, Public Policy, and Law, 1996. **2**(2): p. 293-323.

141. Grove, W.M., D.H. Zald, B.S. Lebow, B.E. Snitz, and C. Nelson, *Clinical Versus Mechanical Prediction: A Meta-Analysis.* Psychological Assessment, 2000. **12**(1): p. 19-30.
142. Meehl, P.E., *Causes and Effects of My Disturbing Little Book.* Journal of Personality Assessment, 1986. **50**(3): p. 370-375.
143. Dawes, R.M., *The Robust Beauty of Improper Linear Models in Decision Making.* American Psychologist, 1979. **34**(7): p. 571-582.
144. Highhouse, S., *Stubborn Reliance on Intuition and Subjectivity in Employee Selection.* Industrial and Organizational Psychology, 2008. **1**(3): p. 333-342.
145. Guyatt, G., J. Cairns, D. Churchill, D. Cook, B. Haynes, J. Hirsh, et al., *Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine.* Jama, 1992. **268**(17): p. 2420-2425.
146. Zimerman, A.L., *Evidence-Based Medicine: A Short History of a Modern Medical Movement.* AMA Journal of Ethics, 2013. **15**(1): p. 71-76.
147. Guyatt G., R.D., Meade M.O., Cook D.J., *Users' Guide to the Medical Literature: A Manual for the Evidence-Based Clinical Practice*. 2nd ed. 2008, New York: McGraw-Hill.
148. Timmermans, S. and A. Mauck, *The Promises and Pitfalls of Evidence-Based Medicine.* Health Affairs, 2005. **24**(1): p. 18-28.
149. Tonelli, M.R., *In Defense of Expert Opinion.* Academic Medicine: Journal of the Association of American Medical Colleges, 1999. **74**(11): p. 1187-1192.
150. Cohen, A.M., P.Z. Stavri, and W.R. Hersh, *A Categorization and Analysis of the Criticisms of Evidence-Based Medicine.* International Journal of Medical Informatics, 2004. **73**(1): p. 35-43.
151. Dietvorst, B.J., J.P. Simmons, and C. Massey, *Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err*. Journal of Experimental Psychology: General, 2015. **144**(1): p. 114-126.
152. Logg, J.M., J.A. Minson, and D.A. Moore, *Algorithm Appreciation: People Prefer Algorithmic to Human Judgment.* Organizational Behavior and Human Decision Processes, 2019. **151**: p. 90-103.
153. Arkes, H.R., R.M. Dawes, and C. Christensen, *Factors Influencing the Use of a Decision Rule in a Probabilistic Task.* Organizational Behavior and Human Decision Processes, 1986. **37**(1): p. 93-110.
154. Dietvorst, B.J., J.P. Simmons, and C. Massey, *Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms if They Can (Even Slightly) Modify Them.* Management Science, 2018. **64**(3): p. 1155-1170.
155. Kleinmuntz, B., *Why We Still Use Our Heads Instead of Formulas: Toward an Integrative Approach.* Psychological Bulletin, 1990. **107**(3): p. 296-310.
156. Dawes, R.M., D. Faust, and P.E. Meehl, *Clinical Versus Actuarial Judgment.* Science, 1989. **243**(4899): p. 1668-1674.
157. *SAE Standard J 3016_201806: Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems*. 2018, SAE International.
158. National Highway Traffic Safety Administration (NHTSA), *2016 Fatal Motor Vehicle Crashes: Overview*, in *Traffic Safety Facts*. 2017, United States Department of Transportation.

159. Schoettle, B. and M. Sivak, *Motorists' Preferences for Different Levels of Vehicle Automation: 2016,* in *Report No. SWT-2016-8*. 2016, University of Michigan: Sustainable Worldwide Transportation.
160. Abraham, H., C. Lee, S. Brady, C. Fitzgerald, B. Mehler, B. Reimer, et al. *Autonomous Vehicles and Alternatives to Driving: Trust, Preferences, and Effects of Age*. in *Proceedings of the Transportation Research Board 96th Annual Meeting (TRB'17)*. 2017. Washington, D.C., USA.
161. NTSB, *Collision Between a Car Operating With Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida, May 7, 2016*, in *Accident Report, Report No. NTSB/HAR-17/02, PB2017-102600*. 2017, National Transportation Safety Board (NTSB): Washington, D.C., USA.
162. Kalra, N. and D.G. Groves, *The Enemy of Good: Estimating the Cost of Waiting for Nearly Perfect Automated Vehicles*. 2017: Rand Corporation.
163. Jing, P., G. Xu, Y. Chen, Y. Shi, and F. Zhan, *The Determinants Behind the Acceptance of Autonomous Vehicles: A Systematic Review.* Sustainability, 2020. **12**(5): p. 1719-1745.
164. *Summary Report of the G7 Multistakeholder Conference on Artificial Intelligence*. 2018: Montreal, Canada.
165. Simonite, T., *Google's AI Guru Wants Computers to Think More Like Brains*, in *Wired*. 2018: https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/.
166. Jones, H., *Geoff Hinton Dismissed the Need For Explainable AI: 8 Experts Explain Why He's Wrong*, in *Forbes*. 2018: https://www.forbes.com/sites/cognitiveworld/2018/12/20/geoff-hinton-dismissed-the-need-for-explainable-ai-8-experts-explain-why-hes-wrong/#40bef765756d.
167. *Artificial Intelligence and Robotics for Law Enforcement*. 2019, United Nations Interregional Crime and Justice Research Institute (UNICRI) and The International Criminal Police Organization (INTERPOL).
168. *Toward Responsible AI Innovation: Second INTERPOL-UNICRI Report on Artificial Intelligence for Law Enforcement*. 2020, United Nations Interregional Crime and Justice Research Institute (UNICRI) and The International Criminal Police Organization (INTERPOL).
169. Dupont, B., Y. Stevens, H. Westermann, and M. Joyce, *Artificial Intelligence in the Context of Crime and Criminal justice*. 2018: Université de Montréal.
170. *Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems.* AI Now Institute, 2018.
171. Richardson, R., J.M. Schultz, and V.M. Southerland, *Litigating Algorithms 2019 US Report: New Challenges to Government Use of Algorithmic Decision Systems.* AI Now Institute, 2019.
172. *Frye v. United States*. 293 F. 1013 (D.C. Cir. 1923).
173. *Daubert v. Merrell Dow Pharmaceuticals, Inc.* 509 U.S. 579 (1993).
174. *General Electric Co. v. Joiner*. 522 U.S. 136 (1997).
175. *Kumho Tire Co. v. Carmichael*. 526 U.S. 137 (1999).
176. *Fed. R. Evid. Rule 901(b)(9)*.
177. *Fed. R. Evid. Rule 702*.
178. *U.S. Const. amend. V*.
179. *U.S. Const. amend. VI*.

180. Langenburg, G., C. Champod, and T. Genessay, *Informing the Judgments of Fingerprint Analysts Using Quality Metric and Statistical Assessment Tools.* Forensic Science International, 2012. **219**(1-3): p. 183-198.
181. *Models & Algorithms in Forensic Science: Recommendations for Innovation*. 2013, Horsham, UK: Models & Algorithms Advisory Group, Forensic Science Special Interest Group (FoSciSIG), Knowledge Transfer Network.
182. Tully, G., K. Sullivan, A. Vidaki, and A. Anjomshoaa, *Taking Forensic Science R&D to Market*. 2013, Horsham, UK: Forensic Science Special Interest Group (FoSciSIG), Knowledge Transfer Network.
183. *Software Validation for DNA Mixture Interpretation (FSR-G-223)*. 2018, Forensic Science Regulator: Birmingham, UK.
184. Meuwly, D., D. Ramos, and R. Haraksim, *A Guideline for the Validation of Likelihood Ratio Methods Used for Forensic Evidence Evaluation.* Forensic Science International, 2017. **276**: p. 142-153.
185. Dror, I.E. and J.L. Mnookin, *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, 2010. **9**(1): p. 47-67.
186. Garrett, B.L., W.E. Crozier, and R. Grady, *Error Rates, Likelihood Ratios, and Jury Evaluation of Forensic Evidence.* Journal of Forensic Sciences, 2020. **65**(4): p. 1199-1209.
187. Montani, I., R. Marquis, N.E. Anthonioz, and C. Champod, *Resolving Differing Expert Opinions.* Science & Justice, 2019. **59**(1): p. 1-8.
188. Swofford, H.J., Chu, S., *Facilitating a Culture of Improvement in a Safe and Non-Adversarial Environment*, in *100th Annual Conference of the International Association for Identification*. 2015: Sacramento, C.A., USA.
189. *Position Description: Physical Scientist (Forensic Latent Print Examiner), PD#: DZ392452*. 2012, Department of the Army, United Sates Government: USA.
190. Swofford, H.J., *Individualization Using Friction Skin Impressions: Scientifically Reliable, Legally Valid.* Journal of Forensic Identification, 2012. **62**(1): p. 62-79.
191. Swofford, H., *The Emerging Paradigm Shift in the Epistemology of Fingerprint Conclusions.* Journal of Forensic Identification, 2015. **65**(3): p. 201-213.
192. Swofford, H.J., A.J. Koertner, and J. Salyards, *Evaluation and Validation of a Model to Quantify the Weight of Fingerprint Evidence*, in *68th Annual Conference of the American Academy of Forensic Sciences*. 2016: Las Vegas, N.V., USA.
193. Swofford, H.J., *Integrating Statistical Thinking and Methods into Practice – Latent Print Examination*, in *Center for Statistics and Applications to Forensic Evidence Florida Pattern Science and Statistics Workshop*. 2016: West Palm Beach, F.L., USA.
194. Swofford, H.J., Koertner, A.J., Salyards, M.J. , *Development and Evaluation of a Model to Quantify the Weight of Fingerprint Evidence*, in *2016 Statistical and Applied Mathematical Sciences Institute Program on Pattern Evidence Transition Workshop*. 2016: Research Triangle Park, N.C., USA.
195. Swofford, H.J., *Towards Integrating Probabilistic Logic and Quantitative Data into Practice – Latent Print Examination*, in *National Institute of Standards and Technology Technical Colloquium on Quantifying the Weight of Forensic Evidence*. 2016: Gaithersburg, M.D., USA.

196. Swofford, H.J., A.J. Koertner, and M.J. Salyards, *Evaluation of a Model to Quantify the Weight of Fingerprint Evidence*, in *101st Annual Conference of the International Association for Identification*. 2016: Cincinnati, O.H., USA.
197. Swofford, H.J., *Integrating Quantitative Approaches into Practice: Method Development, Validation and Implementation*, in *International Symposium on Likelihood Ratio Methods*. 2016: Netherlands Forensic Institute, The Hague, Netherlands.
198. Swofford, H.J., *Towards Reform: Implementing Quantitative Methods Into Practice for Latent Print Examination*, in *69th Annual Conference of the American Academy of Forensic Sciences*. 2017: New Orleans, L.A., USA.
199. Swofford, H.J., *Towards Reform: Demonstrating Validity in Fingerprint Examinations*, in *Forensic Conference presented by the Cook County Public Defender's Office and Loyola Law School*. 2017: Chicago, I.L., USA.
200. Swofford, H.J., Wortman, T. M., *Technology Transition Workshop: Friction Ridge Statistical Interpretation Software (FRStat)*, in *Arizona Forensic Science Speaker Series*. 2017: Phoenix, A.Z., USA.
201. Swofford, H.J., Koertner, A. J., *Technology Transition Workshop: Friction Ridge Statistical Interpretation Software (FRStat)*, in *New York Office of Criminal Justice Services Forensic Training*. 2017: Albany, N.Y., USA.
202. Swofford, H.J., *Statistical Interpretation Software (FRStat)*, in *Internal Revenue Service Forensic Training*. 2017: Chicago, I.L., USA.
203. Swofford, H.J., *Friction Ridge Statistical Interpretation Software (FRStat)*, in *Federal Bureau of Investigation Forensic Training*. 2017: Quantico, V.A., USA.
204. Swofford, H.J., *Towards Reform: Demonstrating Validity in Fingerprint Examinations*, in *Forensic Conference presented by the National Association of Criminal Defense Lawyers and Cardozo Law School*. 2017: New York City, N.Y., USA.
205. Swofford, H.J., LeCroy, J.E., *So We Implemented a Statistical Model . . . What Happened Next?*, in *102nd Annual Conference of the International Association for Identification*. 2017: Atlanta, G.A., USA.
206. Swofford, H.J., *Litigating Fingerprint Evidence: Ensuring a Sound Scientific Foundation*, in *Minnesota Public Defenders Professional Development Meeting*. 2018: Brainerd, M.N., USA.
207. Swofford, H.J., *"Identification" Roundtable (Panel Discussion)*, in *18th Annual Meeting of the European Network of Forensic Science Institutes (ENFSI) Fingerprint Working Group*. 2018: Lausanne, Switzerland.
208. Swofford, H.J., *Litigating Fingerprint Evidence: Ensuring a Sound Scientific Foundation*, in *Forensic Conference of the National Association of Criminal Defense Lawyers and Cardozo Law School*. 2018: New York City, N.Y., USA.
209. Swofford, H.J., Hunt, T., King, P., Stolorow, M., Loudon-Brown, M., *To Err is Human (Panel Discussion)*, in *From the Crime Scene to the Courtroom: The Future of Forensic Science Reform*. 2018, Georgia State University College of Law: Atlanta, G.A., USA.
210. Swofford, H.J., *Statistics in the Crime Lab: Towards Reform – Implementing Stronger Scientific Foundations, Statistical Thinking, and Cultural Change*, in *Center for Statistics and Applications in Forensic Science Forensics, Statistics, and the Law Conference*. 2018, University of Virginia: Charlottesville, V.A., USA.

211. Swofford, H.J., Koertner, A.J., Wortman, T.M. , *Statistical Interpretation Software for Friction Ridge Skin Impressions (FRStat) Workshop*, in *103rd Annual Conference of the International Association for Identification*. 2018: San Antonio, T.X., USA.
212. Swofford, H.J., LeCroy, J.E., *Statistical Interpretation and Reporting of Fingerprint Evidence at the United States Army Criminal Investigation Laboratory*, in *103rd Annual Conference of the International Association for Identification*. 2018: San Antonio, T.X., USA.
213. Swofford, H.J., *Statistical Interpretation and Reporting of Fingerprint Evidence at the United States Army Criminal Investigation Laboratory*, in *70th Annual Conference of the American Academy of Forensic Sciences*. 2018: Seattle, W.A., USA.
214. Swofford, H.J., *Statistical Interpretation and Reporting of Fingerprint Evidence at the United States Army Criminal Investigation Laboratory*, in *Impression, Pattern, and Trace Evidence Symposium*. 2018: Washington, D.C., USA.
215. Swofford, H.J., Wortman, T. M., *Statistical Interpretation Software for Friction Ridge Skin Impressions (FRStat) Workshop*, in *Impression, Pattern, and Trace Evidence Symposium*. 2018: Washington, D.C., USA.
216. *Escobar v. State*. No. AP-76,571 (Tex. Crim. App. Nov. 20, 2013).
217. *Ex parte Escobar (No. WR-81,574-01)*. No. WR-81,574-01 (Tex. Crim. App. Feb. 24, 2016)
218. *Ex parte Escobar (No. WR-81,574-02) (Application of Writ of Habeas Corpus Cause No. D-1-DC-09-301250 in the 167th Judicial District Court, Travis County)*. 2017, No. WR-81,574-02 (Tex. Crim. App., Feb. 10, 2017).
219. *Tex. Code Crim. Proc. art 11.073*.
220. *Final Audit Report For Austin Police Department Forensic Services Division DNA Section*. 2016, Texas Forensic Science Commission.
221. *Ex parte Escobar (No. WR-81,574-02) (Order on Application of Writ of Habeas Corpus Cause No. D-1-DC-09-301250 in the 167th Judicial District Court, Travis County)*. 2017, No. WR-81,574-02 (Texas Court of Criminal Appeals, Oct. 18, 2017).
222. *Ex parte Escobar (No. WR-81,574-02) (Forensic Latent Print Examination Report, Dec. 5, 2018)*. No. WR-81,574-02 (167th Judicial District Court, Travis County, Texas).
223. *Ex parte Escobar (No. WR-81,574-02) (Motion to Exclude the Expert Testimony of Henry Swofford Related to the FRStat Evidence Pursuant to Daubert v. Merrill Dow Pharmaceuticals, Mar. 12, 2019)*. 2019, No. WR-81,574-02 (167th Judicial District Court, Travis County, Texas).
224. *Ex parte Escobar (No. WR-81,574-02) (Applicant's Motion for Disclosure of Materials Related to FRStat and Additional Latent Print Analysis, Dec. 18, 2018)*. 2018, No. WR-81,574-02 (167th Judicial District Court, Travis County, Texas).
225. *Ex parte Escobar (No. WR-81,574-02) (Renewed Motion for Discovery of Materials Regarding FRStat and Motion to Continue Hearing Regarding Latent Print Evidence, Feb. 25, 2019)*. 2019, No. WR-81,574-02 (167th Judicial District Court, Travis County, Texas).
226. *Ex parte Escobar (No. WR-81,574-02) (Renewed Motion to Continue Hearing Regarding Admissibility of FRStat Evidence, Mar. 5, 2019)*. 2019, No. WR-81,574-02 (167th Judicial District Court, Travis County, Texas).
227. *Ex parte Escobar (No. WR-81,574-02) (Response in Opposition to Applicant's Renewed Motion to Continue Hearing Regarding Admissibility of FRStat Evidence, Mar. 6, 2019)*. 2019, No. WR-81,574-02 (167th Judicial District Court, Travis County, Texas).

228. Neumann, C., *Defence Against the Modern Arts: The Curse of Statistics: Part I—FRStat.* Law, Probability and Risk, 2020. **19**(1): p. 1-20.
229. Swofford, H., F. Zemp, A. Liu, and M. Salyards, *Letter to the Editors regarding Neumann, C.'Defence against the modern arts: the curse of statistics: Part 1—FRStat.* Law, Probability and Risk, 2020. **19**(1): p. 1–20.
230. *Ex parte Escobar (No. WR-81,574-02) (Affidavit of Henry Swofford, Sep. 16, 2019)*. 2019, No. WR-81,574-02 (167th Judicial District Court, Travis County, Texas).
231. *Ex parte Escobar (No. WR-81,574-02) (Affidavit of Dr. Simone Gittelson, Aug. 28, 2019)*. 2019, No. WR-81,574-02 (167th Judicial District Court, Travis County, Texas).
232. *Ex parte Escobar (No. WR-81,574-02) (Affidavit of Dr. Karen Kafadar, Aug. 31, 2019)*. 2019, No. WR-81,574-02 (167th Judicial District Court, Travis County, Texas).
233. *Ex parte Escobar (No. WR-81,574-02) (State's Argument Regarding Latent Print Evidence, Nov. 20, 2019)*. 2019, No. WR-81,574-02 (167th Judicial District Court, Travis County, Texas).
234. *Ex parte Escobar (No. WR-81,574-02) (Findings of Fact and Conclusions of Law and Order to Transmit Habeas Corpus Record (Article 11.071 and 11.073 Post Conviction Application), Dec. 31, 2020)*. 2020, No. WR-81,574-02 (167th Judicial District Court, Travis County, Texas).
235. *Ex parte Escobar (No. WR-81,574-02) (Order on Application of Writ of Habeas Corpus Cause No. D-1-DC-09-301250 in the 167th Judicial District Court, Travis County)*. 2022, No. WR-81,574-02 (Texas Court of Criminal Appeals, Jan. 26, 2022).
236. Champod, C. and J. Vuille, *Scientific Evidence in Europe — Admissibility, Evaluation and Equality of Arms.* International Commentary on Evidence, 2011. **9**(1).
237. ISO/IEC, *17025:2017 General Requirements for the Competence of Testing and Calibration Laboratories*. International Organization for Standardization (ISO) / International Electrotechnical Commission (IEC).
238. U.S. Department of Justice. Office of the Inspector General, *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.* 2011.
239. United Kingdom, *Forensic Science Regulator Act of 2021*. *https://www.legislation.gov.uk/ukpga/2021/14/contents/enacted*. 2021.
240. State of Texas, *Forensic Science Commission. https://www.txcourts.gov/fsc/*.
241. State of New York, *Commission on Forensic Science.*
242. State of Illinois, *Illinois Forensic Science Commission (SB0920).* 2021.
243. National Institute of Standards and Technology (NIST), *Organization of Scientific Area Committees for Forensic Science (OSAC). https://www.nist.gov/osac.*
244. Team, R.S., *Integrated Development Environment for R (Version 1.2.5033, R Studio Inc., Boston, M.A., http://www.rstudio.com/).* 2015.

12 Appendix A: Glossary of Acronyms and Terms

- AAAS: The acronym referring to the American Association for the Advancement of Science. An American international non-profit organization and the world's largest general scientific society.

- ACE-V: The acronym referring to the friction ridge examination methodology and stands for Analysis, Comparison, Evaluation and Verification.

- AFIS: The acronym referring to automated computer algorithms designed to compare and rank impressions based on similarity and stands for Automated Fingerprint Identification System.

- Categorical Reporting: A form of reporting forensic results which do not formally recognize or articulate the uncertainties inherent in forensic interpretation. Categorical reporting is often conveyed as statements either of certainty or of some state of quasi-certainty that can be treated as tantamount to certainty that a particular proposition is true.

- Clarity: The distinctiveness and ability to resolve aspects of features within an impression.

- Close non-match: A general concept that refers to comparisons of two friction ridge impressions that bear high similarity but are known to have been made by different sources (i.e., non-mated sources).

- Complexity: A general concept that refers to the characteristic(s) of an impression or comparison in which the attributes of one or both impressions may require additional consideration and quality control measures. Characteristics of impressions or comparisons designated as complex often have lower quality and quantity of features impacting their ability to be reliably interpreted thereby rendering them more vulnerable to erroneous outcomes or inconsistency between experts (e.g., reliance on distorted or low-quality features). Note: characteristics that render an impression complex can be similar to those that render an impression difficult.

- Complexity_{GQS}: The GQS result from the DFIQI software that measures the complexity of an impression along a continuum. A result of -1.0 indicates the impression is designated as "highly complex," and a result of 1.0 indicates the impression is designated as "non-complex."

- Conclusion: The result of a forensic examination that is the outcome of a system controlled by a laboratory's quality assurance program, which can encompass the analyst's opinion, among other inputs and criteria, such as data and calculations.

- DFIQI: The acronym referring to Defense Fingerprint Image Quality Index software. The DFIQI is a quality assessment software for friction ridge skin impression evidence. It measures the clarity of friction ridge impression minutiae and provides a quantitative

assessment of the overall quality of an impression for comparison and evaluation purposes as it relates to determinations of Value, Complexity, and Difficulty of the impression during the Analysis phase of the examination methodology.

- Difficulty: A general concept that refers to the characteristic(s) of an impression or comparison in which the attributes of one or both impressions may require additional consideration and quality control measures. Characteristics of impressions or comparisons designated as difficult often have subtle distinguishing features that are challenging to resolve thereby rendering them more vulnerable to erroneous outcomes or inconsistency between experts (e.g., reliance on edge-shapes or pore structures, or close non-match comparisons). Note: characteristics that render an impression difficult can be similar to those that render an impression complex.

- Difficulty_{GQS}: The GQS result from the DFIQI software that measures the difficulty of an impression along a continuum. A result of -1.0 indicates the impression is designated as "high difficulty," and a result of 1.0 indicates the impression is designated as "low difficulty."

- Discriminating features: The extent to which features are able to differentiate between mated and non-mated sources.

- EBM or EBP: The acronym referring to a movement in the medical community which emphasized the integration of research evidence into the scheme of clinical decision making and stands for Evidence-Based Medicine or Evidence-Based Practice.

- Fingermark: A general term used to describe impressions of friction ridge skin on the underside of the fingers, palms, toes, and soles of the feet left under accidental or chance conditions. For purposes of this thesis, it is used interchangeably with "latent print" and "mark."

- Fingerprint: A general term used to describe impressions of friction ridge skin on the underside of the fingers, palms, toes, and soles of the feet. Impressions can be left under accidental or chance conditions (i.e., "fingermark," "latent print," "mark") or can be left under controlled conditions and for which the identity of the source is known (i.e., "print").

- FRS: The acronym referring to the Friction Ridge Subcommittee of the Organization of Scientific Area Committees for Forensic Science (OSAC).

- FRStat: The acronym referring to Friction Ridge Statistical Interpretation software. The FRStat is statistical interpretation software for friction ridge skin impression evidence. The software utilizes established statistical methods to calculate a statistic (GSS) summarizing the similarity between feature configurations on two separate images of friction ridge skin impressions. Using this statistic, the software calculates the conditional probabilities of a given GSS value or more extreme among datasets of values from mated and non-mated impressions of friction ridge skin.

- GQS: The acronym referring to the Global Quality Score, a generic reference to a score produced by the DFIQI software that measures the quality of a friction ridge impression on the basis of the quality and quantity of ridge detail against three different scales: Value, Complexity, and Difficulty.

- GSS: The acronym referring to the Global Similarity Statistic, a statistic produced by the FRStat software that summarizes the similarity between feature configurations on two separate images of friction ridge skin impressions.

- GYRO: The acronym referring to a method for color coding feature annotations to convey levels of uncertainty in the existence of annotated minutiae and stands for "Green," "Yellow," "Red," and "Orange." The color annotations, Green, Yellow, and Red indicate high, medium, and low confidence, respectively, in the presence of a feature. The color annotation Orange indicates the feature was observed after the analyst viewed the fingerprint standard.

- Identification: The conclusion by an analyst that two impressions were made by the same source of friction skin as a result of their examination.

- Latent Print: A general term used to describe impressions of friction ridge skin on the underside of the fingers, palms, toes, and soles of the feet left under accidental or chance conditions. For purposes of this thesis, it is used interchangeably with "fingermark" and "mark."

- Lights-out: A term referring to fully automated fingerprint identification systems with no human intervention.

- LQS: The acronym referring to the Local Quality Score, a score produced by the DFIQI software that measures the quality of a friction ridge feature within an impression on the basis of its clarity of surrounding ridge detail.

- Mark: A general term used to describe impressions of friction ridge skin on the underside of the fingers, palms, toes, and soles of the feet left under accidental or chance conditions. For purposes of this thesis, it is used interchangeably with "fingermark" and "latent print."

- Mated sources: The condition in which it is known that two impressions were made by the same friction ridge skin.

- NAS: The acronym referring to the National Academy of Science, a non-profit, non-government organization chartered by the United States Government to provide independent, objective advice to the nation on matters related to science and technology.

- NIJ: The acronym referring to the National Institute of Justice, an agency of the United States Government.

- NIST: The acronym referring to the National Institute of Standards and Technology, an agency of the United States Government.

- Non-mated sources: The condition in which it is known that two impressions were not made by the same friction ridge skin (i.e. impressions were made by different friction ridge skin).

- NRC: The acronym referring to the National Research Council, the operating arm of the National Academies of Science (NAS).

- Opinion: The result of a forensic examination that is the outcome of an analyst's subjective assessment of the evidence based on their knowledge, skills, training, and experience.

- OSAC: The acronym for the Organization of Scientific Area Committees for Forensic Science. The OSAC consists of several committees and subcommittees that facilitate and promote the development and use of high-quality, technically sound standards in forensic science. These standards define minimum requirements, best practices, standard protocols, and other guidance to help ensure that the results of forensic analysis are reliable and reproducible.

- PCAST: The acronym referring to the President's Council of Advisors on Science and Technology. A council chartered by the President of the United States to advise the President on matters related to science and technology.

- Print: A term used to describe an impression of friction ridge skin left under controlled conditions and for which the identity of the source is known.

- Probabilistic Reporting: A form of reporting forensic results which formally recognize and articulate the uncertainties inherent in forensic interpretation using probabilistic logic.

- Quality: The utility of an impression based on the clarity and quantity of details available for examination.

- Significance: A generic term that refers to the quality of being worthy of attention; importance. This term is distinct from its use in classic frequentist hypothesis testing.

- Suitability: The determination that an impression is of sufficient quality to perform a comparison. For purposes of this thesis, it is used interchangeably with "Value."

- SWGFAST: The acronym referring to the Scientific Working Group on Friction Ridge Analysis, Study and Technology. The SWGFAST was replaced by the OSAC Friction Ridge Subcommittee in 2014.

- Utility: The quality or state of being useful. This term should not be confused with the "utility" as used in decision theory (i.e., the expected value of an action or agent).

- Value: The determination that an impression is of sufficient quality to perform a comparison. For purposes of this thesis, it is used interchangeably with "Suitability."

- $Value_{GQS}$: The GQS result from the DFIQI software that measures the value of an impression along a continuum. A result of -1.0 indicates the impression is "no value" and a result of 1.0 indicates the impression is "of value."

13 Appendix B: Supplemental Material for Chapter 2

13.1 Appendix B-1

This appendix provides details related to the raw model diagnostics and uncertainty values for the multinomial regression model provided by the *nnet* package in R [58] against a training/test-dataset of feature measurements from impressions for which latent print examiners previously analyzed and categorized based on their "value," "complexity," and "difficulty" for comparison. The multinomial model was selected after testing a range of machine learning techniques with the variables LQS_{sum} and *n*FEAT (naïve based classifier, tree-based classifiers, discriminant analysis techniques, neural networks and support vector machines). Overall, the multinomial regression offers a competitive accuracy while maintaining easy explainability.

Machine learning and subsequent statistical analysis were carried out in R version 3.6.3 [58] coupled with RStudio Version 1.2.5033 [244] using the following packages:

- caret (Caret: Classification and Regression Training. https://CRAN.R-project.org/package=caret.) and the libraries associated with each tested model.
- pROC (pROC: Display and Analyze ROC Curves. https://CRAN.R-project.org/package=pROC.).

Figure B-1-1 illustrates the performance of the multinomial regression model compared to other classifiers. Raw classification performance related to the multinomial regression model are provided in Tables B-1-1a through B-1-1c (confusion matrices) and Table 2-2 (overall classification accuracy). Tables B-1-3a through B-1-3c provide the estimated model coefficients (represented in Tables 2-2a through 2-2c in the body of the paper) along with their standard error.

Note: The performance metrics presented in this appendix correspond to raw classification performance of the model against all three outcomes for each judgment class (value, complexity, and difficulty). These values do not correspond to the performance of the GQS scores when applied to a binary decision to flag an impression for further quality assurance review or not. The performance of the GQS scores is presented in the body of the paper.

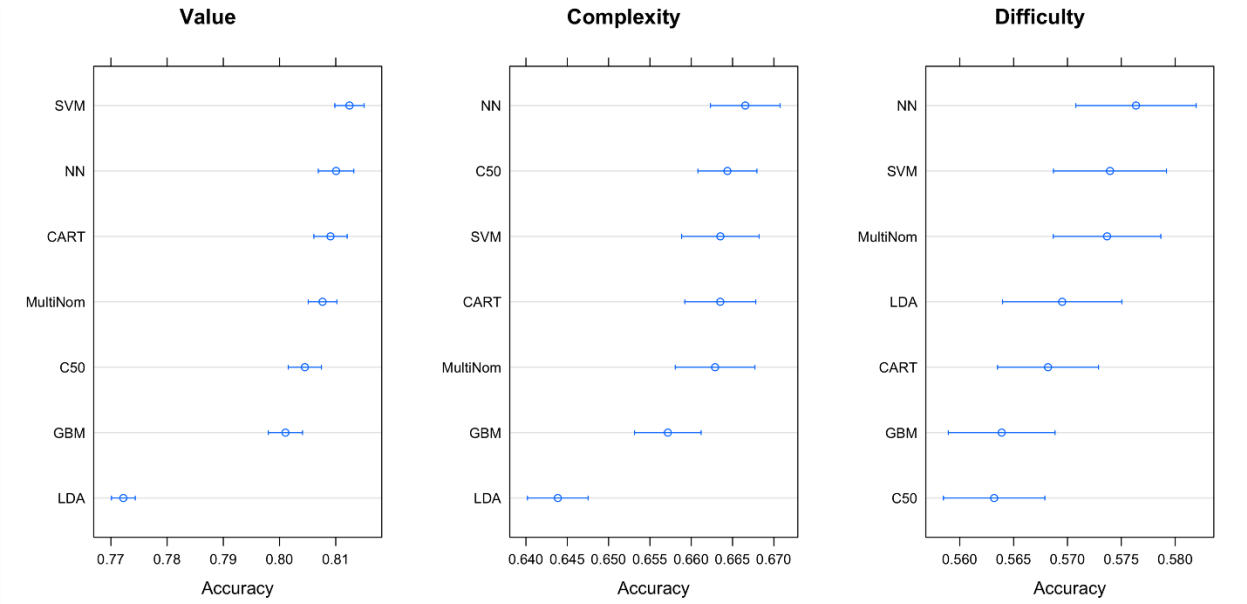


Figure B-1-1: Comparison of the raw classification accuracy of various machine learning techniques tested against a training/test-dataset described by GQS Test Dataset-1.

Judgment / Prediction	NV	VEO	VID	Total
NV	169	84	13	266
VEO	17	22	15	54
VID	66	121	1,113	1,300
Total	252	227	1,141	1,620

Table B-1-1a: Confusion matrix resulting from raw model classifications for each outcome in the Value judgment class.

Judgment / Prediction	Highly-Complex	Complex	Non-Complex	Total
Highly-Complex	193	72	13	278
Complex	68	160	126	354
Non-Complex	30	220	738	988
Total	291	452	877	1,620

Table B-1-1b: Confusion matrix resulting from raw model classifications for each outcome in the Complexity judgment class.

Judgment / Prediction	High	Medium	Low	Total
High	324	186	62	572
Medium	131	266	166	563
Low	32	104	349	485
Total	487	556	577	1,620

Table B-1-1c: Confusion matrix resulting from raw model classifications for each outcome in the Difficulty judgement class.

Class	Overall Accuracy	95% Confidence Interval (lower bound – upper bound)
Value	0.805	(0.785 - 0.824)
Complexity	0.674	(0.650 - 0.696
Difficulty	0.580	(0.555 - 0.604)

Table B-1-2: Overall accuracy and associated uncertainty (95% confidence interval) from raw model classifications for each judgment class (Value, Complexity, Difficulty).

"Value" Coefficients	Intercept	LQS_{sum}	*n*FEAT
NV	0.000 (SE: N/A)	0.000 (SE: N/A)	0.000 (SE: N/A)
VEO	-1.736 (SE: 0.268)	-0.051 (SE: 0.214)	0.277 (SE: 0.073)
VID	-6.042 (SE: 0.389)	0.495 (SE: 0.200)	0.726 (SE: 0.074)

Table B-1-3a: Multinomial coefficients for each outcome class probability of the Value judgment represented in Table 2-2a along with their standard errors (SE). NOTE: The nnet package estimates the coefficients relative to one outcome class which has an assigned coefficient of 0.000. SE is represented as N/A for those values.

"Complexity" Coefficients	Intercept	LQS_{sum}	*n*FEAT
Highly Complex	3.325 (SE: 0.267)	-0.100 (SE: 0.185)	-0.459 (SE: 0.061)
Complex	0.000 (SE: N/A)	0.000 (SE: N/A)	0.000 (SE: N/A)
Non-Complex	-1.781 (SE: 0.198)	0.741 (SE: 0.089)	-0.025 (SE: 0.028)

Table B-1-3b: Multinomial coefficients for each outcome class probability of the Complexity judgment represented in Table 2-2b along with their standard errors (SE). The nnet package in R estimates the coefficients relative to one outcome class which has an assigned coefficient of 0.000. SE is represented as N/A for those values.

"Difficulty" Coefficients	Intercept	LQS_{sum}	*n*FEAT
High	0.000 (SE: N/A)	0.000 (SE: N/A)	0.000 (SE: N/A)
Medium	-1.896 (SE: 0.171)	0.289 (SE: 0.105)	0.125 (SE: 0.033)
Low	-3.071 (SE: 0.198)	0.965 (SE: 0.108)	-0.004 (SE: 0.035)

Table B-1-3c: Multinomial coefficients for each outcome class probability of the Difficulty judgment represented in Table 2-2c along with their standard errors (SE). The nnet package in R estimates the coefficients relative to one outcome class which has an assigned coefficient of 0.000. SE is represented as N/A for those values.

14 Appendix C: Supplemental Material for Chapter 3

14.1 Appendix C-1

As described earlier, the resulting value of the global similarity statistic is dependent upon the manual selection and annotation of features by fingerprint experts. This appendix (1) evaluates the precision of feature annotations with respect to their location and angle for a given feature by fingerprint experts as well as (2) describes how such variation is accounted for in the resulting global similarity statistic using simulated variations of feature annotations.

Empirical variability of feature annotations

The variability of feature annotations (for a given feature) was evaluated with respect to the differences in the location and angle of each annotated feature compared to a specified reference point. This was evaluated separately for latent impressions and reference impressions due to the general clarity differences between the two types of impressions.

Latent Impressions:

The variability of feature annotations in latent impressions was evaluated using five practicing latent fingerprint experts employed by a federal crime laboratory in the United States. Each expert was provided five sets of fourteen images of latent fingerprints. Each set contained the same fourteen images. Considering the intent of this evaluation is to capture the reproducibility of annotations for a given feature, a template image was also provided indicating which features the experts should annotate. The template did not, however, indicate exactly where or how to annotate the feature. Although the specific number of features varied across images, the total number of features annotated by each expert per set was 100 (n = 2,500 annotations in total). The overall quality of the images used was subjectively considered representative of the quality of typical latent impressions received during normal casework. Of the available features, those that were subjectively evaluated as “low” or “medium” clarity (on a scale of “low,” “medium,” and “high” clarity) were specified on the template image. Experts were advised to annotate each set during normal business hours using the same software and hardware as they normally would in actual casework. Furthermore, experts were given one week to complete the five sets and were advised to ensure at least four hours lapsed between each set.

The X, Y coordinates and angle for each feature in each image was extracted. The mean X, Y coordinate and angle for each feature was calculated across all experts and sets and served as the “consensus” reference location and angle. The difference between the “consensus” feature location and angle compared to each annotation was calculated. Figure C-1-1 illustrates the empirical distribution of variations in X, Y, and angle annotations, respectively.

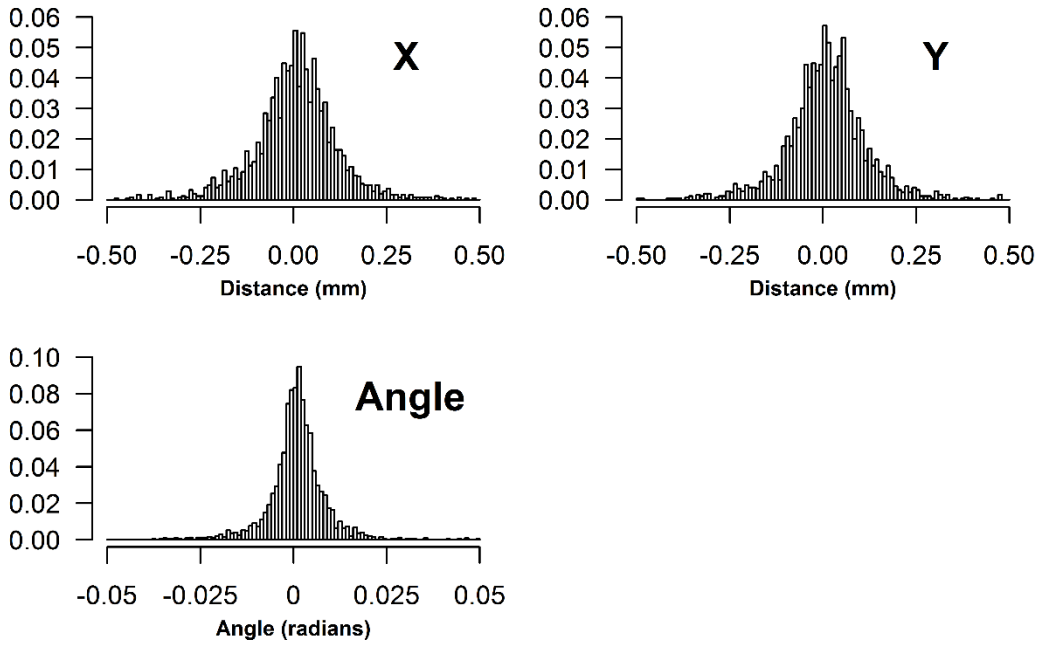


Figure C-1-1. Empirical density distributions of the X, Y, and angle differences as a result of variations in feature annotations on latent impressions (n = 2,500).

Reference Impressions:

The variability of feature annotations in reference impressions was evaluated using ten practicing latent fingerprint experts employed by a federal crime laboratory in the United States. Each expert was provided ten replicate images of a single reference fingerprint. Considering the intent of this evaluation is to capture the reproducibility of annotations for a given feature, a template image was also provided indicating which features the experts should annotate. The template did not, however, indicate exactly where or how to annotate the feature. The template specified ten features to annotate resulting in a total of 100 annotations by each expert (n = 1,000 annotations in total). All specified features were subjectively evaluated as "high clarity" (on a scale of "low," "medium," and "high" clarity) and representative of the quality of typical reference impressions received during normal casework. Experts were advised to annotate each image during normal business hours using the same software and hardware as they normally would in actual casework. Furthermore, experts were given approximately one week to complete the annotations and were advised to ensure at least four hours lapsed between annotations for each image.

The X, Y coordinates and angle for each feature in each image was extracted. The mean X, Y coordinate and angle for each feature was calculated across all experts and sets and served as the "consensus" reference location and angle. The difference between the "consensus" feature location and angle compared to each annotation was calculated. Figure C-1-2 illustrates the empirical distribution of variations in X, Y, and angle annotations, respectively.

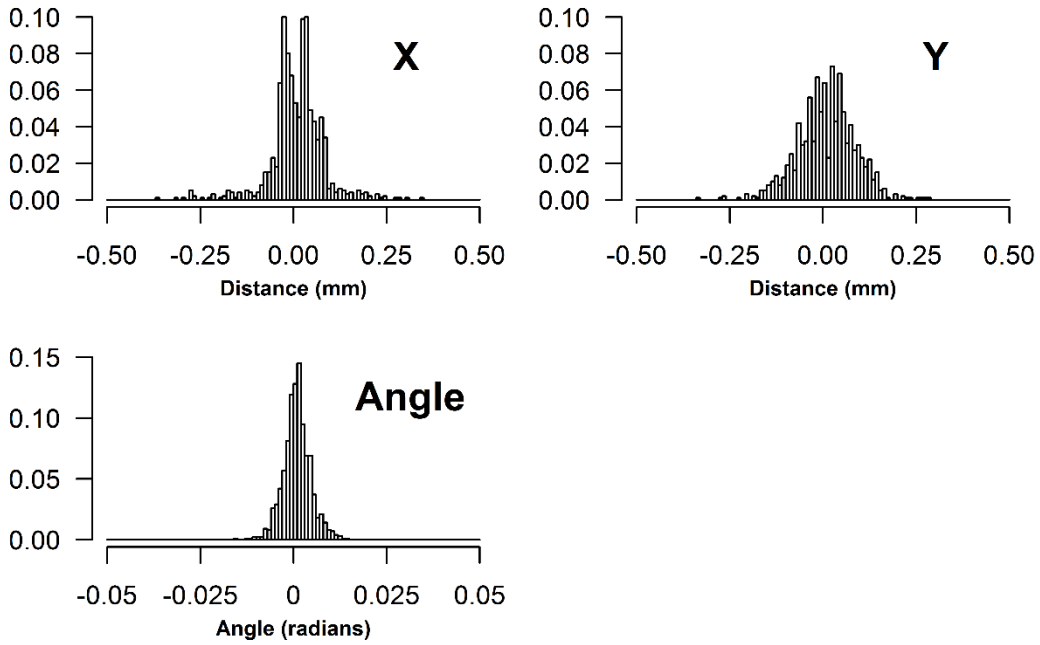


Figure C-1-2. Empirical density distributions of the X, Y, and angle differences as a result of variations in feature annotations on reference impressions (n = 1,000).

Simulated variability of feature annotations

The global similarity statistic value is initially calculated based on the locations and angles of features annotated by the user. To simulate the impact of normal variations by analysts re-annotating the same features, the similarity statistic is recalculated *k* iterations (*k* = 100) using randomly displaced feature locations and angles on both image #1 and image #2 normally distributed around the initial annotation made by the user. The displacements are calculated using a scaled approximation of the inverse of the normal cumulative distribution. The parameter values were estimated using the empirical distributions and manually optimized to minimize the differences between the empirical distribution and a randomly generated sample distribution. For latent impressions, the parameter value for distance variations was estimated with respect to both X and Y-value differences since the two empirical distributions appear very similar to one another. For the reference impressions, however, the parameter value for distance variations was estimated with respect to Y-value differences only since the X-values indicate *less* variation than Y-values. This observation seems to suggest there is some other factor influencing the feature annotations with respect to the X-axis compared to the Y-axis on the reference impressions. As a result of this observation, the template image used to specify which features to annotate was examined to determine whether the feature selection could have biased the variations in one direction vs. another (i.e., X-value vs. Y-value displacements). Indeed, it was observed that 80% of the specific features selected for the expert all occurred in the north-south direction. Accordingly, the differences in variation appear to be related to the uncertainty associated with annotating the specific end-point location of the ridge (spanning north to south) rather than the precise center of the ridge, which is more clearly interpretable. While the explanation seems plausible, the more

important consideration is that the parameter value for distance variations be estimated using the Y-value differences, which exhibited greater variation.

The empirical distributions of X, Y, and angle differences were compared to a randomly generated sample distribution. Figure C-1-3 illustrates the comparison between the randomly generated sample distributions and the empirical distributions for X, Y, and angle displacements for latent impressions. Table C-1-1 provides the two sample Kolmogorov-Smirnov (K-S) test statistics as well as the resulting *p*-values under the null hypothesis that the empirical distributions and randomly generated distributions were drawn from the same distribution.

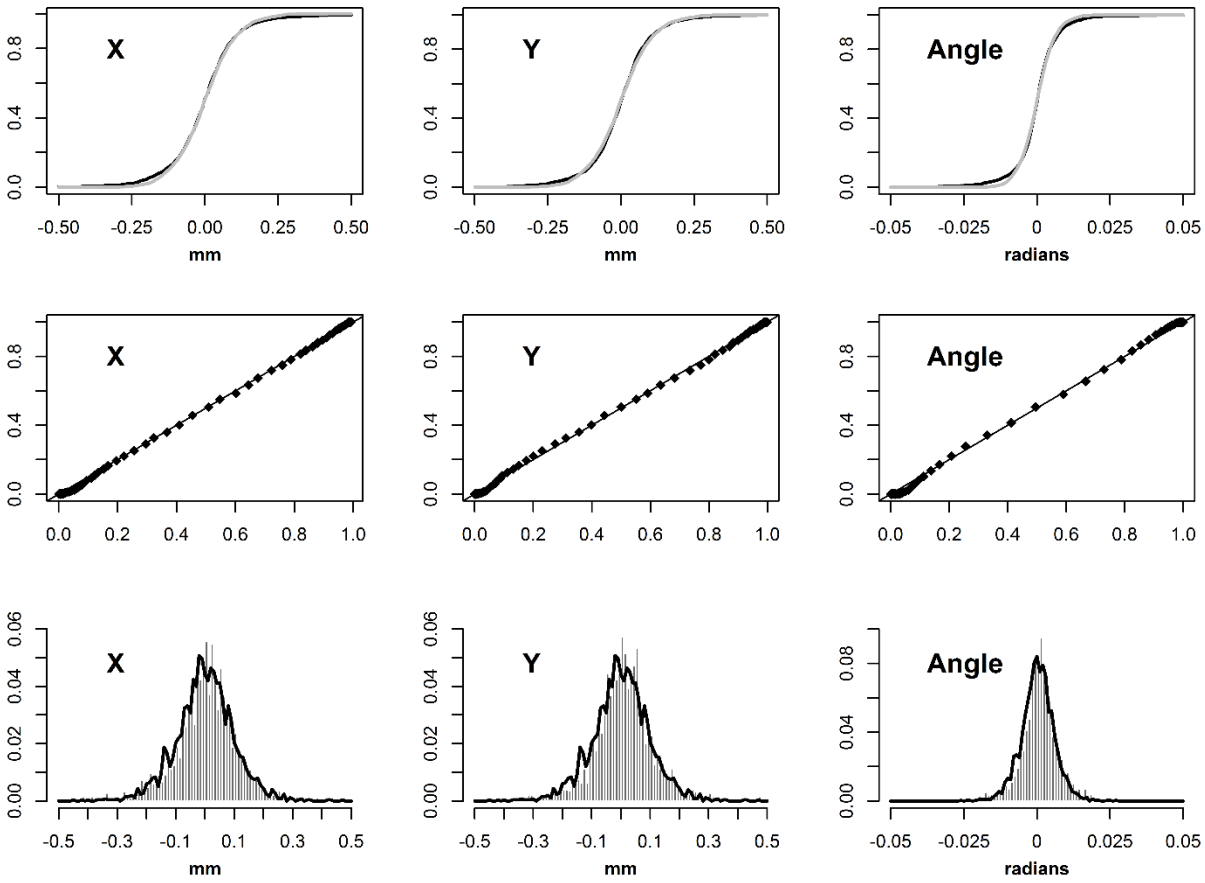


Figure C-1-3. Empirical distributions of the X, Y, and angle differences as a result of variations in feature annotations compared to a randomly generated distribution of X, Y, and angle differences using a scaled approximation of the inverse of the normal cumulative distribution for latent impressions. Top row illustrates overlays of the empirical cumulative distribution (black) and randomly generated dataset (grey). Middle row illustrates the P-P plots between the empirical dataset (X-axis) and randomly generated dataset (Y-axis) (the black dots represent the P-P plot and the grey line represents the ideal slope of 1). Bottom row illustrates overlays of the empirical density distribution (grey histogram) and randomly generated dataset (black line).

Feature quantity	***n* sample 1** (empirical)	***n* sample 2** (randomly generated)	**K-S test statistic**	***p (null)***
X-value difference	2,500	2,500	0.025	$p >> 0.05$
Y-value difference	2,500	2,500	0.023	$p >> 0.05$
Angle difference	2,500	2,500	0.028	$p >> 0.05$

Table C-1-1. Summary of the Kolmogorov-Smirnov test results between empirical distribution and randomly generated distribution of displacements for the X, Y, and angle for latent prints. Statistical significance is based on a p-value decision threshold of 0.01.

Figure C-1-4 illustrates the comparison between the randomly generated sample distributions and the empirical distributions for X, Y, and angle displacements for reference impressions. Table C-1-2 provides the two sample K-S test statistics as well as the resulting *p*-values under the null hypothesis that the empirical distributions and randomly generated distributions were drawn from the same distribution.

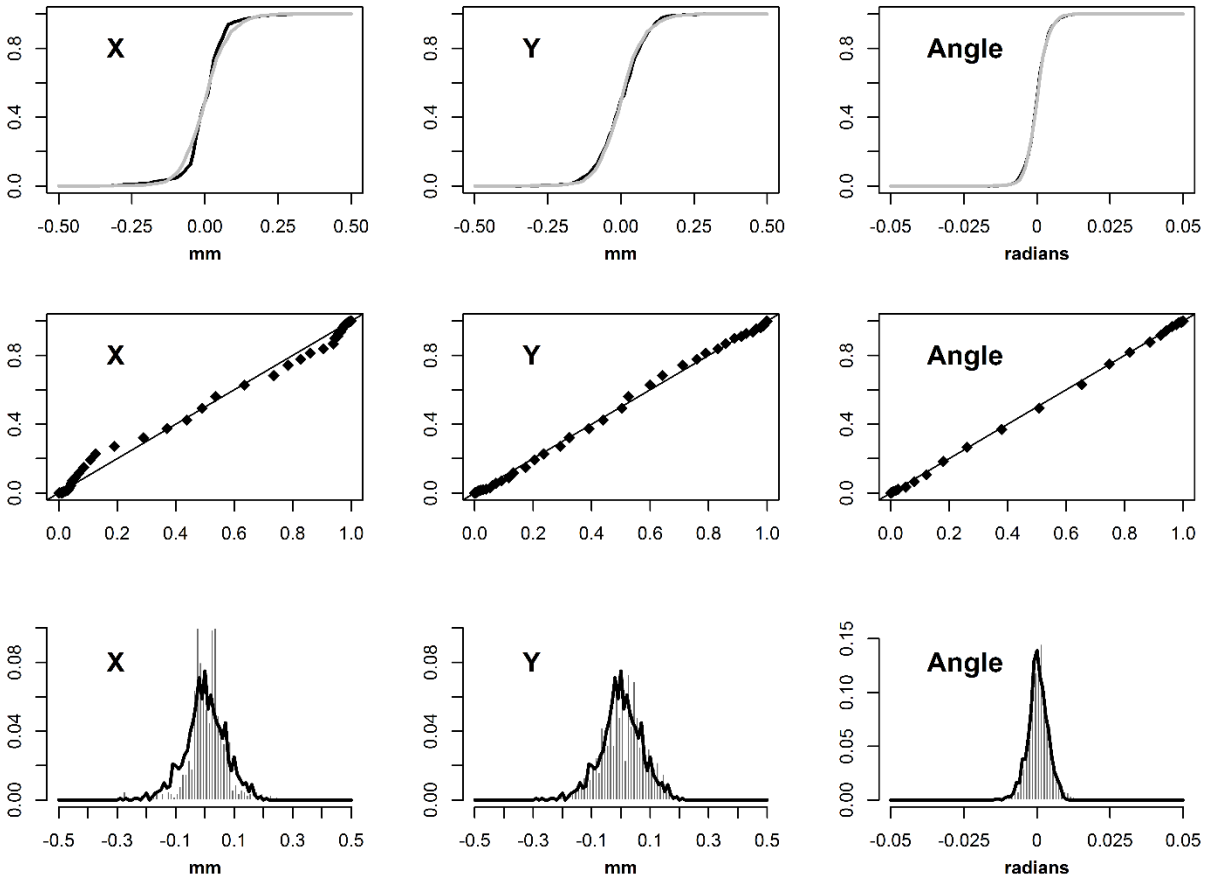


Figure C-1-4. Empirical distributions of the X, Y, and angle differences as a result of variations in feature annotations compared to a randomly generated distribution of X, Y, and angle differences using a scaled approximation of the inverse of the normal cumulative distribution for reference impressions. Top row illustrates overlays of the empirical cumulative distribution (black) and randomly generated dataset (grey). Middle row illustrates the P-P plots between the empirical dataset (X-axis) and randomly generated dataset (Y-axis) (the black dots represent the P-P plot and the grey line represents the ideal slope of 1). Bottom row illustrates overlays of the empirical density distribution (grey histogram) and randomly generated dataset (black line).

Feature quantity	***n* sample 1** (empirical)	***n* sample 2** (randomly generated)	**K-S test statistic**	***p (null)***
X-value difference	1,000	1,000	0.103	$p < 0.001$
Y-value difference	1,000	1,000	0.042	$p > 0.05$
Angle difference	1,000	1,000	0.019	$p > 0.05$

Table C-1-2. Summary of the Kolmogorov-Smirnov test results between empirical distribution and randomly generated distribution of displacements for the X, Y, and angle for reference prints. Statistical significance is based on a p-value decision threshold of 0.01.

Based on these findings, with the exception of X-value differences in the reference impressions (for reasons previously discussed), the distributions exhibit little difference and thus the scaled approximation of the inverse of the normal cumulative distribution is proposed as a sufficient means of simulating the impact of variations in feature annotations by fingerprint experts. Figure C-1-5 illustrates the simulated variations of feature annotations.

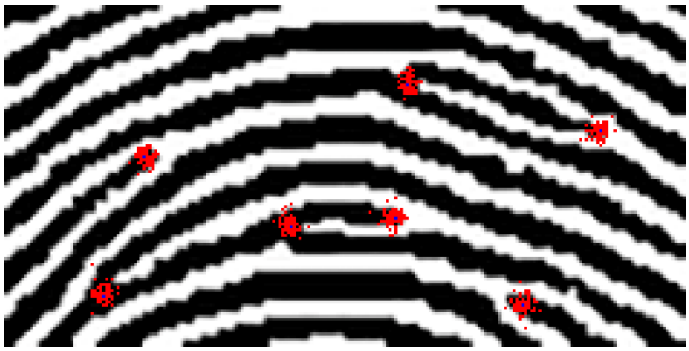


Figure C-1-5. Illustration of the iterative random sampling scheme for the annotated details resulting in random displacements of feature annotations. The blue dot represents the X, Y pixel location of the center of the original annotation by the expert. The red dots each represent separate randomly generated displacements. NOTE: only displacements in terms of Euclidean distance are illustrated in this figure.

14.2 Appendix C-2

This appendix provides greater detail regarding the determination of the region of friction ridge skin which maximizes the opportunities of observing higher similarity statistic values (i.e. which region of friction ridge skin results in more similar configurations of features). Two sets of fingerprints were selected to empirically determine the optimal region for conditioning the non-mated distribution: (1) a sample representing the "delta" region of fingerprints and (2) a sample representing the "core" region of fingerprints. Only the delta and core regions were considered because they provide a known anatomical reference point on the fingerprint and have the highest densities of features with respect to other areas of the friction ridge skin. Each dataset was separated into eleven separate subsets, each containing approximately 100 samples, conditioned on the number of features (*n*) being compared (ranging from 5 features to 15 features). All fingerprint images consisted of reference impressions taken under controlled conditions such that distortions were minimized. Features were manually annotated by practicing fingerprint experts. Features were annotated such that the features closest to the reference point (core or delta depending on the sample) were annotated first and then the remaining *n* features were annotated in a radiating fashion outward. Post annotation, each image was cropped by a bounding rectangle such that only those ridges and features that are part of the annotated configuration remain. These images serve as the "query" print (image #1).

Each query print was then searched using an Automated Fingerprint Identification System (AFIS) against an operational database containing approximately 100 million different fingerprint impressions from approximately 10 million different individuals. The AFIS ranked the top 20

most similar reference fingerprints to the fingerprint image searched. Of the top 20 results, the fingerprint image in rank 1 was confirmed to be a non-mated source with respect to the query print and saved (image #2). For each non-mated rank 1 result, fingerprint features were annotated manually by practicing fingerprint experts and independent of the features annotated on image #1. Features were annotated such that the features closest to the reference point (core or delta depending on the sample) were annotated first then the remaining *m* features were annotated in a radiating fashion (where $m \geq n + 5$). Fingerprints contain only one core, but some fingerprints, depending on the pattern type, may contain up to two deltas. For the sample consisting of two deltas, the AFIS did not indicate which delta resulted in the high similarity ranking. As a result, for the AFIS results which contained two deltas, both were annotated and the query print (image #1) was compared separately against each delta using the method described in Section I. Between the two possible deltas, the delta resulting in the highest similarity statistic value when compared to the query print (image #1) was retained.

Figure C-2-1 illustrates the empirical cumulative frequency distributions of the similarity statistics between the two samples (core vs. delta). Visually, it can be observed that the distribution of similarity statistic values from the delta region consistently resulted in higher similarity statistic values compared to the core region. A two sample Kolmogorov-Smirnov (K-S) test was performed comparing the distributions for each quantity of features. Table C-2-1 provides the K-S test statistics as well as the resulting *p*-value under the null hypothesis that the two samples originated from the same distribution. Based on these findings, the delta region was determined to maximize the opportunities of observing *higher* similarity statistic values among non-mated samples and thus is the optimal region of the fingerprint to condition the empirical distribution of similarity statistic values.

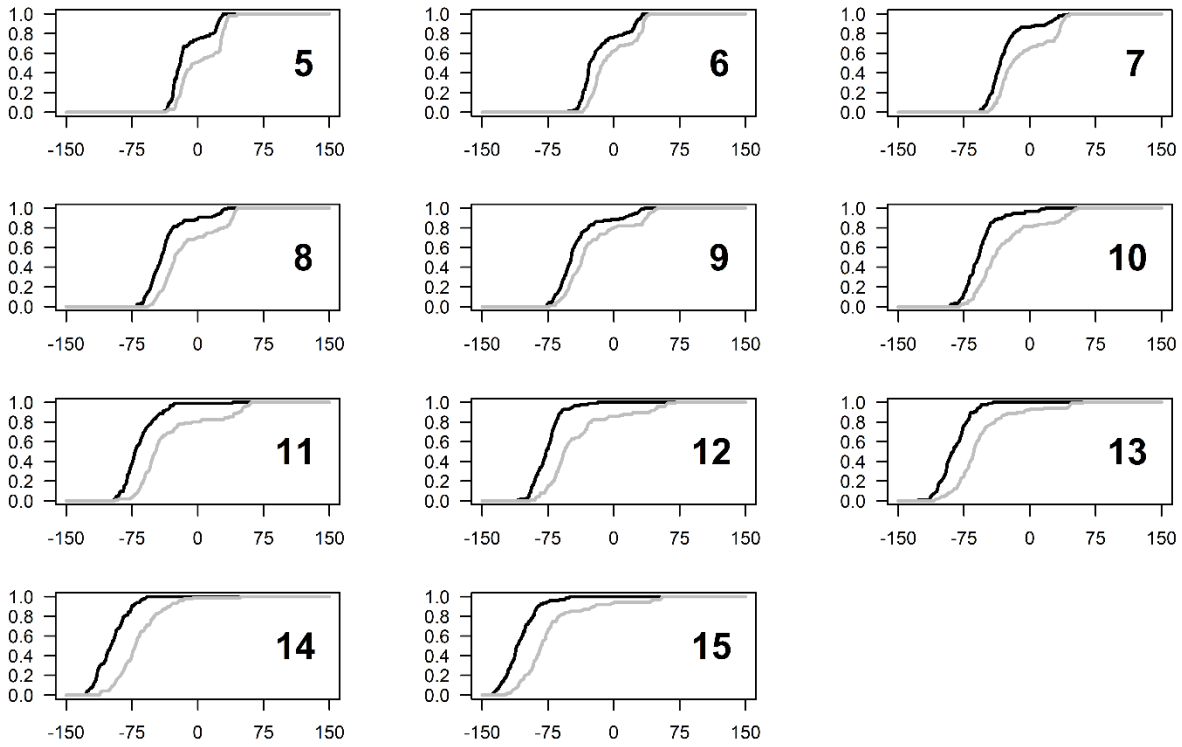


Figure C-2-1. Empirical cumulative frequency distributions of the similarity statistic values from non-mated AFIS "core" comparisons compared to the non-mated AFIS "delta" comparisons for each quantity of features (ranging from 5 to 15). The black line represents the "core" results. The grey line represents the "delta" results. The X-axis represents the global similarity statistic values.

Feature Quantity	***n* sample 1** (core)	***n* sample 2** (delta)	**K-S test statistic**	***p (null)***
5	94	99	0.345	$p << 0.01$
6	96	99	0.370	$p << 0.01$
7	95	96	0.318	$p << 0.01$
8	96	99	0.416	$p << 0.01$
9	95	99	0.310	$p << 0.01$
10	96	97	0.431	$p << 0.01$
11	95	96	0.487	$p << 0.01$
12	97	98	0.549	$p << 0.01$
13	97	99	0.520	$p << 0.01$
14	96	100	0.552	$p << 0.01$
15	95	100	0.537	$p << 0.01$

Table C-2-1. Summary of the Kolmogorov-Smirnov test results between the empirical cumulative frequency distributions of the similarity statistic values from AFIS "core" comparisons compared to the AFIS "delta" comparisons for each quantity of features (ranging from 5 to 15). Although each set initially consisted of 100 samples, some failed search results caused a few images to be discarded prior to calculating the similarity statistic values. Statistical significance is based on a p-value decision threshold of 0.01.

14.3 Appendix C-3

As described earlier, the appropriate mated sample is that which results in a distribution of similarity statistic values which is similar to the distribution of similarity statistic values observed in actual casework or biased towards lower similarity statistic values thus ensuring that the empirical distributions represent the full range of plausible similarity statistic values that could reasonably be observed in casework when impressions are subject to various distortions during deposition. This appendix provides greater detail regarding the determination.

A casework dataset of 605 latent and reference impressions were collected from casework during the course of routine operations by fingerprint experts in a federal crime laboratory in the United States and reported as "positive associations." The impressions were collected from a wide variety of cases, substrates, and assigned fingerprint experts. The corresponding features (ranging between 7 and 15) were manually annotated by the assigned fingerprint expert during the initial case examination. The selected features were then annotated in the proper format at a later time by the same fingerprint expert for purposes of this evaluation. The distribution of similarity statistic values from this casework sample were compared to the empirical distribution of similarity statistic values described earlier from mated sources in which extreme distortions were deliberately produced during deposition on a livescan device. To ensure the casework sample had sufficient similarity statistic values for each quantity of features to compare against the mated sample, the distribution of similarity statistic values was calculated by randomly selecting one combination of *n* features out of *m* available. Figure C-3-1a illustrates a comparison of the cumulative frequency distributions of similarity statistic values between both samples for each quantity of features. Figure C-3-1b illustrates the P-P plots of the two empirical cumulative frequency distributions for each quantity of features. Figure C-3-1c illustrates a comparison of the two empirical density distributions for each quantity of features.

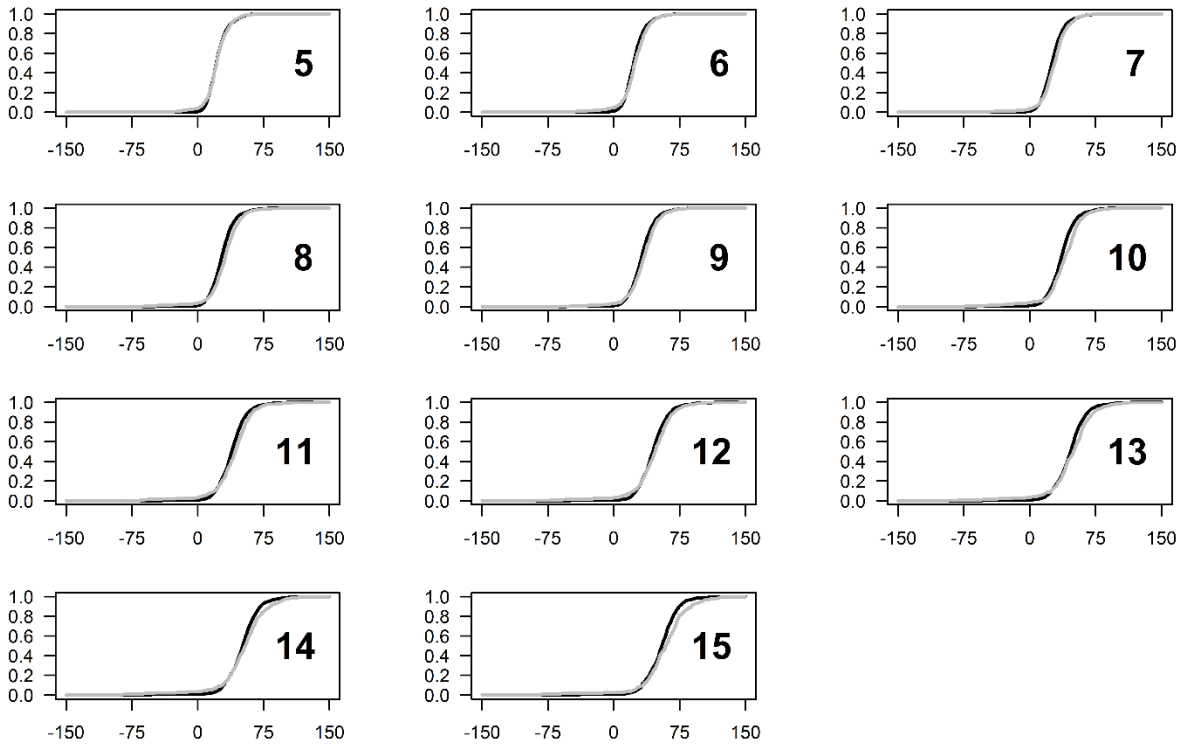


Figure C-3-1a. Empirical cumulative frequency distributions of the similarity statistic values from the casework sample (believed to be mated) and the mated sample (manually distorted – known to be mated) for each quantity of features (ranging from 5 to 15). The grey line represents results from the casework sample. The black line represents results from the mated sample. The X-axis represents the global similarity statistic values.

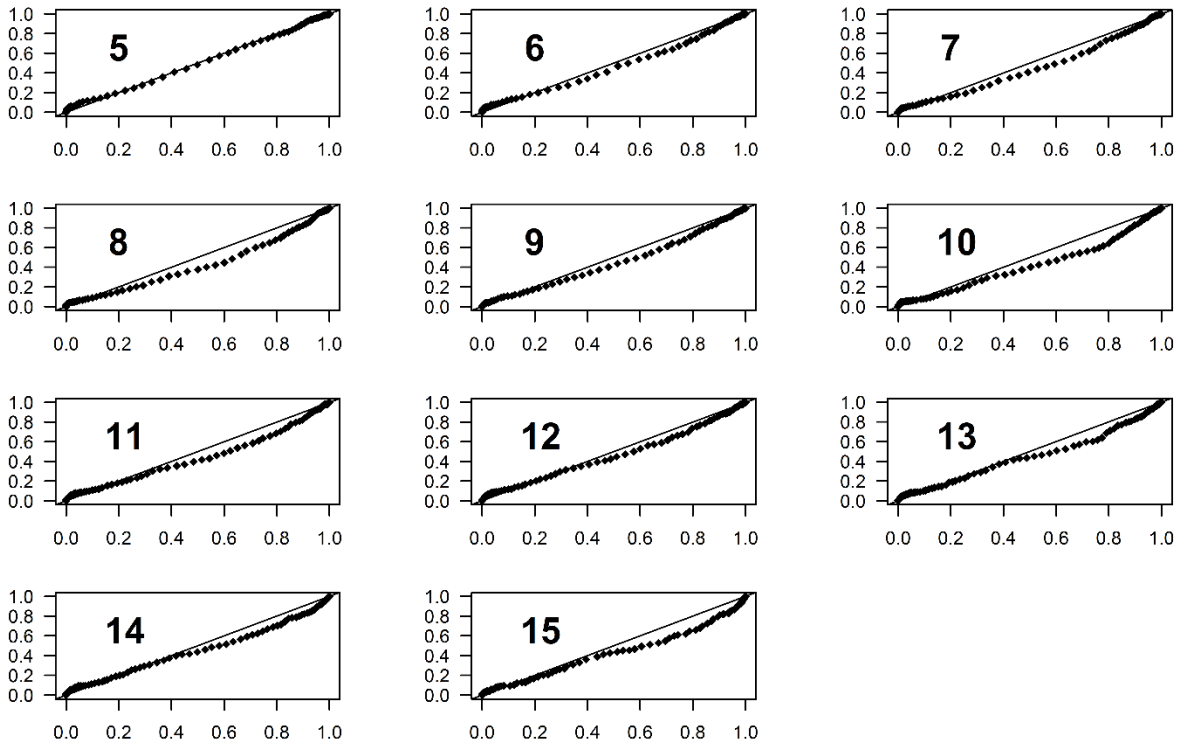


Figure C-3-1b. P-P plots of the empirical cumulative frequency distributions of the similarity statistic values from the casework sample (believed to be mated) (vertical axis) and the empirical cumulative frequency distributions of the similarity statistic values from the mated sample (manually distorted – known to be mated) (horizontal axis) for each quantity of features (ranging from 5 to 15). The black dots represent the P-P plot. The grey line represents a slope of 1 (perfect correspondence).

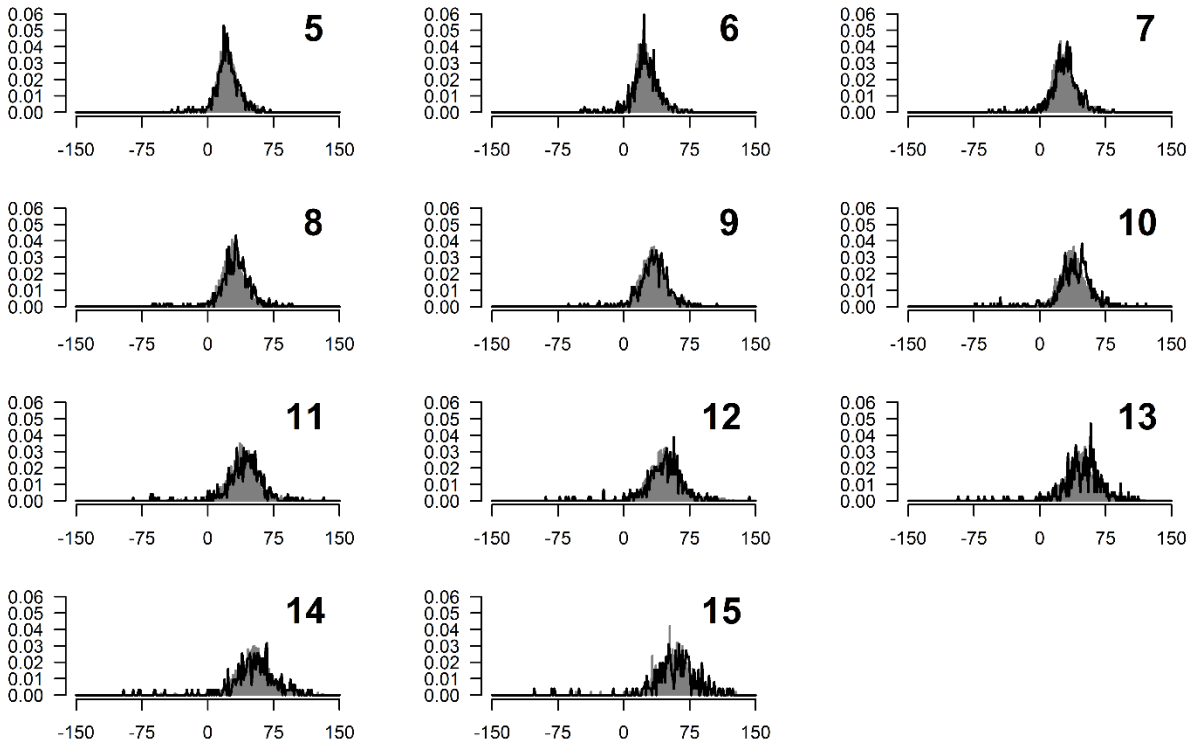


Figure C-3-1c. Empirical density distributions of the similarity statistic values from the casework sample (believed to be mated) and the mated sample (manually distorted – known to be mated) for each quantity of features (ranging from 5 to 15). The black line represents the results from the casework sample. The grey histogram represents the results from the mated sample. The X-axis represents the global similarity statistic values.

A two sample Kolmogorov-Smirnov (K-S) test was performed comparing the two distributions for each quantity of features. Table C-3-1 provides the K-S test statistics as well as the resulting *p*-value under the null hypothesis that the two samples originated from the same distribution.

Feature Quantity	***n* sample 1** (casework)	***n* sample 2** (manually distorted)	**K-S test statistic**	***p (null)***
5	605	1,996	0.046	$p > 0.05$
6	605	1,996	0.083	$p < 0.01$
7	605	1,996	0.112	$p < 0.01$
8	601	1,996	0.155	$p < 0.01$
9	585	1,996	0.101	$p < 0.01$
10	549	1,996	0.167	$p < 0.01$
11	499	1,996	0.124	$p < 0.01$
12	438	1,996	0.092	$p < 0.01$
13	382	1,996	0.139	$p < 0.01$
14	316	1,996	0.104	$p < 0.01$
15	258	499	0.153	$p < 0.01$

Table C-3-1. Summary of the Kolmogorov-Smirnov test results between the empirical cumulative frequency distributions of the similarity statistic values from the casework sample (believed to be mated) and the mated sample (manually distorted – known to be mated) for each quantity of features (ranging from 5 to 15). Statistical significance is based on a p-value decision threshold of 0.01.

From the K-S test, the null hypothesis is rejected for each quantity of features except 5. Although the null hypothesis is largely rejected by the K-S test, the distributions are not substantially different from one another in terms of appearance. More importantly, however, the means of the similarity statistic values from the mated samples (manually distorted – known to be mated) are consistently lower, to a marginal degree, than the casework samples (believed to be mated) thereby satisfying the criteria set forth above. Based on these data, the mated sample (manually distorted – known to be mated) is proposed as a plausible mated source distribution.

14.4 Appendix C-4

This appendix provides greater detail regarding the determination that the parametric models are plausible estimations of the population distributions of similarity statistic values for both non-mated and mated friction ridge skin impressions for each quantity of features.

Non-mated distribution parameter estimation

The empirical distributions for the non-mated samples exhibit a distinct bimodal appearance for several of the lower feature quantities. As the feature quantities increase, however, the proportion of the distribution represented by the second mode decreases until the distributions for higher feature quantities appear unimodal. Although initially, the bimodal appearance may seem perplexing, it is actually quite straightforward as a mathematical consequence of the weight algorithms; however, it is outside the scope of this paper to go into innate detail on the actual weight algorithms themselves. Nevertheless, recognizing that the weighting functions are based on a mixture of functions, it seems natural that the resulting distribution is also a mixture of distributions. Taking this into consideration, the empirical distributions for all quantities of

features were each modeled using *k*-component (where $k = 2$ or 3) mixtures of Gaussian distributions. Component weights and parameter estimates were determined using maximum likelihood estimation methods within commercially available statistical analysis software (JMP). Manual adjustments to the estimated weights and parameters were made to smooth trends among parameter estimates between feature quantities.

Although *k*-component Gaussian mixtures are more common, logistic distributions were applied on the basis of their heavier tails compared to Gaussian distributions. The heavier tails provide more conservative estimates of probabilities in the extreme ends of the distributions. The parameters for the logistic distribution were approximated using the estimated parameters of the Gaussian distributions. This was accomplished by setting the location parameter of the logistic distribution equal to the mean parameter of the Gaussian distribution as well as applying a coefficient to the standard deviation parameter of the Gaussian to approximate the scale parameter of the logistic distribution such that the difference between the two densities is minimized.

Prior to estimating the component weights and parameter values, the empirical distributions were partitioned into two groups. For each bin of feature quantities (ranging from 5 to 15), three-fourths ($n = 1{,}500$) of the sample was randomly selected and used to estimate the population distribution parameters. The remainder of the sample was used to evaluate the goodness of fit of the estimated parameters for the population distribution. Once the optimal parameters were estimated, a one-sample Kolmogorov-Smirnov (K-S) test was performed to evaluate the goodness of fit between the estimated theoretical logistic mixture distribution and the empirical distribution of the partition of similarity statistic values that was not used to estimate the theoretical distribution parameters. This process was repeated for each quantity of features (ranging from 5 to 15). Figures C-4-1a through C-4-1c illustrate the comparison between the theoretical distributions (*k*-component logistic mixtures) and the complete empirical distribution. Figure C-4-1a overlays the cumulative frequency distributions, figure C-4-1b illustrates the P-P plots between the cumulative frequency distributions, and figure C-4-1c overlays the density distributions. Table C-4-1 provides the K-S test statistics as well as the resulting *p*-values under the null hypothesis that the theoretical mixture distribution is representative of the distribution of which the non-modeled partition was drawn. Based on these findings, the distributions exhibit little difference and thus the parametric models are proposed as plausible estimations of the population distributions for each quantity of features.

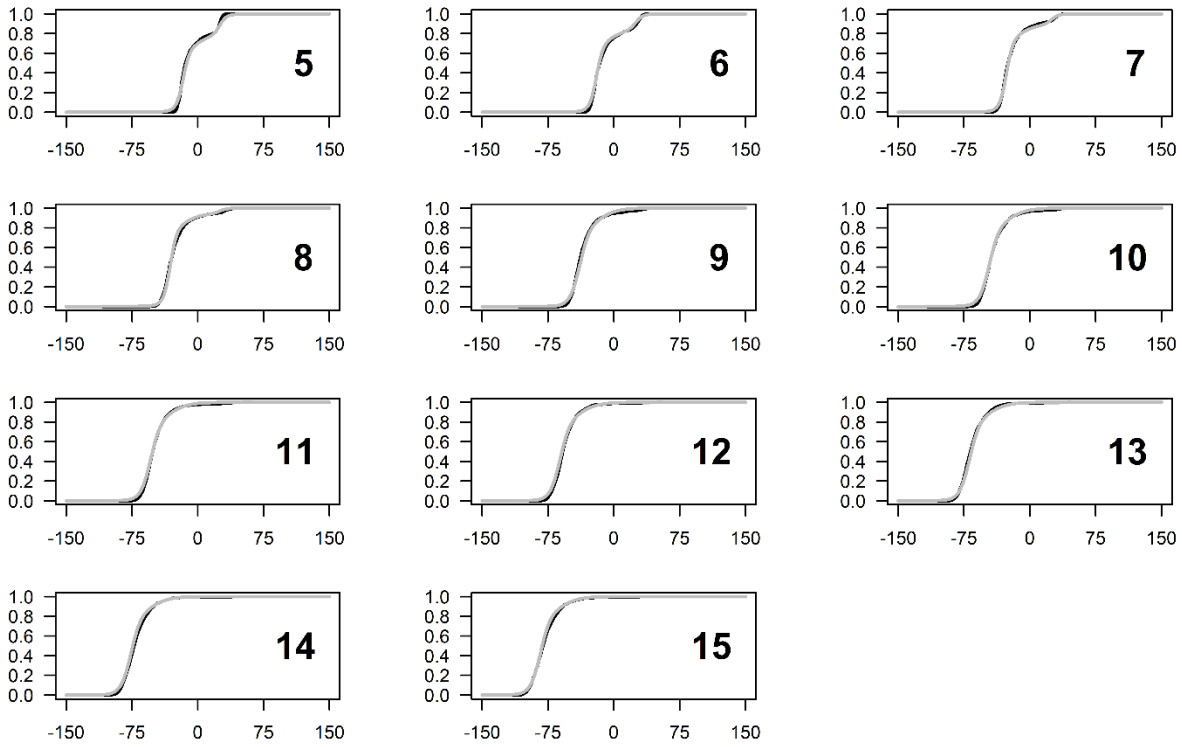


Figure C-4-1a. Cumulative frequency distributions of the similarity statistic values for the non-mated sample (empirical) compared to the theoretical (k-component logistic mixture) distribution for each quantity of features (ranging from 5 to 15). The black line represents the empirical distribution. The grey line represents the theoretical distribution. The X-axis represents the global similarity statistic values.

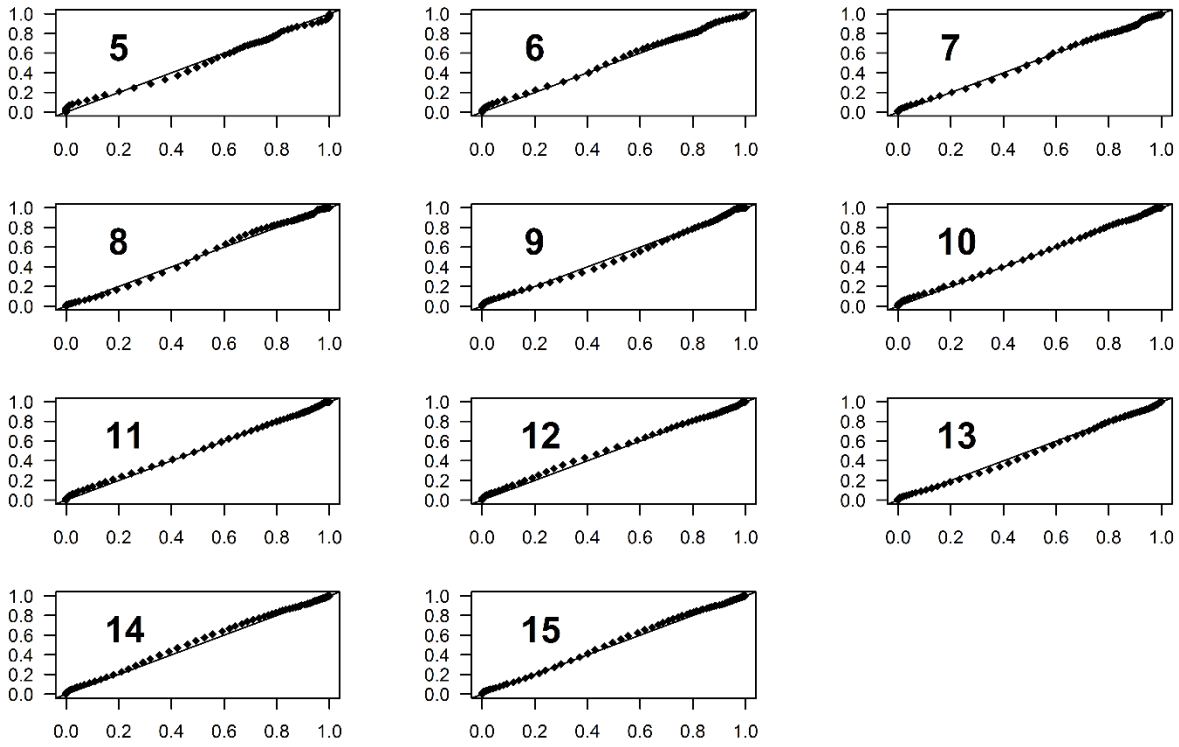


Figure C-4-1b. P-P plots of the empirical cumulative frequency distributions of the similarity statistic values (horizontal axis) vs. theoretical (k-component logistic mixture) (vertical axis) cumulative frequency distributions for the non-mated sample for each quantity of features (ranging from 5 to 15). The black dots represent the P-P plot. The grey line represents an ideal slope of 1.

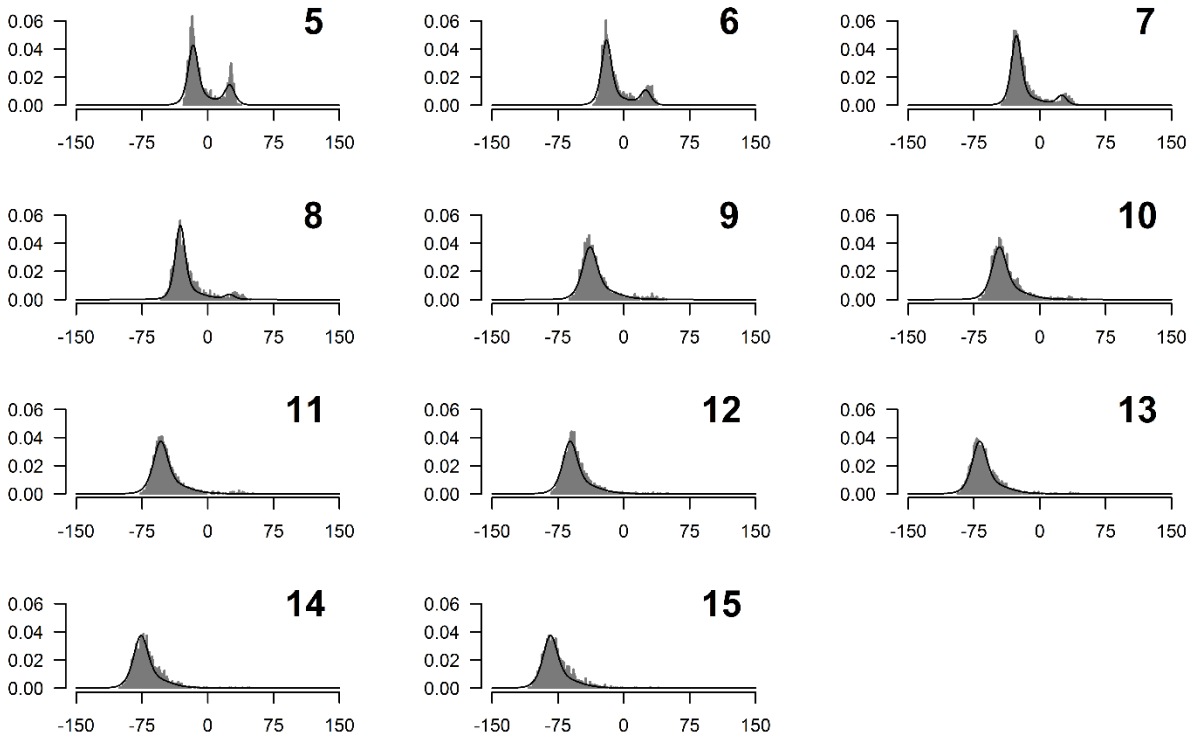


Figure C-4-1c. Empirical density distributions of the similarity statistic values for the non-mated sample (grey) compared to the theoretical (k-component logistic mixture) distribution (black) for each quantity of features (ranging from 5 to 15). The X-axis represents the global similarity statistic values.

Feature Quantity	***n* sample** (non-mated; non-estimated partition)	**K-S test statistic**	***p (null)***
5	500	0.067	$0.01 < p < 0.05$
6	500	0.071	$0.01 < p < 0.05$
7	500	0.042	$p > 0.05$
8	500	0.077	$p \sim 0.01$
9	500	0.045	$p > 0.05$
10	500	0.041	$p > 0.05$
11	500	0.055	$p > 0.05$
12	500	0.058	$p > 0.05$
13	500	0.057	$p > 0.05$
14	500	0.058	$p > 0.05$
15	500	0.070	$0.01 < p < 0.05$

Table C-4-1. Summary of the Kolmogorov-Smirnov test results between the distribution of similarity statistic values representing the partition not used to estimate the population parameters of the theoretical (k-component logistic mixture) distributions for each quantity of features (ranging from 5 to 15). NOTE: 1,500 sample statistic values were used to estimate the distribution parameters. The remainder of each sample was used to evaluate the goodness of fit. Statistical significance is based on a p-value decision threshold of 0.01.

Mated distribution parameter estimation

The empirical distributions for the mated samples appear unimodal; however, they exhibit a slight right-skew for several of the lesser feature quantities. As the feature quantities increase, the skew decreases, tending towards more symmetrical distributions. Recognizing that the same weighting functions were utilized for the mated source distributions, it seems natural that the resulting distributions are also a mixture of distributions. Taking this into consideration, the empirical distributions for all feature quantities were each modeled using *k*-component (where $k = 2$) mixtures of Gaussian distributions. Component weights and parameter estimates were determined using maximum likelihood estimation methods within commercially available statistical analysis software (JMP). Manual adjustments to the estimated weights and parameters were made to smooth trends among parameter estimates between feature quantities.

Although *k*-component Gaussian mixtures are more common, logistic distributions were applied on the basis for their heavier tails compared to Gaussian distributions. The heavier tails provide more conservative estimates of probabilities in the extreme ends of the distributions. The parameters for the logistic distribution were approximated using the estimated parameters of the Gaussian distributions. This was accomplished in the same manner as described above for the non-mated dataset by setting the location parameter of the logistic distribution equal to the mean parameter of the Gaussian distribution as well as applying a coefficient to the standard deviation parameter of the Gaussian to approximate the scale parameter of the logistic distribution such that the difference between the two densities is minimized.

Prior to estimating the component weights and parameter values, the empirical distributions were partitioned into two groups. For each bin of feature quantities (ranging from 5 to 14), approximately three-fourths ($n = 1{,}500$) of the sample was randomly selected and used to estimate the population distribution parameters. Due to the fewer samples in the bin for the feature quantity equal to 15, half ($n = 250$) were randomly selected and used to estimate the population distribution parameters. The remainder of the sample was used to evaluate the goodness of fit of the estimated parameters for the population distribution. Once the optimal parameters were estimated, a one-sample K-S test was performed to evaluate the goodness of fit between the estimated theoretical logistic mixture distribution and the empirical distribution of the partition of similarity statistic values that was not used to estimate the theoretical distribution parameters. This process was repeated for each quantity of features (ranging from 5 to 15). Figures C-4-2a through C-4-2c illustrate the comparison between the theoretical distributions (*k*-component logistic mixtures) and the complete empirical distribution. Figure C-4-2a overlays the cumulative frequency distributions, figure C-4-2b illustrates the P-P plots between the cumulative frequency distributions, and figure C-4-2c overlays the density distributions. Table C-4-2 provides the K-S test statistics as well as the resulting *p*-values under the null hypothesis that the theoretical mixture distribution is representative of the population distribution of which the non-modeled partition was drawn. Based on these findings, the distributions exhibit little difference and thus the parametric models are proposed as plausible estimations of the population distributions for each quantity of features.

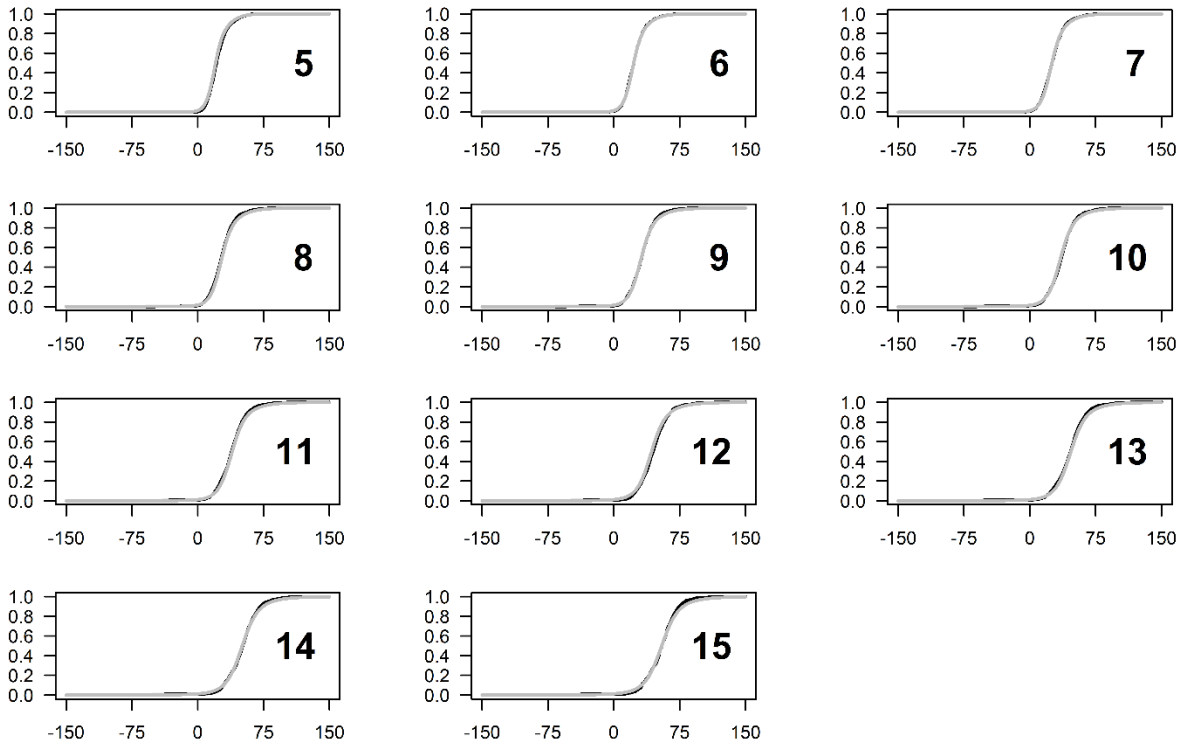


Figure C-4-2a. Cumulative frequency distributions of the similarity statistic values for the mated sample (empirical) compared to the theoretical (k-component logistic mixture) distribution for each quantity of features (ranging from 5 to 15). The black line represents the empirical distribution. The grey line represents the theoretical distribution. The X-axis represents the global similarity statistic values.

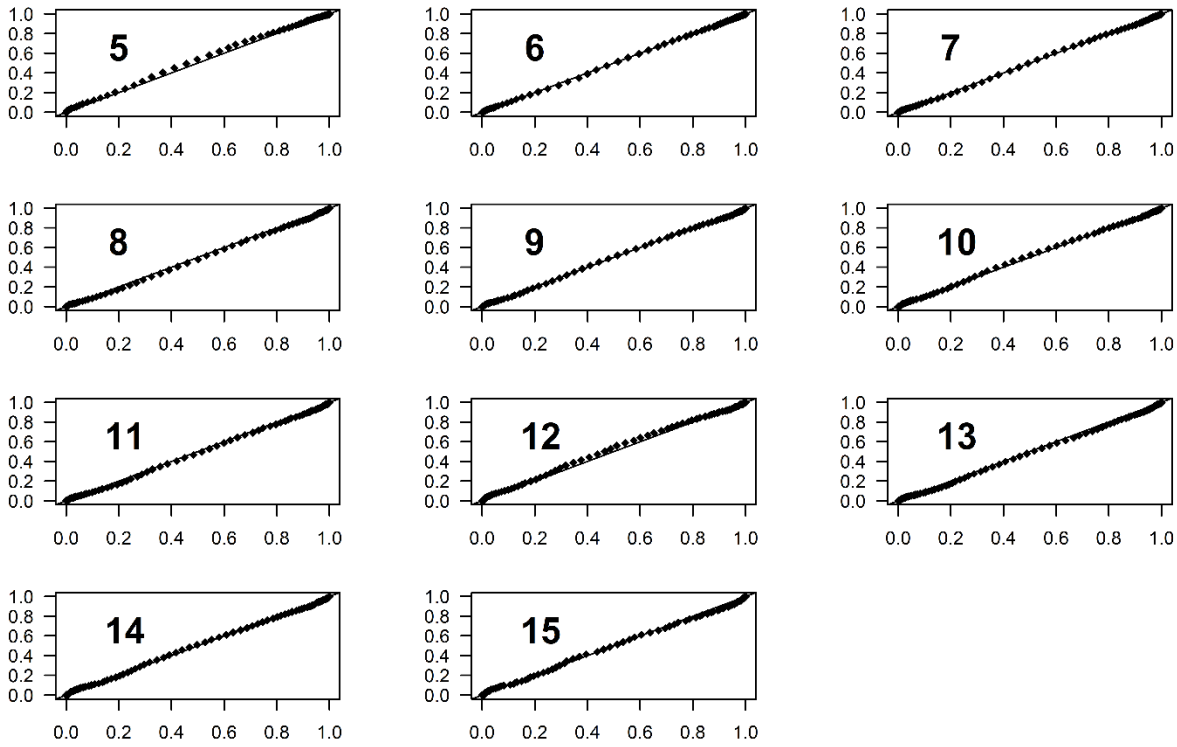


Figure C-4-2b. P-P plots of the empirical cumulative frequency distributions of the similarity statistic values (horizontal axis) vs. theoretical (k-component logistic mixture) (vertical axis) cumulative frequency distributions for the mated sample for each quantity of features (ranging from 5 to 15). The black dots represent the P-P plot. The grey line represents an ideal slope of 1.

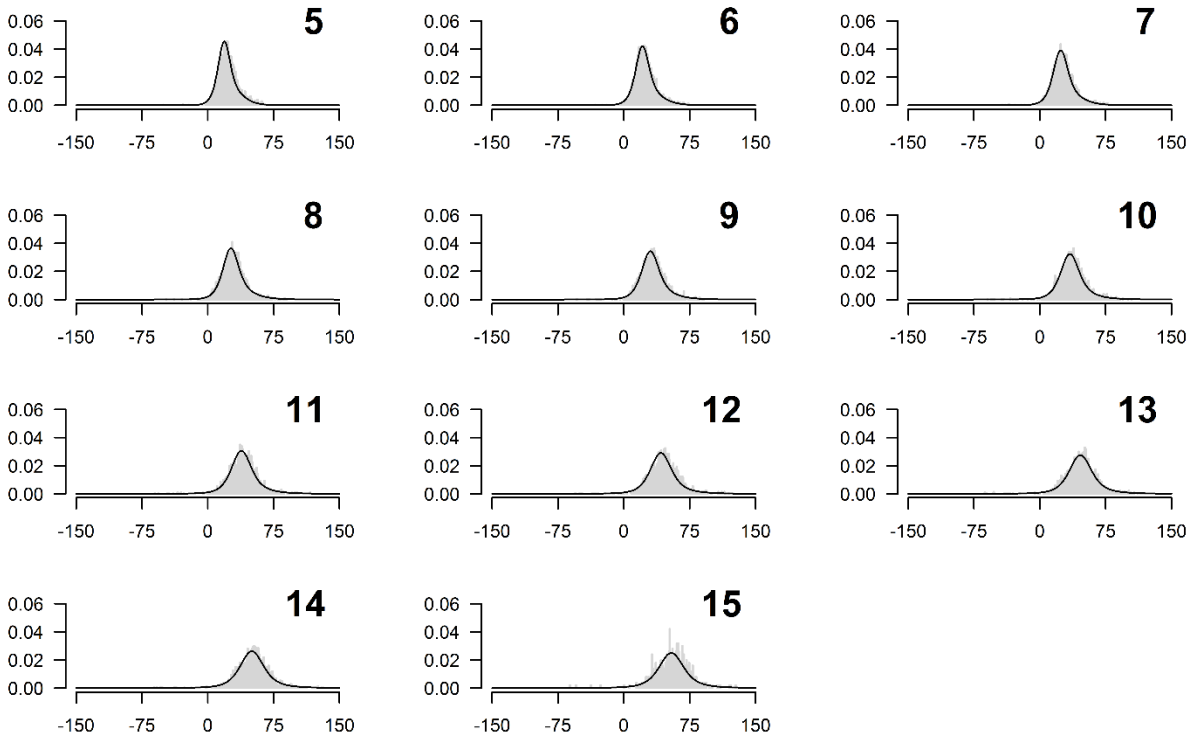


Figure C-4-2c. Empirical density distributions of the similarity statistic values for the mated sample (grey) compared to the theoretical (k-component logistic mixture) distribution (black) for each quantity of features (ranging from 5 to 15). The X-axis represents the global similarity statistic values.

Feature Quantity	***n* sample** (mated; non-estimated partition)	**K-S test statistic**	***p (null)***
5	496	0.052	$p > 0.05$
6	496	0.028	$p > 0.05$
7	496	0.032	$p > 0.05$
8	496	0.053	$p > 0.05$
9	496	0.032	$p > 0.05$
10	496	0.064	$0.01 < p < 0.05$
11	496	0.049	$p > 0.05$
12	496	0.073	$p \sim 0.01$
13	496	0.048	$p > 0.05$
14	496	0.034	$p > 0.05$
15	249	0.051	$p > 0.05$

Table C-4-2. Summary of the Kolmogorov-Smirnov test results between the distribution of similarity statistic values representing the partition not used to estimate the population parameters of the theoretical (k-component logistic mixture) distributions for each quantity of features (ranging from 5 to 15). NOTE: 1,500 sample statistic values were used to estimate the distribution parameters for feature quantities ranging from 5 to 14 and 250 were used for feature quantity = 15. The remainder of each sample was used to evaluate the goodness of fit. Statistical significance is based on a p-value decision threshold of 0.01.

14.5 Appendix C-5

This appendix provides an example demonstrating the use of *FRStat* on a fingerprint image pair from NIST Special Database 27 [65]. Figure C-5-1 illustrates the manual annotation of fifteen fingerprint features believed to correspond. Figure C-5-2 demonstrates the *FRStat* results output.

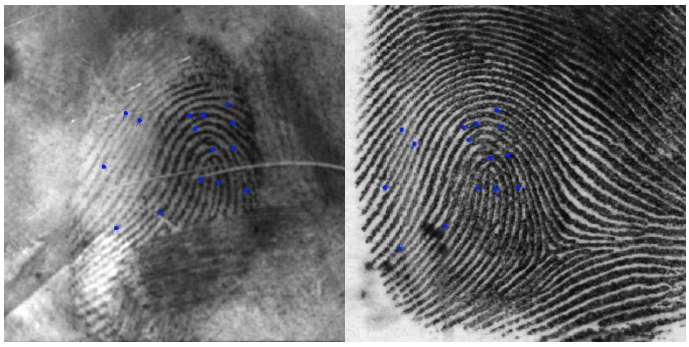


Figure C-5-1. Example latent fingerprint image (left) and corresponding reference fingerprint image (right) with fifteen features annotated (blue) believed to correspond.

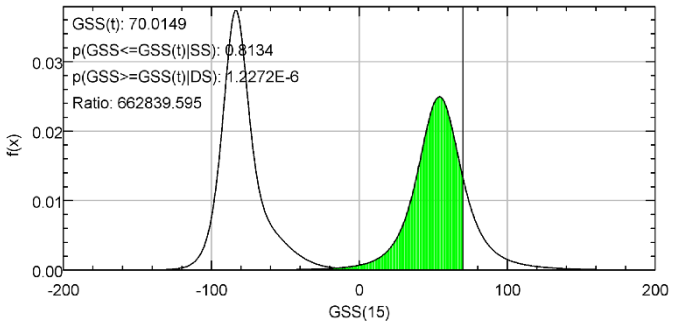


Figure C-5-2. FRStat results output from the annotated features in figure C-5-1. The Global Similarity Statistic, GSS(t), is 70.0149. The probability of observing a GSS(t) value of 70.0149 or lower among mated (same) source impressions is approximately 0.8134. The probability of observing a GSS(t) value of 70.0149 or greater among non-mated (different) source impressions is approximately 1.2272 E-6 (0.0000012272). Taken together, the probability of observing a GSS(t) value of 70.0149 is approximately 662,838 times greater when impressions are made by the same source rather than by different sources.

15 Appendix D: Supplemental Material for Chapter 5

15.1 Appendix D-1

Participant recruitment email from author #2 (Cole):

Dear Analyst,

I am an investigator for CSAFE, the Center for Statistical Applications in Forensic Evidence (http://forensicstats.org/), and a Professor in the Department of Criminology, Law & Society at the University of California, Irvine. CSAFE has partnerships with many forensic laboratories including the Midwest Forensic Resource Center. As you probably know, CASFE was named a Forensic Science Center for Excellence by the National Institute of Standards and Technology to "support NIST's efforts to advance the utility of probabilistic methods to enhance forensic analysis." As part of this mission, we are conducting a survey of forensic practitioners about current practices regarding probabilistic reporting of forensic results. We are particularly interested in your experiences with probabilistic reporting. We request your participation in this survey. You are eligible to participate in this study if you are 18 years of age of older and a forensic practitioner. Please note that your participation in this study is completely voluntary and declining to participate will not in any way affect your standing with your employer. There are no direct benefits from participation in the study. However, this study may help us better understand the impact of probabilistic reporting on practitioners. We expect the survey to take around 60 minutes. We thank you in advance for your cooperation.

15.2 Appendix D-2

The closed-response survey question in which participants were asked to choose which of three sample statements most closely resembled the wording they currently used in reports of an association between two friction ridge impressions.

Question: Which of the following most closely resembles the language used in your written examination reports when reporting an association between two prints in your practice?

*A) "The latent print on Exhibit ## was identified to the standards bearing the name XXXX." //**OR**// "There is sufficient quality and quantity of detail in agreement to conclude that the latent print on Exhibit ## and the standards bearing the name XXXX originated from the same source." //**OR**// "The latent print on Exhibit ## matched the standards bearing the name XXXX."*

*B) "The latent print on Exhibit ## and the standards bearing the name XXXX have corresponding ridge detail. The likelihood of observing this amount of correspondence when two impressions are made by different sources is considered extremely low." //**OR**// "The latent print on Exhibit ## and the standards bearing the name XXXX have corresponding ridge detail. This amount of correspondence provides extremely strong support for the proposition the two impressions were made by the same source rather than*

*by different sources." //**OR**// "The latent print on Exhibit ## and the standards bearing the name XXXX have corresponding ridge detail. The probability of observing this amount of correspondence is approximately ## times greater when impressions are made by the same source rather than by different sources."*

C) *"The features within the impressions are clear and there is an abundant amount of consistency (much more than needed to arrive at a conclusion) with no inconsistencies. The conclusion is easily demonstrable to others and others would be hard pressed to find any reasons to doubt the conclusion. There is a large amount of data that infers that the latent impression was deposited by XXXX."*

15.3 Appendix D-3

A full listing of all the statements for which researcher coding conflicted with self-report.

Number	**Statements self-reported as "categorical" but coded by researchers as "probabilistic"**
1.	*Item ###. The partial latent print of value has characteristics in agreement with the fingerprint/palm print impressions of XXX, Item ###, (finger # & label/left or right palm). In the opinion of this examiner the likelihood that the impressions were made by a different source other than the one listed is very small.*
2.	*I am of the opinion that the latent prints from item #X and exemplar fingerprints from XXX likely originated from the same source.*
Number	**Statements self-reported as "probabilistic" but coded by researchers as "categorical"**
1.	*Amount of agreement between two impressions is compelling, inferring both originated from the same individual (i.e. it is not plausible that the impression originated from a different source).*
2.	*"The likelihood the ff. print impressions were made by another source is so remote that it is considered a practical impossibility"*
3.	*Print Quality: Sufficient for a comparison and an ABIS search. The XX finger of (name/DOB/record#) was identified as having made this print. This is followed by a qualifying statement: Identification: the decision by a qualified examiner, that two friction ridge impressions originated from the same source. The features present in the two impressions are in sufficient correspondence, and the probability the questioned impression was made by a different source is so small, it is negligible.*
4.	*he latent print on exhibit ## and the exemplars bearing the name XXX have corresponding ridge detail. The likelihood of encountering this much correspondence between two different individuals is considered extremely low and I have discounted it.*
5.	*The likelihood the following print impressions were made by another (different) source is so remote that it is considered as a practical impossibility:*
6.	*The impression on the latent card was made by the same source as the known fingerprint/palm print standards as XXXX*
7.	*Prints recovered from Item XXXX have been analyzed, compared evaluated and identified with the known prints of XXXX. The card bearing the know prints of XXXX used in this*

	identification was obtained on "DATE" by "NAME", an employee of "WORKPLACE". Source identification is the opinion of the examiner that two friction ridge skin impressions originated from the same source. This opinion is the decision that the features are in sufficient correspondence and that the probability the questioned impression was made by a different source is so small that it is negligible.
8.	*The latent lift labeled L-1 was labeled as being lifted from the _____ by Officer ____. It is my conclusion that the latent lift labeled L-1 was made by the right thumb of _____.*
9.	*Identification to Name, SID # along with this statement, An identification does not necessarily eliminate the possibility that another person in the world could leave a print with areas of similar agreement. Identification means that within the examiner's experience and knowledge, no other prints with this much similarity have come from different people.*
10.	*The latent prints of comparison value were compared to the standards for XXX with the following conclusions: F1A - Identified to the right middle finger*
11.	*Latent print card 1 of 1, collected from " insert", has been identified to the right thumb of the fingerprint cared bearing the name XYZ*
12.	*The latent print, L-X, was identified as the XXXX XXXXX of XXXX XXXX XXXX.*
13.	*Latent impression on lift 0450 originated from the same source as the known right index finger of .....*
14.	*I formed the opinion that fingerprint R1 and the known fingerprint of XXXX were made by the same finger.*
15.	*Latent # is identified as the number ... finger of John Doe*
16.	*A visual examination of LL#1 and the ### finger of fingerprint card bearing the name XXXX have sufficient detail to conclude an identification.*
17.	*"identified"*
18.	*A match was made between John Doe ********* and the latent print lifted from the inside passenger door's window by **********. The match was verified by **********.*
19.	*Suitable detail within "R2", has been compared with the fingerprint images associated with a Form RCMP C- 216/booking record dated 2017-04-27 bearing the name and particulars of subject... SUBJECT KO ##### d.o.b. 1992-05-21 ...specifically with the impression image(s) of the LEFT THUMB. In the Evaluation phase of the ACE-V process, I considered all of the information gathered during Analysis and Comparison to reach conclusions about the origin of the latent print. As a result of the evaluation of corresponding ridge features observed during comparison, the writer concluded INDIVIDUALIZATION (2) , in other words that the impressions from both sources (R2 <> Lt. Thumb print from record of SUBJECT) were caused by the same donor person. The Verification step of the ACE-V (1) process consists of an independent and blind application of the ACE process by a subsequent examiner to either support or refute the conclusions of the original examiner. In the case of the individualized impression(s) above, a vetted version of the case file was forwarded to the FSU Blind Verification Coordinator for assignment to a second analyst for verification or invalidation. D/Cst. *******conducted this review and reported conclusion(s) back to the coordinator. There were no conflicts of opinion. *Note #1***"Notice(s) of Intention - Expert Opinion / Report" under s.657.3 CCC should be served upon the accused or his/her counsel at least 30 days prior to trial date for the introduction of expert testimony. See images for Draft copy of this Form. ***Note #2***Should the introduction of fingerprint evidence be required at trial, the following individuals should be included on the list of witnesses; Cst.*

	********* - recovered prints - submitted for analysis S/Cst. ********* - analyzed, compared, identified fingerprint R2 D/Cst *****conducted independent comparison process verifying conclusion(s) S/Cst. tba- Cell Officer who fingerprints the subject , if arrested, for this general occurrence.***Note #3***The presence of a friction ridge print on an item of evidence indicates contact was made between the source and the item of evidence. The presence of a friction ridge print alone does not necessarily indicate the significance of either the contact or the time frame during which the contact occurred.(1) ACE-V - The acronym for a scientific method: Analysis, Comparison, Evaluation, and Verification (see individual terms at www.swgfast.org).(2) Individualization -The decision by an examiner that there are sufficient features in agreement to conclude that two areas of friction ridge impressions originated from the same source (donor). 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.*
20.	*The writer has compared fingerprint impression"R#" described as from "location of impression", with the fingerprint images appearing on a form RCMP-C216 bearing the name, particulars, and portrait image of Name (DOB: YYYY-MM-DD) KO #######, FPS ########. As a result of the evaluation of corresponding detail observed during this comparison phase, the writer has concluded individualization" in other words that the fingerprints from the aforementioned sources, the latent print "R#", and the "specified digit" fingerprint purported to be that of Name, were caused by the same donor-person. Individualization - The decision by an examiner that there are sufficient features in agreement to conclude that two areas of friction ridge impressions originated from the same source (donor). 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.	*the Latent print was identified to XXXX*
22.	*impression x was identified as the such and such finger/palm of John Doe (sid#). dentification is the opinion of an examiner that there is sufficient quality and quantity of detail in agreement to conclude that two impressions originated from the same source.*
23.	*Latent print X was identified to the exemplars bearing the name X*
24.	*I compared the latent print from card (#) to the tenprint card of (subjects name) using the ACE-V method. I identified the latent print as being made by the (# finger/palm) of (subjects name). The identification was verified by (Forensic Specialist).*
25.	*Latent(s) XX was/were compared to the exemplar prints bearing the name(s) of the above listed subject(s) and was/were IDENTIFIED in the following manner:*
26.	*Comparison of the latent print to the known prints listed above revealed there is sufficient information in agreement based upon features, sequence, and spatial relationship to conclude that the latent fingerprint of Lab Item #1 was made by the same individual whose known prints appear on Lab Item 2.*

16 Appendix E: Supplemental Material for Chapter 6

16.1 Appendix E-1

PARTICIPANT INVITATION LETTER

Dear Participant,

I am a doctoral candidate pursuing a degree in forensic science through the University of Lausanne, Switzerland, under the direction of Dr. Christophe Champod, Professor of Law, Criminal Science and Public Administration. The focus of my research is on the development and implementation of computational algorithms for forensic science. As part of this research, I am conducting a study to explore the perspectives of criminal justice stakeholders (laboratory managers, prosecuting attorneys, defense attorneys, judges, and other stakeholders [e.g., academic scholars]) as it relates to issues concerning the use of probabilistic reporting (with or without algorithmic tools) and the use of computational algorithms in forensic science for court purposes.

I am writing to invite you to participate in this study and contribute to this broader effort. Our interest in this topic is broad and includes technical, operational, and legal dimensions. Our ultimate objective is to characterize various stakeholder perspectives on these issues to enable a path forward for the forensic science community as it relates to the use of probabilistic reporting and computational algorithms in forensic science. We aim to enroll approximately fifteen participants (three from each stakeholder group). This study will be conducted as a semi-structured interview of each participant lasting approximately one-hour. Identities of participants will be kept confidential and not publicly disclosed. Participation is by invitation only and selections are based on participants having been actively engaged in issues concerning forensic science policies, procedures, and practices.

For your convenience, I have attached the following items:

(1) Participant information and consent form
(2) A short description of the purpose and background of the study
(3) A one-page guide outlining the structure and questions that will guide the interview
(4) My curriculum vitae outlining my professional background and experiences related to issues concerning forensic science

I hope you will accept my invitation to participate. Please let me know if I can answer any questions or be responsive to any concerns you might have.

NOTE: Although my academic pursuits are through the University of Lausanne, I am physically located in Washington, D.C. and the research is focused on issues concerning forensic science practices in the United States.

16.2 Appendix E-2

PARTICIPANT INFORMATION AND CONSENT FORM

Title and Summary of the research

Probabilistic Reporting and Algorithms in Forensic Science: Stakeholder Perspectives within the American Criminal Justice System

Over the last decade, with increasing scientific scrutiny on forensic examination and reporting practices, there have been several efforts to introduce probabilistic reasoning and computational methods (i.e., algorithms) into forensic practice. Although various approaches have been proposed, reactions to probabilistic reporting and the use of algorithms in forensic science have been mixed. This research is aimed at exploring the perspectives of key criminal justice stakeholders (laboratory managers, prosecuting attorneys, defense attorneys, judges, and other stakeholders [e.g., academic scholars]) to improve our understanding of the issues related to the use of probabilistic reporting practices (with or without algorithmic tools) and the use of algorithms in forensic science for court purposes.

The current study is conducted by Henry Swofford, under the supervision of Dr. Christophe Champod, Professor at the School of Criminal Justice, University of Lausanne, Switzerland.

A) PARTICIPANT INFORMATION

1. **Objective of the Study**
   This study focuses on the issues concerning the use of probabilistic reporting and computational algorithms in forensic science for court purposes by exploring the perspectives from various key stakeholders in the criminal justice system (e.g., laboratory managers, prosecuting attorneys, defense attorneys, judges, and other academic scholars). Our interest in this topic is broad and includes technical, operational, and legal dimensions. Our ultimate objective is to characterize various stakeholder perspectives on these issues to enable a path forward for the forensic science community as it relates to the use of probabilistic reporting and computational algorithms in forensic science.

2. **Procedure**
   If you agree to take part in this study, you will be asked to participate in a semi-structured interview lasting approximately one hour, at a time of your convenience using a virtual meeting platform.

   This study aims to enroll approximately fifteen participants (three from each stakeholder group). Participation is by invitation only and selections are based on participants having been actively engaged in issues concerning forensic science policies, procedures, and practices.

3. **Confidentiality**
   The interview will be digitally recorded to facilitate subsequent analysis. The information gathered will be kept strictly confidential and will be destroyed at the end of the study. You will be assigned a unique identifier within your respective stakeholder group; however, your personal identity will not be disclosed or publicly attributed to any specific statements.

4. **Voluntary Participation/Withdrawal from the Study**
   Your decision to participate in this study is entirely voluntary and no compensation will be provided. You are free to withdraw at any time by simple verbal notice. Any personal data gathered prior to your withdrawal will be destroyed.

5. **Conflict of Interest**
   No one on the study team has a financial interest related to this research.

B) STATEMENT OF CONSENT

1. **Declaration by Participant**

   I have read the preceding information thoroughly. I have had the opportunity to ask questions, and all of my questions have been answered to my satisfaction. I understand the purpose as well as the nature of the study.

   I hereby consent to take part in this study:

   Participant's Name:_____

   Signature:_____Date:_____

2. **Declaration by Researcher**

   I have given a verbal explanation of the research project to the participant, and have answered the participant's questions.

   Researcher's Name:_____

   Signature:_____Date:_____

C) CONTACT

For additional information about the research, please contact Henry Swofford at henry.swofford@unil.ch or Dr. Christophe Champod at christophe.champod@unil.ch.

16.3 Appendix E-3

STUDY PURPOSE AND BACKGROUND

Title and Abstract of the research

Probabilistic Reporting and Algorithms in Forensic Science: Stakeholder Perspectives within the American Criminal Justice System

Over the last decade, with increasing scientific scrutiny on forensic examination and reporting practices, there have been several efforts to introduce probabilistic reasoning and computational methods (i.e., algorithms) into forensic practice. Although various approaches have been proposed, reactions to probabilistic reporting and the use of algorithms in forensic science have been mixed. This research is aimed at exploring the perspectives of key criminal justice stakeholders (laboratory managers, prosecuting attorneys, defense attorneys, judges, and other stakeholders [e.g., academic scholars]) to improve our understanding of the issues related to the use of probabilistic reporting practices (with or without algorithmic tools) and the use of algorithms in forensic science for court purposes.

The current study is conducted by Henry Swofford, under the supervision of Dr. Christophe Champod, Professor at the School of Criminal Justice, University of Lausanne, Switzerland.

Purpose:

To explore perspectives from key criminal justice stakeholders related to interpretation and reporting practices (with or without algorithmic tools) and the use of computational algorithms in forensic science for court purposes.

Background:

For purposes of this study, we focus on the use broad use of computational algorithms in traditional forensic science disciplines (e.g., DNA, facial recognition, fingerprints, footwear, firearms, handwriting, etc.) for court purposes. Computational algorithms used outside of traditional forensic science disciplines or for non-court purposes, such as for investigatory or intelligence purposes, are outside the scope of this study.

For purposes of this study, the following terms and definitions apply:

- Appropriateness: the quality of being suitable or proper in the circumstances.
- Validity: the quality of being logically or factually sound. E.g., the extent to which a conclusion is logically justified and factually supported by the measurements resulting from the examination performed.

- Categorical (reporting): The expression of results in a manner which does not formally recognize or articulate the uncertainties inherent in the interpretation or the possibility for an alternative proposition to be true. NOTE: categorical reporting includes unadorned statements of "match," "identification," or "individualization" without articulation of an associated error rate, statements that attribute two samples as being made by the same source, and statements that encourage uncertainties to be disregarded (e.g., "practical impossibility," "negligible," "discounted").

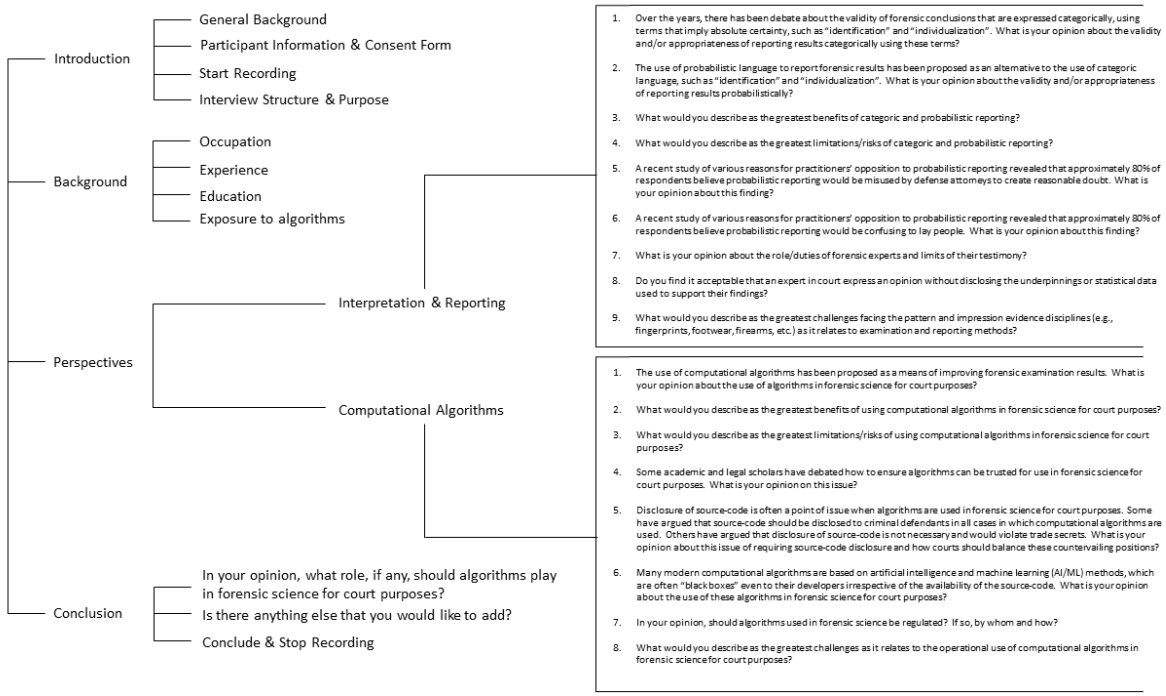
- Probabilistic (reporting): The expression of results in a manner which formally recognizes and articulates the uncertainties inherent in the interpretation and the possibility for an alternative proposition to be true. NOTE: probabilistic reporting includes statements that are expressed as a likelihood ratio, posterior probability (of a proposition), match probability, or any other expression which is accompanied by an explicit statement of recognizing the possibility for an alternative proposition to be true.

- Computational Algorithm: A broad term to describe any computer and mathematically-based prediction method, such as statistical models or other defined sets of computer implementable mechanical processes used for forecasting, predictions, statistical evaluations and decision making. NOTE: computational algorithms include human-interpretable rules or processes as well as non-human interpretable processes, such as those developed through artificial intelligence and machine learning (AI/ML).

For purposes of this study, the following terms are intentionally not defined to ensure we do not limit participants' responses: "benefit(s)," "limitation(s)," "risk(s)," and "trust."

16.4 Appendix E-4

INTERVIEW GUIDE



16.5 Appendix E-5

PARTICIPANTS' ELABORATED RESPONSES RELATED TO INTERPRETATION AND REPORTING PRACTICES

Participants' responses related to the validity and/or appropriateness of reporting results categorically:

Laboratory Managers

I know there's been a lot of controversy on [expressing conclusions categorically]. ... And I do understand there's comparisons where people have maybe overstated their data, a bite mark, smudged fingerprints, [for example; however,] if I've got a broken taillight with tremendous detail where I can fit it back together, I'm not doing anybody a service of saying "it could have come from there or really hedging my opinion when anyone, even a layman, can look at it and go, 'that came from there'" ... it is my opinion. Even though it's considered maybe by some, a little taboo, it's still individualizing. There are still cases where that's the case. ... [but] I do believe people are overgeneralizing [to lump all comparisons together] (LM#1).

Nothing's absolute. I get the drive of why those answers came to be, [such as] "beyond a scientific certainty," or whatever other dumb language gets used. The reality is nothing, nothing is absolute. Everything has some level of uncertainty to it. So fundamentally, I guess, I disagree with the kind of categorical absolute statements. It's just not the case. That being said, I do understand how that gets driven there because the question in court is "how certain are you?" Well, how do you put some kind of framework around that? And I think every discipline actually struggles with that. ... Not to jump off into anything too entirely philosophical [but] there is uncertainty in the entire system. I think it is an inherent obligation to try and express the limitation of the result. And that limitation can take a lot of different forms, but that you have at least attempted to express the limitation of it. ... I've read through a number of [pattern evidence examiners'] testimonies ... [and] they say, "my opinion is," not "the result is" or "these two are a match," but [they say] "in my opinion ... this is my opinion." They have been very careful to qualify it as an opinion. I can live with that because they are the instruments and their assessment of these images is their trained opinion. It can be a very weighty opinion, but given their opinion how do you provide any kind of uncertainty on how good is their opinion? Well, blind testing starts to give you a bit [of insight] (LM#2).

Absolutes and conclusions, I think, are probably inappropriate. I, however, do not have a problem with experts giving their opinion. I think we have very good experts. I think expertise matters. I think exposure to casework matters. I do agree with a lot of the defense experts and the academics that we need a reasonably good way to express uncertainty. ... It's a little fuzzy right now [of how to express that uncertainty, but] this is the nature of black box studies, [as well as] ... talking about the person's experience, talking about the number of samples this person has looked at, or the number of positives they've made, the

number of exclusions they've actually made, the amount of training [they have had], the breadth [of their training]. … The black box studies, I think, are an attempt to provide numerical, objective measures of those conclusions that experts actually draw. … It's really a holistic approach, not just an absolute number that you get with digital data (LM#3).

Prosecutors

I don't think saying identification implies absolute certainty. And, individualization, I don't think that to a layman that implies absolute certainty. ... And I know that in the past, specifically in latent prints, and then also I believe in firearms and tool marks, there would be additional certainty statements beyond that, like, to the exclusion of all other firearms in the world, or a hundred percent certainty. I agree that those types of statements should not be used, those outside certainty statements. But I actually think that the categorical approach and pattern evidence, at least based on what I've seen so far, is probably the most appropriate, scientifically appropriate, and also easily understandable way of reporting conclusions to both litigants and ultimately the fact-finder, be it a judge or a jury (P#1).

I'm not sure that is the case. I think there's a debate that's going on in the forensic community and certainly amongst some of the legal community about whether those terms imply [absolute certainty]. I don't think that they imply that. ... Now there's going to be limitations to what they can also say related to that identification. And I think that's well put by the Department of Justice Uniform Language with Testimony and Reports. ... But the statement itself, that this is this expert's opinion, that this is an identification. I have no pause about that (P#2).

I don't have a problem with the use of a categorical response. It's easy to understand. It's easy for the jury to grasp, and I believe that it is the true opinion of the scientist who's giving us that opinion. I think that the backing away from the certainty and the different locutions that have been proposed, I think those are imposed on the analysts or the scientist. ... I think that given their own druthers, they would say forthrightly, it's a match. And I don't think they would say it's a match if they didn't believe it was a match (P#3).

Defense Attorneys

I think it's particularly problematic to be offering something like an individualization conclusion. ... My personal opinion is that fingerprints probably are unique, but that's not a scientific validated fact, and we shouldn't express those opinions in court as though they were. I think what's probably more important is that we don't know how similar prints can be and in latent print examination we are dealing with very small amounts of data. ... I don't think it ever should be that person to the exclusion of every other human ever born on the planet. You know, it's funny that virtually anybody will agree with it when it's couched like that in this hyperbolic language. But the reality is that's virtually every opinion that you ever get introduced in court is essentially that. ... If you're going to make

an association at all, it should never be categorical, and the association should always allow for the possibility of error or the possibility of a random match (D#1).

These terms as they have been used by forensic examiners have always been misleading and they have resulted in systematic overstatement of the forensic evidence in ways that have been detrimental to people charged with crimes all over the country and in a way that most in the forensic science community have failed to come to grips with even today. … The desire to cling to claims of certainty have been so overwhelming and blinding that it largely took scientists and other concerned parties from outside of the forensic science community to tell forensic examiners to stop claiming such certainty (D#2).

There's a tremendous amount of concern. Specifically, because there's essentially no scientific foundation for the claims of identification that are being made in almost all of the pattern disciplines. We've been objecting to that for many, many years and trying to explain to the courts why the validation studies that are put forward in some of these disciplines don't support the statements that are being made (D#3).

Judges

As I think many people know, bad forensic science has been an element in the conviction of innocent people. ... One of the reasons that those inaccuracies [in forensic science] came about [was] because the science itself was much more subjective than was represented to courts and to juries, [and] because they were presented as being certain conclusions. [For example,] in many states to this day, forensic scientists cannot testify unless he or she is prepared to state that they have reached their conclusion to "a reasonable degree of scientific certainty." Now, if you look at the case law, the case law often says that the courts regard that as nothing more than more probable than not—very far from certain—but what the jury is hearing and what judges who are not familiar with this case [law] are hearing is "scientific certainty." There's almost no part of science that can claim certainty. If you talk to physicists or chemists or whatever, they won't claim that. Yet here it is, in effect, being claimed by forensic science. ... [Although experts are allowed to express their opinion,] in the federal system, it's not good enough for someone to say, this is what I believe. They have to show that it has a scientific basis through a Daubert analysis. ... [Among several other factors] you have to show that it doesn't have a significant error rate. This is where a lot of forensic science fails. You have to show that the methodology used is reliable so that it could be used again and again, with the same results and with two different examiners coming up with the same results. That often is not correct or not true in certain types of forensic science. ... [This is applicable] [i]f it's being offered as science. Now, if it's not offered as science, all those requirements still apply, but they are applied more flexibly (J#1).

I think it's very challenging to use [categorical statements] for purposes of how to report a result. Part of that is because I don't know that they're particularly well-defined and my experience seems to be that attempts to define those categorical statements do not always align with what a lay person's understanding of the definitions would be. ... So, from that standpoint, I do think that it can be challenging. ... [Further,] as far as an expert testifying

based on the "training and experience" [and] "to a reasonable degree of blank certainty," whether that's "ballistic certainty" or whatever, I think that has no place in a courtroom, period. I don't think it has a definition. ... But, as far as source attribution itself, ... as long as [the examiners] can back it up, especially if, as they explain[ed] their analysis, they explained the things about their opinion that have to do with the uncertainty so that they're saying the parameters and limitations of what it is that [they] can tell you, [then] that that's fair game. ... [But] do we really know what ground truth is and how do we get there? How do I know that there's the foundational science to be able to say that, as we're doing this comparison, that I can make the statement, "yes, this impression came from this source?" We get into [things] like, "well, it's a match." Well, okay. It may be, [but] how do you know that (J#2)?

Other (Academic Scholars)

I would rather see a complete move away from using those terms. Accepting that that move takes time, in the interim I think there [should] be a clear explanation given alongside the terms, that they are only opinions and that two examiners wouldn't necessarily come to the same opinion in a more complex case (O#1).

I think it's clearly not justified scientifically. It's an overstatement of the value of the evidence. We know it's simply not plausible for a discipline, like fingerprinting, that a trained examiner can determine the rarity of the set of features observed [based solely on human judgment] with the precision necessary to know whether it's probability in the population is low enough to support the claim that it's a unique observation. From my standpoint, the psychology of human judgment, it's totally implausible that anybody could make that kind of judgment. ... It's not just unjustified, from my point of view it's a prosperous claim. It's a laughable claim. ... Even then, black box studies show the error rates are low, but not so low as to justify a claim of identification, so it's empirically debunked too. ... [Further,] Daubert [admissibility standard] doesn't or shouldn't allow you to get away with saying, "well, there's no scientific basis for this opinion, but it's my opinion anyway, so I'm going to say it." ... From my point of view, forensic scientists should not be claiming they can identify things because it's a clear overstatement of the value of the evidence. It's a claim that they can do something that they clearly cannot do. It claims too much. They need to acknowledge the potential for error and uncertainty. ... Even couching it as an opinion should not be allowed. I guess the question is do the legal standards for admissibility of evidence, such as Daubert, allow experts to present something which does not have an adequate scientific basis so long as the experts say, "but it's just my opinion." Answer: No! Saying "it's just my opinion" sort of softens the claim a little bit, [but] I think we'd be much better off if we had some algorithm that would put a weight or strength of evidence statistic on it, even if it supported claims far less extreme (O#2).

I think it is wrong. I think it's immoral to stand in front of a jury and make categorical statements if you are a forensic scientist because the word "scientist" confers in the minds of the jury that you are, well, one way that I heard it expressed is that the words have totemic power. I think it's wrong to abuse that level of trust. ... I'm all for making sure that

statements in front of the jury, in front of any trier of facts, when presented as forensic science must be accompanied by how certain I am based on the scientific principle about the statement. ... If someone is an expert then you can give an opinion, but if you claim your opinion is based in science, then that constrains you to give a scientific opinion. ... To say that I have absolute certainty on the basis of my experience, you have gone right outside of the role of science. ... Look, the way I view it, we can make categorical statements, but don't claim it's backed up by science (O#3).

Participants' responses related to the validity and/or appropriateness of reporting results probabilistically:

Laboratory Managers

I like [numbers] because it provides [context]. On the other hand, even numbers have their limitations. … How do you throw somebody just a number and expect them to understand it? ... It's still not standalone. You just don't lob the grenade of the number in [the courtroom] and run out. You have to explain it. ... The danger is people just start to rely on a number, [but] what does that mean? That's where the expert [comes in and explains] what this means. … Our job is as witnesses, as opinion witnesses, as expert witnesses, is to provide that interpretation. … So, [the] danger is you let [numbers] do too much of the talking, but it's tremendously good when you use it to inform your opinion. And, that's part of the whole package. It's better with it than without it, if done correctly (LM#1).

From a philosophical standpoint, I think it is more appropriate. What I see though, is a hell of a lot of confusion on the part of the lay person and lawyers and juries. There is a seductive, unfortunate, seductive nature to it that it seems like it should be an appropriate way to express things that allow the fact-finder to understand how reasonable it is to make their decision on those data. The reality is as humans, we suck at basic risk assessment, and you are giving people a number to base against their personal experience and their biases to assess risk. So, I guess it's better than categorical, but it has its own pitfalls (LM#2).

I don't have a problem with [probabilistic reporting] per se. ... I think some standardization in the industry is warranted and good for juries so that when one person says it versus another person, we're at least talking on the same playing field rather than completely different contexts. Some of the terminology is difficult for juries to understand. ... I think we have to be concerned about how juries are going to take the information. I think this is why a forensic scientist as an expert is so important because they bring the human element to interpretation and explanation that I think is so critical. ... [That said,] I struggle a little bit with reporting a number unless you have a numerical basis for it. ... I have no problem with subjective interpretations [such as] "in my experiences," [or] "is very likely," just as a subjective conclusion, but if you're going to put a number on it, I think you need to have some basis [of] where you're pulling the number from (LM#3).

Prosecutors

So obviously probabilistic language has been used in reporting DNA results forever. ... I don't have any information or knowledge as to how something similar would be done in a pattern discipline. ... I would be open to considering it ... but I think that the analyst's report or the scientist giving their conclusion, about whether, for example, a particular tool could be the source of a particular tool mark is sufficient. ... [However,] ... prior to trial, the risks are the reader is going to misunderstand the value of the opinion. I think that risk exists, whether you use categorical language, or whether you use more likelihood kind of language in a report. And actually, I would say in my experience, the more complicated the written opinion in the report gets, the more likely that law enforcement and prosecutors, at least at first blush, are going to be confused by it without calling and having a conversation with the analyst. ... [Ultimately,] ... if the experts themselves decide that articulating their opinions in that manner is more scientifically accurate than in, let's say a categorical kind of statement, if they decide that they think that that more correctly states their scientific opinion, then I would be fine with that (P#1).

A probabilistic conclusion is a lot looser and as a result is much less clear what that means. So, when you say more likely, what does that really mean? More likely to whom? To you? To me? To the juror? What if we have different standards about what's more likely and what's less likely? What's negligible to me may not be negligible to you. So, I think we all have very different standards there and that looseness, I think gets us into muddy waters, and can cause more problems for both defense and for prosecution because both sides may manipulate that language in a way that's unfair since it's looser (P#2).

I think both of the statistics that we use in DNA, the RMP, which, describes the rarity of a profile in the population or a likelihood ratio, which pits two hypotheses against each other and tells you which one is more probable given the evidence. I think both of those are basically easy to explain to a jury and easy to explain to prosecutors and the court and it was relatively easy for me to understand, and I have no science background. ... But I don't see where the numbers come from in pattern matching. ... I just don't see how that's possible. ... Or where even if you're not using a numeric value, if you're using a verbal scale, I just don't know where that data comes from. ... I think it's less precise and I think it's actually way more subjective because, you know, strong support, very strong support, ... what does that mean? I think it means different things to different people and numbers mean different things to different people. ... So, I think it's problematic (P#3).

Defense Attorneys

I think as we sit here today, when those types of conclusions are offered at trial, there's no real empirical support for that. ... What I think is that there would be appropriate black box testing that shows that with the quality of evidence at issue in a particular case with the same examiner or an examiner of similar level of experience and expertise that those examiners get the "right" answer X percentage of the time, and given that, in these circumstances, I would expect to get the appropriate conclusion x percentage of the time. And I think that as we sit here today, that's about what is the most scientifically valid

statement that you can make. ... In the meantime, if we can develop a fingerprint database, we can do the statistical work that needs to be done and develop some empirical evidence for the uniqueness or the rarity of particular types of prints and we can have some objective thresholds onto what is an appropriate print for analysis at all, then I would feel a lot more comfortable to talk about what kind of numbers that we were going to be offering for probabilistic testimony in latent print analysis (D#1).

I think the move towards probabilistic language for any forensic discipline that doesn't have reliable rarity data is really problematic. ... The other problem with that is that I know fingerprint examiners won't know what source of error is included in that statement and what sources of error are not included in that statement. They will make that statement without either recognizing or admitting to the other sources of error that could adjust that. ... If you gave me the choice between probabilistic reporting or categorical reporting accompanied by accurate statements about the weaknesses and limitations of the forensic method, I would take the latter every time (D#2).

As a litigator, my primary concern is what the jury does with the testimony that they receive. There's a significant concern that jurors, number one, don't really understand probabilistic language and that prosecutors will misuse it. ... The second concern is, I think there is also a dearth of validation demonstrating the validity of probabilistic language being used, either in numerical form or by verbal texts. ... At the end of the day, if there were studies to support that type of language, and if there was some way to ensure that jurors understood what it meant and it was not misstated by either the examiner or by the prosecutor, I think probabilistic language is probably preferable. ... I think probabilistic language better conveys the limitations to a jury, which is obviously essential and necessary. Whether it's reliably and accurately presenting the limitations, I think lies in the validation (D#3).

Judges

Well, I think [probabilistic statements] would be an improvement, but I worry again about two things. First, the ability of judges and juries to really scrutinize, in a meaningful way, when someone says it's this probability or that probability. And secondly, the validity of the underlying statistical methodology used, which varies considerably. ... [Additionally,] most forensic science involves a high degree of subjectivity. I don't think you can easily translate that into statistical probabilities. [What that means is that] there is not only the possibility of human error, but that there will be a considerable range between good practitioners and bad practitioners. So, it's not a question of overall statistical probability. Nevertheless, I think expressing it as a probability would still be better than expressing it as a certainty. But I do think it still has a great potential to confuse (J#1).

Probabilistic [approaches], I think, provides us with an overall framework that [can] cross disciplines and allows us to be able to talk about how we are actually putting a statistical model on something to give it value. At first, when I started working with them, I was like, this is way too confusing and there's no way we're going to be able to do this in a way that's meaningful to people, but in some ways, I think there are some things about it that makes

it more approachable. ... Probabilistic and mathematical statements having to do with the probability that one source is from the other provide a different layer of meaning, which, although far more confusing in many ways, it seems like once you start working with them, [they are] almost easier to define. [For example,] if I'm a juror, and I hear the word match and that word has been defined within this particular community as meaning whatever it's defined as, but I hear it as "they're the same and you know they're the same, like two pairs of socks match." I know what match means [colloquially] so I may not listen to the nuances. ... [That said,] I think that if you're going to use statistical modeling and probabilistics, you should be using numbers. If you want to assign a category, like a word to a category of numbers to make something easier, fine, but I think you have to be able to have some kind of modeling that you can do to be able to get to that answer (J#2).

Other (Academic Scholars)

I strongly believe that [probabilistic reporting] is the appropriate approach to take. I also firmly believe that we must have more effort to improve the data sets that we have available to us to make that probabilistic approach more robust. But even in the interim, where we don't have necessarily the best datasets, it's still better to use a logically robust framework and be open about the lack of data in some cases. I really think that we need to push forward both with the probabilistic approach and in parallel with the generation of data sets to help us with that probabilistic interpretation. ... It is much more scientifically correct and defensible to acknowledge that uncertainty in a probabilistic form (O#1).

I really don't see any way to avoid probabilistic presentations because there's not a scientific justification for categorical determinations. We live in a probabilistic world, so if we're going to be honest about it, the source of our knowledge that we gain as forensic scientists, we have to acknowledge the probabilistic nature of it. ... From my point of view, we basically are at a point where we have to do it and the real discussion should be what's the best way of doing it. I'd say, knowing what's the best way of doing it is an issue on which more research is needed. ... [As for the basis for examiners' conclusions,] I think there are a lot of situations where we just will need to continue relying on examiners' judgment. I do think that examiners' experience gives them some basis for making judgments. There's a whole literature on the accuracy of human judgments and particularly people's abilities to estimate the rarity of frequency of events. [However,] we know that people tend to be overconfident in their ability to do this kind of thing, so we know experts are going to tend to think they can do more than they can do, but that's not the same thing as saying that they can't do it. So, do we allow that person to present their results in court? I've sort of come around on this and said, "okay, well, I think that we probably should allow them to say something about it, but we need to be very careful about what we allow them to say because of this overconfidence problem." ... I would allow experts to give testimony in these areas just because I think it can sometimes be useful and maybe necessary, so I think it's evidence that we want to have in criminal cases because I do think these experts know more than non-experts about lots of things, but I would really hem them in, and it needs to be transparent and the uncertainty needs to be fully acknowledged (O#2).

I have problems with [probabilistic reporting] too, but the problems don't lie on the side of the forensic science community, it lies on the side of the triers of fact. [For example,] I know for a fact, most people don't understand fractions … I'm not sure why a statement of probabilistic interpretation of the data taken at a crime scene actually is better than making a categorical statement. To me, [it doesn't] pay deference to the capacity of the trier of fact to integrate properly [the] information. So, I'm not sure if probabilistic is better, but I know a lot of people are in favor [of it]. I don't think that's a great improvement. … What I love to see is a commitment to study [and] education. ... I'm willing to give [forensic examiners] the benefit of the doubt [as to their expertise] as we perhaps move to a more efficient system, but the thing that I would also say is "let's test it." We should be testing this. I think the expertise is actually there, but I would love to be able to have a scientific test where I could stand up for my scientific colleagues to say, "we know that these people, although you don't understand how they're arriving at these right answers, that test showed that they can do what they say they can." … To me, the most useful thing we could do while we're kind of waiting for the scientific foundations is to simply ask the question, "how accurate is the system performing right now?" It's a measure that will give me some comfort and say, "this is the best we can do in the current situation." … [Ultimately,] if you can tell me how often you make a mistake, then that's something that a trier of fact can get their hands on and say that the scientific evidence tells us you make a mistake one in every 100 times or even one in a million times. That gives me confidence and trust [in] the other things that you're saying (O#3).

Participants' responses related to the benefits and limitations/risks of reporting results categorically and probabilistically:

Laboratory Managers

N/A

Prosecutors

When thinking about probabilistic reporting, especially when numbers were applied in statistics, it goes back to if you have a likelihood ratio, let's say the likelihood ratio considering this evidence is 10,000 times more likely to be this defendant versus some unknown person. Let's say that that's the conclusion. 10,000 means what to me? And what does it mean to you and what does it mean to a juror? And even sliding scales that give you that qualitative statement about the strength of that support? What is that based on? So, I think it gets messier the more you start complicating the conclusions in pattern matching disciplines (P#2).

The benefit for categorical is the certainty of the opinion. And that's one piece of evidence that the jury has. ... If you're not using numbers the way we do in DNA, then I think the weakness is that the descriptors are built on sand. I think there's way less foundation for those kinds of answers than for a categorical answer (P#3).

Defense Attorneys

I don't think either approach is more beneficial, absent accurate statements of the limitations and weaknesses of the method. ... If with a fingerprint probabilistic reporting scheme, large-scale studies showed that examiner agreement with each other a high percentage of the time when selecting probabilistic reporting language, and further showed that the language selected was an accurate statement of the value of the evidence, and further showed the jurors understood all of this accurately, we could then be in a position to discuss benefits. But any discussion of benefits in the absence of that research is nonsense (D#2).

The positive is that [categorical statements] are easy to understand. It's very clear what an examiner is saying, and from the perspective of a litigator and a jury, I would think that categorical statements are much easier for them to understand. But it doesn't really accurately convey the weight of the evidence. ... I think very clearly categorical statements overstate the evidence, and that is always a significant danger. I think that's the biggest problem. ... [On the other hand,] I think probabilistic statements they more accurately convey the weight of the evidence, [but] I think they are very difficult for judges, juries and litigators to understand. … The prosecutor's fallacy was always a real problem for us when we were dealing with DNA litigation and that was, in my view, a simpler concept to convey and the prosecutors consistently got it wrong (D#3).

Judges

N/A

Other (Academic Scholars)

The greatest benefit of categorical reporting is [its] simplicity and ease of understanding, perhaps ease of people thinking they understand anyway. It's a very easy concept for a lay person to say "this expert thinks these two are from the same source." ... [However, the] greatest limitation [of categorical reporting] is that it doesn't acknowledge or is transparent about the fact that it isn't a black and white decision. … Another weakness is that it doesn't allow the expert to give any indication to the court about what value there may be [for] that great big number [of comparisons] in the middle that are inconclusive [e.g., whether those comparisons support exclusion, or support inclusion]. [On the other hand,] when it comes to probabilistic methods, you're not trying to force a continuum into boxes with artificial distinction. You're not trying to make something that is all the way from white to black through gray, into white, black, or "I don't know," and you're able to give more information about more samples with more transparency. … [However,] the greatest disadvantage [of probabilistic reporting] is that we have few well-developed, deployable, validated methods to help practitioners to deploy those sorts of methods. Although I'm a strong believer in the likelihood ratio is the appropriate framework for reporting of evidence, my gut feel is that it is misunderstood widely by lay people and that lay people routinely transpose [the] conditional. So, if I had one biggest fear, it's that (O#1).

[Categorical reporting] is cut and dry—perfectly clear. It's easy for the jury to understand that the expert is telling them "it's him." It's easy to understand. The cognitive demands are low. [However,] from a scientific perspective, these [categorical statements] are problematic claims. So, do we allow experts to say things that are not scientifically justified because they're really, really easy to understand? Well, I don't think so. In fact, if that's your strongest argument, you're on shaky ground. … I think this argument that we should continue making unjustifiable claims because they're easy to understand just becomes laughable. I mean, you're violating the prime directive, which is be a scientist. … [On the other hand, probabilistic reporting] acknowledges the reality that our evidence is probabilistic. If we're going to be honest about the nature of the evidence that forensic scientists can offer, we're inherently going to be in a probabilistic world. … [However, probabilistic reporting is] difficult to understand and interpret properly. It takes people outside of their common experience and comfort zone to be dealing with numbers of this sort and there are known biases and misinterpretations that people are susceptible to. So, it has to be presented with care and, I don't think there's a clear indication in the literature about the best way to do that yet. … [Overall,] I think the standards that we're developing for presentation should require careful acknowledgements of uncertainty and transparency about where the numbers are coming from. So, I think given that we're in this uncertain period where we don't know, we don't have a consensus on the best way to present things, let's be really open, honest, and maybe a little conservative about how we approach it (O#2).

Participants' responses related to findings from a recent survey of forensic friction ridge practitioners indicating 80% of respondents believe probabilistic reporting would be confusing to lay people:

Laboratory Managers

I understand the concerns. It doesn't help anybody if that is not understood, but when you think about what our job is, in the shortest term, [it is to] maximize the value of evidence. If you have this car that is faster, quicker, does things. It could be more dangerous, [but] you still can't not use it. It's a better car, learn to drive it, apply the tools (LM#1).

Watching what I've seen happened with biology, yes, it will be confusing. Is it irrevocably confusing? No. I think everybody in the system can learn how to deal with it and how to explain it. We've been through multiple iterations already of refining how we explain it. You can see the growth curve in the lawyers that are involved. You can see the growth curve in the courts. Obviously, people may serve on one jury in their lifetime. It's not like they're going to come back and have learned something from the last time they were on a jury, but people get better over time about being able to explain it. That window of that confusion narrows down. So yes, I agree, the lay people will be confused. The practitioners are confused by it right now. But that is (1) not a reason to not go there, and (2) not an indelible absolute. The confusion will subside. The confusion will abate and people will get better about explaining it (LM#2).

It probably will be. This is why I don't want it to be only probabilistic reporting. I think the type of testimony that we're currently giving plus this is the best model for the future. I think as it's used more and as we can standardize some things it will be easier because attorneys will know how to ask it and it will be used in a more standard fashion. So, it will be very frustrating, just like likelihood ratios still are right now, with juries (LM#3).

Prosecutors

I think that they should be worried about it to a certain extent. They should be cognizant of whether what they are saying at trial is an accurate description of their opinion (P#1).

There've been cases where prosecutors misunderstood DNA results in a case where there wasn't a match and inappropriately argued those results to juries and those cases were reversed. So, a lot of this is on the part of the scientists and the prosecutor to foundationally understand what the conclusion is, and the limitations of the conclusion are appropriately presented. When both of those things happen, I'm less concerned that a jury is going to misunderstand it (P#2).

Defense Attorneys

[My opinion is] [t]hat 80% of latent fingerprint examiners find it confusing (D#1).

It's clear that people struggle with probabilistic notions. Most of the studies are pretty clear on this. People don't know what's encompassed in a random match probability. People have illogical uses of probabilistic statements as well as verbal expressions of probabilistic statements and sometimes do use those in incongruent ways. So yes, I agree (D#2).

I actually tend to agree with that. Until there is a lot of effort put into: number one, doing the underlying validation, and then number two: figuring out how this is presented and how it's made clear, that we are not saying that X matches Y. That is a real danger. ... I actually do think that the forensic science community does have some obligation for thinking through how information should be accurately reporting. I actually do think it is within their purview because I think that, again, that's something that for years has not been, either intentional or unintentional, but there have been overstatements made in every discipline for years and years and years. So, I think it's important for them to understand that they need to convey information accurately and clearly (D#3).

Judges

Well, I do think there is a potential for confusion, but it's not as bad as the view that the jury will take otherwise, that it's an absolute fact. When the jury hears the opinion it's a match, their natural reaction is to say, "okay, it's been scientifically found that it's a match. Period." When they hear that, "well, there's an error rate of [X]%." Yeah, that may be hard for them to fully digest, but at least it is better than the clearly erroneous view that

they are otherwise presented with that it's a hundred percent and no error or zero error rate and a hundred percent correct. I do think, if I had a magic wand, everyone would take a course in statistics in first year of college because it permeates so much of modern life. But, there's no reason why a good expert, on either side, you can't give the jury the basics of statistics so that they can appreciate what those statistics mean. I also think, even more globally, even jurors or judges without that background know the difference between a 2% error rate and a 20% error rate. And the last thing I say on this subject is, remember, we're not really talking in 97% of the cases about judges or jurors. We're talking about prosecutors, because 97% of all cases plead out so the person who needs to be educated here is the prosecutor so that [the prosecutor] doesn't fall into the kind of errors that a judge or jury might fall [into] because of their lack of education. I don't see any reason, for example, why, as part of the training that the prosecutors go through you couldn't have a half day on statistics. I think that would be a very valuable thing (J#1).

At first, when I started working with them, I was like, this is way too confusing and there's no way we're going to be able to do this in a way that's meaningful to people, but in some ways, I think there are some things about it that makes it more approachable (J#2).

Other (Academic Scholars)

I'm sympathetic with that perspective. I think it is the case. We know that even those of us who've been in this field for many years can, on a bad day, transpose [the] conditional and it seems inconceivable to me that lay people will not transpose conditionals, and that is just one symptom. I think of the fact that it is much harder for a lay person to really fully understand what they're being told in probabilistic terms. But on the other side of that, what they understand when they're being told in categorical terms is an over-simplification to the point of being untrue, to some extent. If someone is really reporting well and making sure that they say this is just my opinion and other people's opinions could vary, then no, it's not untrue. But, yes, I have sympathy with that perspective. I think there is a real difficulty in making sure that we explain these things well enough, but I don't think it is a strong enough reason not to do it (O#1).

I think they are right. It may be confusing to a lot of people, but I don't think that's a sufficient reason to go back to an unjustifiable alternative form of reporting. ... It's not easy to present statistics. We'll need to do it carefully and we need more research on how to do it best. But, from my point of view, we're stuck in a statistical world and we need to make the best of it. I don't think that the evidence supports claims that people are so hopeless that it's a hopeless task. ... I would say, although there is considerable evidence that people can misunderstand statistics, and that is a problem that has to be dealt with, I don't think the errors will systematically favor one party or the other (O#2).

[I agree,] just ask someone on the corner and say, "I have this problem with fractions. I want you to solve it" and see what kind of reaction you get. So that informs me that for the average person who finds themselves on the jury, a deep understanding of probability is it's like asking them to solve Einstein's equations. It's just not going to occur (O#3).

Participants' responses related to findings from a recent survey of forensic friction ridge practitioners indicating 80% of respondents believe probabilistic reporting would be misused by defense attorneys to create "reasonable doubt":

Laboratory Managers

We want to make sure whatever we put out there is not going to be misused by either side so I want to make it very clear and I can understand people being concerned about it. … The last thing I want is to put something out there that can be misused. So, I understand the concern, but then that's up to us to write the reports in such a way that they're clear as much as possible through training that they can't be misused. … That's why you should have the opinion that we believe that this has a likelihood of association, then you throw in the number but you give the whole package as opposed to just reporting a number that potentially could be misinterpreted (LM#1).

I can understand that fear. I don't think it's a realistic fear. I think there will be a period of adjustment just as there has been with biology. There is a period of adjustment that people learning to understand what it means, but I can understand the logic of that when you're going from a circumstance of same things, categorically, without uncertainty to any uncertainty, how is that not reasonable doubt? ... [Whether this should be a factor that practitioners take into account,] I would like to say that people can be rational enough that this shouldn't be something that would be the driving factor, but reality is even people that pride themselves on being rational, really aren't. And that irrational fear of what may happen with a big fundamental shift like that, you can't ignore it. You have to respect the fact that as irrational as it may be, it's where people are at, so you can't exactly tell them they shouldn't worry about that because they're going to worry about it. … I am routinely struck with, even again, in rational laboratorians that seemingly are able to hold the dichotomy in their head of "I'm a rational scientist, I'm going to follow the evidence where it is, but let's go get the bad guy," or "I'm a rational scientist, I follow the evidence where it leads, but I'm an advocate for the downtrodden and those that are wrongfully convicted." Everybody struggles with that. I think there is a huge grade of the concerns that all come back to the fear of the uncertainty and they insert whatever their particular uncertainty is there. Whether it's, "we're going to lose cases that we shouldn't lose," or "I'm going to get beat up on the stand," or "I'm going to lose my job because I can't answer that question certainly enough" or "the juries are going to make the wrong decisions even though it's clear what this should be," their fear is if we change this, I don't know what's going to happen on the other side of it (LM#2).

It can be misused, yes. I don't know what the consequences of that will be. ... Could they be misused? Yes. Does that negate the importance of them? No. They need to be used appropriately. … I think we as forensic scientists, we're always worried about how things are going to be misused. Things are always misused, for the benefit of trial and at least with the attorneys. If the probability studies had no benefit, then yeah, I'd say get rid of them because they're always misused, there's no benefit. That's not the case here. The probability studies are important. We're going to have to navigate through how they're

used so that they're not misused. ... If you're implying that it's going to be misused, so it's bad. That's not what I'm saying at all. Everything can be misused by attorneys, so that's part of our business. Anything that's not complete truth can be contextualized to the point where you're not saying the whole truth. ... I think the staff need to know the limitations and the broad picture and how to explain, respond to certain objections, make sure that the assumptions the attorneys are making are correct versus false assumptions, so that [the] truth can come out. I think it's important for us to know how to deal with the information, but I don't think it means we should exclude it (LM#3).

Prosecutors

I guess that their concern is more like, oh, defense attorneys are going to use this as a way to try to undermine my opinion. ... Like, you're going to use a probability that maybe I don't really think is necessary or isn't really valid for whatever reason and you're going to try to make that seem like my opinion. You're going to try to use it to attack my opinion (P#1).

A defense attorney has an obligation to defend the interests of their clients. So, they can take anything in a case and try to create reasonable doubt. That's their job. So, whether they use a statistic and they get to take advantage of that statistic or the probabilistic reporting and use that as reasonable doubt, so be it. If I decide that this way of reporting is scientifically valid, I'm going to offer that. So, just because in any given case, any type of evidence may be an area of a reasonable doubt, we, as a prosecutor, who's going to offer that evidence needs to evaluate, well, is this truthful? Is it accurate? Is it something I should offer? And there's a lot of layers to the evaluation, whether we're going to use evidence, but for me, it would be improper to say, I don't want to use this type of science because a defense attorney is going to argue reasonable doubt. That's not part of the calculus for a prosecutor. ... The reality is, as a scientist, you should do what good science dictates. As a lawyer, we're going to argue on the law side, it's almost like, you know, scientists, you stay in your lane, you produce good stuff. Lawyers, we're going to argue stuff in court. And as much as you may try to please everybody, that will never happen, unfortunately. So, scientists keep doing good science. Lawyers, we're going to keep arguing (P#2).

I think that that we should be worried about conveying information clearly and cleanly. And so, if anything gets in the way of that then I think that's a problem. ... Shame on the prosecutor if they can't counter what the defense is trying to say. I mean, we see that all the time, you know, a defense attorney will say, well, reasonable doubt, they said they were 99% sure. That 1%, that's a reason for doubt, you know, there's all kinds of locutions that lawyers use to try to create doubt. So, I don't see that as being particularly troubling (P#3).

Defense Attorneys

I think [forensic scientists] should stick to the science and let the lawyers worry about what we're going to say (D#1).

I would call those results laughable if they didn't concern me so much. As we know for decades, forensic examiners have overstated the value of forensic evidence in just about every discipline. I don't remember seeing surveys of examiners concerned about overstatements at that time. … So, the fact that 80% of examiners are fearful of that is not only laughable, but it's also concerning. Why are forensic examiners concerned about the outcome of the case? Why are they concerned at all about what jury determinations are in the case? Why are they concerned about reasonable doubt when that's the very thing that examiners should not be concerned about? ... The fact that 80% of the examiners in a survey are concerned about case outcomes based on shifts of how we report language to me shows the power of the unconscious bias in the criminal justice system. And it's really concerning that examiners are even worried about case outcomes. ... [This] would be shocking if I did not already believe that an overwhelming amount of pro-law enforcement bias exists in the forensic sciences despite the limited efforts of a few to identify and address it. … Forensic examiners should be concerned about accurately stating the meaning of the forensic evidence and possibly whether the jury received the intended meaning (D#2).

I think that probably the opposite is more likely. I think that it's much more likely that the jurors hear a probabilistic statement and they take it to be a categorical statement. I think it is far less likely that somehow defense attorneys would use it to present to the jury an argument that there should be less weight given to the evidence than what the underlying science shows. … I just think that is ludicrous. I think that it's much more likely that the jurors are going to hear a probabilistic statement and take it to mean, you know, that X matches Y. ... I think that it's not for scientists to be opining on how the adversary system is going to understand or misunderstand the evidence. They need to present the science. Their concern that somehow something's being misused by the defense seems to be out of their lane, so to speak (D#3).

<u>Judges</u>

I don't understand that objection at all. If you say it's my opinion this is a match and that's all you said, that's conveying a quality of certainty to something that in fact is not certain. If you, under cross-examination, are asked, "well, what's the error rate," what are you going to say? You're going to say, "I don't know," just probably the usual reaction. Then the jury is deprived of information that is available, that is out there, that if you had required a probabilistic response, the practitioner would have boned up on in advance and could give a response. I'm not sure what is meant by the objection that this might create a reasonable doubt. Well, that's what the system is all about, is finding out whether there is, or is not, a reasonable doubt. It sounds like those respondents didn't have much faith in juries (J#1).

I think we would need to stop being afraid of defense attorneys. I really do think that we just need to stop that nonsense. These numbers can be misused by everybody because they aren't being understood properly. I don't think a lot of it is even intentional. I just think that it is what it is. So, I think misuse happens for all sorts of reasons and it doesn't have to do with what side you're on. So no, I don't think that it should be a reason that we should not look at [probabilistic reporting]. Quite honestly, my experience has been that [many]

defense lawyers are far more interested in actually understanding what the numbers mean and how those things are being generated versus the prosecution that seems to want to sort of just come into court and have it serve to them (J#2).

Other (Academic Scholars)

I think they don't want doubt introduced, [and] it scares me actually. It scares me that forensic scientists don't feel confident to talk through uncertainties and anything that is below a hundred percent. We, as scientists, should be comfortable in talking about the limitations of our analysis as much as the strengths of our analysis. It's the job of defense attorneys to introduce reasonable doubt, but it's our job to be sufficiently transparent to allow them to scrutinize the evidence (O#1).

[First of all,] creating reasonable doubt is what defense lawyers are supposed to be doing. If there's some reasons to doubt the finding, then the jury should know about them. So, the wording of the question kind of amused me—it's the presumption that creating reasonable doubt is a bad thing. ... [Second,] from my perspective, this portrays a mindset, which is that the goal of forensic science is to produce convictions and anything that gets in the way of producing convictions is a bad thing. I just have a totally different perspective on this. … We have to ask what is going to make our legal system operate most effectively. We're talking about optimal operation of a system. Usually, the optimal operation of the system requires getting the ultimate decision-makers the evidence they need to make a fair evaluation (O#2).

I would agree with this. Look, … what you have [in our legal system] is a back and forth between two sides presenting evidence. The point of the exercise is to convince the majority of the triers of fact that my side has done better on the argument than yours. So, if you have a tool in that process of back and forth, that lends more credence to the points that [one side is] making [compared to] the other side, then you're not going to want to give that tool up. The way that forensic science is currently structured, mostly that tool is something that prosecuting attorneys can use. … The "danger" of probabilistic reporting is that now the defense attorneys have this tool of creating doubt. So yes, I would agree with [the practitioners] … that it's likely to create opportunities for defense attorneys to abuse it. I don't disagree about that. [However,] that's why I want to put bounds around all this stuff. The bounds are proof of [the] range of reliability. That's what I keep coming back to (O#3).

Participants' responses related to the role/duties of forensic experts and the limits of their testimony:

Laboratory Managers

Our duty is to make sure that our testimony is framed appropriately—not underweighted, not overweighted. We don't want to have one side or the other to misrepresent, or for that matter, overstate what we're saying. At the same time, we don't want it to be lost in terms of [it] really didn't mean anything. … It's got to be clear [and] it's got to stand alone. …

So, I'm a big fan of you really just can't just give a number runaway. You really should have some verbiage with it so people better understand. ... I get that there's been information that people [have conveyed in the past and they] shouldn't go that strong. ... I believe we're getting direction now to say we really shouldn't use these words anymore—it's overstating. So, I think while that might've flown [in the past] and still might be appropriate for [some situations], it's not appropriate anymore for [others]. You still might have that opinion, but guess what? The times have changed, the data has changed. ... I do believe there needs to be just a little more cohesiveness as an enterprise. ... I think we need to kind of come together a little more as an industry to make sure that we don't overstate or we understate (LM#1).

I think it is an inherent obligation on the part of the expert to convey those limitations and do the best they can trying to explain the inherent uncertainty there. Now the tricky part of that is that's not an easy thing necessarily to explain, even when you have quantitative measurements. ... It's almost more important that it is effectively conveyed on the report [rather than just in testimony in court] because [if] you think about it, our system of justice is not actually an adversarial court hearing. Our system of justice is a system of negotiated plea agreements that, at most, the decisions are made off that report. Approximately 97 to 98% of the stuff never sees the inside of a courtroom. ... [To claim plea agreements are made with full understanding of the limitations of forensic results] is bullshit. ... Most defendants that are dealing with that result have a harried, inexperienced, overwhelmed public defender who has no clue. [Compared to defendants that have the resources to hire competent counsel,] that playground is not level at all, not even remotely close to level. [However,] this is not saying that we have effectively managed to accomplish this, we haven't (LM#2).

I think all of us have an ethical obligation to understand the limitations of what we're saying. That's based on self-auditing, our experience, the papers we've read, how we come to our conclusions, following policies and the reasons why we have policy. So, all of that information that we actually use, I think ethically obligates us to present that information to the jury with a foundational uncertainty. ... [However,] most of the time the court hearings won't allow us [to express those limitations] unless they directly ask us. ... The laboratory isn't going to give you a number, because I don't have a way of showing you how I calculated it. So, articulating that uncertainty is something we're not perfect [doing] yet. But, it's also one of the reasons why we don't say to the exclusion of all others [for example] (LM#3).

<u>Prosecutors</u>

The roles and duties of forensic experts are to test the evidence and follow their rules and the best practices within their discipline and to accurately and impartially convey those opinions (P#1).

A scientist, in my opinion, should give their opinion as to what the science can say. The lawyer argues the value of that opinion and that's foundational in every aspect of a trial. ... So, we, as lawyers, argue value of things, the evidence should speak for itself (P#2).

I think that forensic experts should bring their very best skill and training to whatever the task is and do it without bias. That's what I want. That's what I want from a witness. And that's what I want from my forensic scientist at every phase of the investigation and the trial, because sometimes the information is useful. Sometimes it's exculpatory. Sometimes it's inculpatory. Sometimes it doesn't bring anything to the table. … You should testify about what the finding were in this case and that absolutely should be the limit of what your testimony is because you don't know what came before, what came after, you shouldn't know the " prior odds" in the case. So, I say stay in your lane (P#3).

Defense Attorneys

The role and duty is to not overstate the science based on a subjective belief in it, or what you've been told by a mentor that isn't verified in science. … It's also dangerous for an expert to be offering subjective opinions as to the accuracy of their own opinions. That's layers of opinion. If you're giving an opinion in court, the presumption is that you're confident about it. Whether or not it's in your opinion as to this being the same source, that's for the jury to decide based on the associations that you've made and the appropriate reporting that you've done about your error rate under these same or similar circumstances, then it's up to the jury to decide. What we all know and what the social science research demonstrates is that lay jurors turn off their critical thinking when expert witnesses get on the stand, because jurors are looking for objective evidence to tell the story, they want somebody that doesn't have an ax to grind, doesn't have a stake in the outcome of the proceedings, and wants to do the right thing. They're going to be listening to an expert witness much more carefully and accept it much less critically than you would your typical lay witness whose biases were a little bit more easily exposed. ... We don't have a DNA expert coming in after they give the likelihood ratio and saying, and then in my opinion, it's from the defendant. You're just taking the data that exists that the science supports and the jurors are making their conclusion. … And, your duty is if you make a mistake, or you fear that you made mistakes to go back and correct the record and that duty extends your entire career, your entire life, because when you're dealing in criminal law, you're dealing with life and liberty issues. ... There's an ethical obligation to do that (D#1).

Forensic experts have an ethical as well as a legal duty to accurately state the weaknesses and limitations of their forensic method. But forensic examiners don't take this duty seriously. In my 20+ years of litigating many forensic cases, I have never encountered a forensic examiner who took this duty seriously. The limitations and weaknesses are never documented in written reports. And examiners never admit to them on direct exam. It is always a game of hide and seek for examiners. And this game of hide and seek is exacerbated by the fact that many forensic examiners refuse to discuss the fundamental literature in the field. The weaknesses and limitations of every scientific endeavor is reflected in the peer-reviewed literature. While this was not the case for decades as fingerprint examiners were overstating the probative value of fingerprint evidence, a lot of literature is now available. But too many examiners refuse to acknowledge it. They refuse to discuss the literature in pre-trial interviews. And they refuse to engage in meaningful

discussions of it during cross examination, often enabled by judges who don't understand any of it. Examiners will continue to play this game until they are clearly directed to, one: accurately document the weaknesses and limitations of their method, and two: read, understand, and discuss the fundamental literature in the field. Without both of these, examiners will continue to overstate the probative value of forensic evidence and evade real discussion of the science, and whether a jury gets accurate information will depend on the chance that the defense attorney is unusually prepared and whether the judge grants a proper scope of cross by allowing discussion of the fundamental literature (D#2).

The role and ethical obligations are for [forensic scientists] to, one: clearly and accurately report the information that they intend to present. Number two: they have an obligation to be willing to meet with both sides and explain their findings and explain any limitations of their findings. Then, number three: when they present the evidence in court, they need to be clear about what the limitations are of their findings. ... [When] they're answering the question, they're answering the question fully and accurately. Frankly, I think that examiners set up a little bit of a straw horse where they say, "well, we're not asked that question." I think that most of the information they convey, if they think about what the question is, they could present a more robust answer then they do, and in some cases choose not to (D#3).

Judges

No (J#1).

I struggled with this question because I really do think that an expert who was on the stand really does need to be answering the questions that have been put to them by the lawyer, and we have mechanisms for how it is that we want to expound. If there's an issue that's raised by one side that the expert is not allowed to provide additional information on, the other side has the opportunity to elicit that information. So, there's a court process that is sort of layered over top of what it is that I think an expert can do in being proactive about explaining those things. On the other hand, I think that experts that do explain the basis for their underlying conclusions are far more compelling and better experts. I do think, frankly, the rules of court require that you have a foundation for your opinion. So, from that standpoint, I think that they should be allowed. How it is that an expert can be proactive about it, I was thinking is that some of that proactivity should really be being done at the front end and should be considered in what's being provided as part of the report that's provided to counsel in the case, maybe it's part of the trial prep that goes on between counsel and the expert. I think those are places where experts have a lot more opportunity to be able to work with the lawyers about why it is that it's important for them to explain [and] what it is that they'd like to tell the jury. So, I think that it's just limited by the rules of court and the relationship between the expert and the lawyer, which, whatever lawyer it is that might be working with that particular expert. ... [That said,] if someone is being shut down about testifying about the limitations of a particular testing that was done, that is a place where it is that I think it's fair game for an expert to say, "I'd like to be able to answer your question, but the answer to your question is premised on some information that's also important." I do think that there's some of that that really is appropriate and

it's really hard because I know experts get pushed into this all the time, ... lawyers are imposing language upon them [for example, language such as "reasonable degree of scientific certainty"], and they really feel pressured to respond in a way that they think that is what the listener wants to hear. So those are areas where I really do think it's worth pushing back to some degree on and you know what it is that's being elicited in a courtroom (J#2).

My view is that [would be] called ipse dixit—"it is because I said it is," and, under the Daubert standards, the Supreme Court standard for the admissibility of an expert opinion, that's not allowed. So, there would [also] be a good preclusion motion under a state court standard for admissibility, like the Frye standard that's even more exacting. So, it should not be allowed. Every judge should require that an opinion be backed up by the reasons for the opinion and that, if an expert gets up there and says, "based upon my experience, this is just the way it is," ... I would say that that's an unreliable opinion. It's ipse dixit (J#3).

Other (Academic Scholars)

I think the role of a forensic science expert is to assist the court, not the prosecution or the defense but the court, in its evaluating evidence and to use their skill and knowledge that lay people don't have to help evaluate the scientific findings in a way that is helpful to the court—that is transparent about strengths and limitations. ... If I was to balance what should an expert do in terms of expressing uncertainty, I think they have to err on the side of making sure the court really gets the point that there is uncertainty as opposed to erring on the simplification and ease of understanding side of things. ... I think it is the role of the court to conduct that final reasoning in the light of the uncertainty that exists. I don't think it's the role of the expert to take that uncertainty away from the court if they don't have the scientific basis to do so, [but] if they have the scientific basis to do so, [then] fantastic, fire away. I think we just need to be so careful not to try and be so helpful to the court in helping them to get rid of the uncertainty that they don't like [such] that we stray beyond what we can robustly and scientifically say. It's something that I would say I've observed anecdotally over the years. Forensic scientists want to be terribly helpful, and I think that pushes the community sometimes to give an opinion on something that is too uncertain to give an opinion on in the first place. ... I don't think it comes from any desire to do anything wrong. I think it comes from a desire to be helpful, but I think it's dangerous (O#1).

I think the first duty is to get it right—to say things that are justified scientifically [and] to not go beyond their expertise and not claim more than the science will support. That's duty number one. Do not make unjustifiable claims. Then duty number two is, once you've identified the various claims that might be justifiable, try and choose among them in a way that promotes better understanding for a wider range of people. When in doubt, maybe present the evidence in multiple alternative ways and focus on transparency and a fair characterization of uncertainty. ... [It's] not that opinions of forensic scientists aren't valuable, but we have to acknowledge our own limitations. There's a need for scientific humility. The overwhelming tendency of experts in multiple domains is toward overconfidence. If we're going to be good scientists, we need to combat that by adopting

norms that emphasize when in doubt [to] make the more modest claim rather than a claim that may be too bold (O#2).

I would hope that a forensic expert would limit their testimonies so that it was scientifically defensible. That means avoiding statements that you cannot show having an observationally true basis. ... that are repeatably observable. ... We are implicitly talking about repeatability and reliability (O#3).

Participants' responses related to whether it is acceptable for experts to express their opinion in court without disclosing the underpinnings or statistical data to support those opinions:

Laboratory Managers

I would strongly encourage they do it because I feel it makes their opinion better, stronger. ... [However,] I think there are probably some straightforward circumstances [where it is not necessary] ... [and] you don't have to go into data. Other ones, it's probably not acceptable if you don't give that. ... [That said,] it takes a few seconds more. It's just a fuller testimony (LM#1).

There is my answer to this, and then there is where even we are at, [which] are two different places. No, I don't think that's acceptable, but there is an enormous effort between where the world is at, even on the well-funded, well-intended, pushing the envelope end, and where we need to be. There is still a huge gap there. And the gap extends beyond just laboratories. I can whip my people in having that answer every time and it still wouldn't work because on direct, you have a puppy DA that can't find their way out of a paper bag to ask the questions to allow them to make that answer. And on cross, they get cut off and you've got a judge that's hostile to anybody answering squat. And even when you've got an expert sitting on the stand, getting confronted with an inappropriate question that they are trying to say, "I cannot answer your question as a yes, no answer," the court won't let them do that. You're stuck. That aspect of getting all of that underpinning there is not just a laboratory issue. It is a prosecution, defense, and court issue. And all of those things have to get fixed for that to legitimately and routinely be there. ... And, it happens even more that you've got the puppy DA and the inexperienced, overwhelmed public defender. They aren't even asking the questions. They aren't even giving the opening for you to be able to insert the answer. The court room environment is not allowing for that part to be there. ... This is one of the things that I'm finding myself getting a little bit more worked up about these days, of this issue of it was the laboratory that didn't express the extent and limitations of the testing. No, the lab is willing to do that, the lab wants to do that, all the rest of the system cut it off at the knees (LM#2).

[Not disclosing the underpinnings is] not the best answer, and it would be better to talk about the certainty of that conclusion. We're not always allowed to do that. I don't think it's a wrong expert opinion to give. It's just not the best that could be given (LM#3).

Prosecutors

There are specific rules of evidence that govern expert testimony in any jurisdiction, and they differ jurisdiction to jurisdiction. [In my jurisdiction], technically the expert doesn't even have to discuss the basis of their opinion. But they can be asked about it on cross. Again, this is where I'm going to say, you know, it would be bad practice as an attorney to elicit an opinion from an expert without having them discuss the basis of their opinion. But if you're talking about statistical underpinnings and things like that, there aren't always statistical underpinnings factoring into an expert's opinion. So, a firearms examiner, I don't think necessarily has statistical underpinnings, when he's doing a side-by-side comparison, you know. If you're getting more at, do I think it's appropriate for them to articulate their opinion without getting into things like error rates as a condition precedent of them giving their opinion. Yes. I think it's appropriate. And I think that that's for the legal side to decide (P#1).

The rules of every jurisdiction are going to differ slightly in this way. But a lot of these issues that you're now raising are the specific issues that are dealt with in pretrial admissibility litigation and hearings about whether or not there's a sufficient foundation for the evidence to be offered in court. Once that hurdle is overcome, then I don't think there's a need to then further explain the scientific basis for it and all the research that surrounds it in trial. ... I mean, if you think about it, it could really go down a rabbit hole there. If I was going to do that as a prosecutor, I would then offer every single study that refers to the reliability of the discipline. And we would then spend days reviewing that literature potentially then calling the designers of those studies to further delve into the methodologies and reliability of those conclusions. I feel like you start going into a way well beyond the scope of your trial. And that's one part of it, pretrial admissibility challenges. The other part of it is, the job of the defense attorney is to cross examine the witness. So, if they think that their conclusion is unreliable, then use whatever you think is necessary and appropriate to challenge that conclusion. And that's the essence of challenging a witness through cross examination (P#2).

I think there needs to be some data, but the data could be the two items and the fact that, you know, if you're in ballistics, you can explain how metallurgy works and every land and groove mark is going to be different. And you can explain the consecutively manufactured barrel studies. All of those things are possible. I think any expert should be able to do that. Whether that's necessary in every trial? I would say it's not (P#3).

Defense Attorneys

No opinion should be entered into evidence without a thorough examination for the basis of it. The whole reason that we have a confrontation clause and cross examination is to examine the basis of the opinion. If you're just giving an opinion, then there's nothing you can even cross examine about giving that. Well, I just think that's so based on my training and experience, where do you go from there? It is so, because I say so, right. That's why they call ipse dixit, that's why science rejects that (D#1).

No (D#2).

[Forensic scientists] have an obligation to provide the supporting data, but also the limitations on that data. The forensic scientists that are more steeped in science tend to offer there's a hundred studies, but they don't tell you, for example, that 98 of them are closed set studies and what that means, or that in the two studies where the error rates are 0.01, that the majority of the people who returned the answer answered inconclusive and what that means. ... That's a real limitation on these studies that the examiner has never seemed to bring out, ... [and training and experience] are just not a legally sufficient basis for an opinion. It's something that the court considers, but it's one small factor along with a number of things. ... [It's been admitted in the past because] for years and years and years, the defense bar really was, frankly, not educated and did not do a particularly good job of starting to bring to courts the problems with all of these disciplines. So, there's this whole body of case law that's based on either no litigation or very poor litigation. ... Unfortunately, the courts rely on precedent. So, bad precedent builds on that precedent. I think judges and defense attorneys are starting to be much more educated and are beginning to understand many of the limitations of the forensic disciplines. So, the courts are now starting to limit them to what is scientifically shown or proven are valid. ... The education of defense attorneys has sort of upped the game (D#3).

Judges

No (J#1).

I struggled with this question because I really do think that an expert who was on the stand really does need to be answering the questions that have been put to them by the lawyer, and we have mechanisms for how it is that we want to expound. If there's an issue that's raised by one side that the expert is not allowed to provide additional information on, the other side has the opportunity to elicit that information. So, there's a court process that is sort of layered over top of what it is that I think an expert can do in being proactive about explaining those things. On the other hand, I think that experts that do explain the basis for their underlying conclusions are far more compelling and better experts. I do think, frankly, the rules of court require that you have a foundation for your opinion. So, from that standpoint, I think that they should be allowed. How it is that an expert can be proactive about it, I was thinking is that some of that proactivity should really be being done at the front end and should be considered in what's being provided as part of the report that's provided to counsel in the case, maybe it's part of the trial prep that goes on between counsel and the expert. I think those are places where experts have a lot more opportunity to be able to work with the lawyers about why it is that it's important for them to explain [and] what it is that they'd like to tell the jury. So, I think that it's just limited by the rules of court and the relationship between the expert and the lawyer, which, whatever lawyer it is that might be working with that particular expert. ... [That said,] if someone is being shut down about testifying about the limitations of a particular testing that was done, that is a place where it is that I think it's fair game for an expert to say, "I'd like to be able to answer your question, but the answer to your question is premised on some information that's also important." I do think that there's some of that that really is appropriate and

it's really hard because I know experts get pushed into this all the time, … lawyers are imposing language upon them [for example, language such as "reasonable degree of scientific certainty"], and they really feel pressured to respond in a way that they think that is what the listener wants to hear. So those are areas where I really do think it's worth pushing back to some degree on and you know what it is that's being elicited in a courtroom (J#2).

My view is that [would be] called ipse dixit—"it is because I said it is," and, under the Daubert standards, the Supreme Court standard for the admissibility of an expert opinion, that's not allowed. So, there would [also] be a good preclusion motion under a state court standard for admissibility, like the Frye standard that's even more exacting. So, it should not be allowed. Every judge should require that an opinion be backed up by the reasons for the opinion and that, if an expert gets up there and says, "based upon my experience, this is just the way it is," … I would say that that's an unreliable opinion. It's ipse dixit (J#3).

Other (Academic Scholars)

I think it is really important to disclose the basis of your opinion. I think when it comes to the actual courtroom, [however,] it depends on so many things—what you actually say in testimony. When it comes to your written statement of evidence and your case file, that contains all your notes, [however,] I think that underpinning has got to be disclosed so at least it should be available for scrutiny by whoever in the court process wants to scrutinize it. I think that when we just give unqualified opinions, it is almost impossible to challenge really, because if you're not giving a reason for your opinion then it just comes down to, "well, that's my opinion" (O#1).

No (O#2).

No (O#3).

Participants' responses related to what they would describe as the greatest challenges facing the pattern and impression evidence disciplines as it relates to examination and reporting methods:

Laboratory Managers

Keeping abreast of the technology and how the movement of the data and the philosophy of things are happening in a discipline that used to be very manual that is becoming more and more algorithm and computer assisted. Keeping up with that when you you're still giving opinion evidence is a real challenge because you have a mindset [shaped by] what you learned, and whatever you learn, [to you] it becomes right. … The challenge for that group of individuals to keep up, to feel they're part of it, to stay on top of it, especially when you may have some people that are not as savvy in some of these things (LM#1).

The trivial answer is just money. I say that's trivial because there's really a lot of nuances under that. Really it is a matter of laboratories, writ large, the entire system is so wildly under resourced. ... Then the expectation of what people think is occurring, just doesn't match the reality. I keep finding myself in many circumstances saying, "yes, there are very real issues of science that we can't lose sight of, [such as] how many points of minutia make up a sufficient circumstance for identification? [or] quality algorithms to be used on [fingerprints]." That's great, but when we're dealing with simply the evidence coming in the door, being fundamentally flawed, those matters of science, don't matter. ... Most of the stuff coming through the door is illegible, mixed up, damaged, contaminated, and really inappropriate to use. So, yes, there are real science things that we can't miss, [but the] biggest challenge is you've got shit coming in the door—of course, the answer coming out the other end is going to be shit. The biggest challenges are the enormous amount of effort and patience and capital that's sucked away in, at best, modestly competent information systems that the laboratories run on. Basically, you've got analysts working in near third world conditions. Those are really the biggest challenges. Yes, the answer of money is trivial, but until we solve some of these things that affect every case for the entire system, all the way through, honestly, the science-y stuff is a little a bit of a privilege to think about (LM#2).

I think as a group we need to integrate and develop probability-based studies into our work quicker. We are moving too slow in this story, and I think some of this is architectural and practical. ... We need to get these models working in the laboratory side-by-side with the expert witnesses, that's the way we're going to be able to give the best information to the jury—by having both the expert witness and hard objective models. ... [But,] these algorithms aren't easy. There's a separation between academically available data and a crime lab available data. Labs and academics, research groups need to work closer together. There is still a little bit of resistance that you're taking away the expertise [the experts] already have and supplanting it with something else. That, to me, I think is completely false if you agree to integrate them both together. If you want to completely replace an expert, then I'm going to be opposing you because I don't think it's appropriate either. I don't have a problem with the numbers. I just don't think the numbers themselves are the best model. ... The other biggest reason is that [for] crime labs, it's not our mission to do research, unfortunately. I love research and it's wonderful, but we are under so much pressure to get casework done. We just don't have the time, energy or money to do it. It's unfortunate because we're really the best place to do it, but we just don't have the money to do it (LM#3).

<u>Prosecutors</u>

Lawyers. I mean, there's really no other way to say it. So, I'll give the caveat just because I feel I need to give it like, yes, I am a prosecutor, but I want accurate and scientifically sound forensic evidence. I don't want opinions that are inflated. I want you guys to decide what is accurate and what the limits within your discipline are. And then that's my evidence and I deal with it in my case. I think probably the biggest challenge that the pattern disciplines are facing is what I'm going to call sideways attacks. I think that your disciplines, sometimes under the guise of cooperation, are being undermined and

encouraged to render yourself obsolete. I guess this would be the best way to say it. I really do think that lawyers right now are your biggest problem, lawyers and academics. You need to let us into a certain extent, but you got to kick us out too. And that goes for prosecutors too. Like you got to let us in to hear what we have to say, because it does help you. And, I know it looks great to play nice in the sandbox, but there is value to hearing what we have to say, but you have to know when to stop listening, and to realize that not everybody has a crystal-clear agenda (P#1).

I think it's a bigger issue that's happening in the community, is to understand what the conclusions are and what the limitations are, and to ensure that we're staying within those boundaries (P#2).

I think the challenge is that practitioners and people like you are attempting to appease the defense bar and that's never going to happen. ... You are never going to satisfy the defense bar because we are in an adversarial system. You're never going to have the defense bar saying, you know what, we're satisfied. You have done a great job. Because it's part of the adversarial system, but they are trying to dilute the impact of forensic evidence that implicates their clients while at the same time, and nobody ever calls anybody out on this, if the forensic evidence supports their theory of the case or tends to exculpate their client, then it's the gold standard of whatever the discipline is. So, I think that's the biggest challenge. It is this feeling among the disciplines that they're going to be irrelevant if they don't agree with the defense bar, because the defense bar has many seats at the table. You know, I look at the composition of the OSACs and I look at the composition of various committees, and I see as many defense attorneys as I do see scientists and prosecutors combined. So, I think that the challenge is trying not to fold in the face of that kind of pressure (P#3).

Defense Attorneys

This digging in on the way that this has always been done because of subjective belief that there were no problems with it or because there haven't been tons of wrongful convictions associated with it, is sticking your head in the sand. We know what's lacking in these techniques, read the PCAST report, read the NAS report, read the AAA's report fingerprints. These are the top scientists in the country. Just because forensic folks disagree with them doesn't mean that the forensic folks are right. ... I should probably withdraw that . ... Overstating conclusions, you know, is a fast way for wrongful conviction and for us to have to go back and examine tens of thousands of cases where this has been done. The challenge is that courts will . . . . [well, . . .] I don't know, you know, actually, the truth is there may be no challenge, courts just may not care, because we don't care about the rights of the indigent defendants. In your typical criminal cases, the challenge is scientific integrity. The challenge is trying to claim science when you don't have any (D#1).

In pattern matching, I would say it probably continues to be the lack of empirical research. So, take fingerprints, we know so much more about it now than we did a decade ago in terms of all these empirical studies and they have been so important to me, at least in understanding the limitations of the method, and to their credit, the fingerprint discipline

is way ahead of other pattern matching fields in that regard. So, all of those big studies, many of them have not been repeated in any way or in as good a way as fingerprints, and yet, even in the fingerprint discipline, there's more to do (D#2).

To do the research that's necessary and have it done by people who are independent of the discipline, who don't have any interest in the admissibility of it. Let's see what the research shows us, and then let's learn how to present it accurately in the courtroom. I can tell you that there will be far fewer admissibility challenges and far less litigation if that's done. ... I think it's in reach. It seems like there is funding now for this research, there is energy behind it, and there should be the incentive to do it (D#3).

Judges

Good, blind, scientific testing. That's not my conclusion. That was the conclusion of the National Academy of Science. So much of this has just been "seat of the pants." Most of the forensic sciences are developed by police as investigative tools and, for an investigator tool, it doesn't matter whether it's subjective or not. If it gives you a helpful lead that you can trace out and see whether it pans out or not, great. But then, beginning in the early 1900's, with fingerprinting, it began to be introduced as hard evidence in court and people forgot that it had never been subject, with the great exception of DNA, to serious testing. But it's not as if it couldn't be tested. So, I think that's the greatest failing (J#1).

Other (Academic Scholars)

I think that the sort of ongoing narrative of forensic science in crisis can be really unhelpful to these disciplines because there is a huge amount that we can confidently say in these disciplines. If we are honest about our limitations, then [forensic science] can still be of real assistance to the courts. I think one of the challenges, really, is this ongoing "until it's perfect." It's all awful narrative, which I think is just unhelpful. I think if there was more of an acceptance of imperfection, with clarity and disclosure of that imperfection, then we would be able to move ahead in a more step-wise fashion and just keep improving rather than [what seems to be] a desire for a jump from, as I say, "terrible" to "perfect." That's just never going to happen. We just have to keep pushing at this from every direction (O#1).

If the field is going to continue with relying on the human brain and a human assessment of similarity as the major instrument for making assessments, then it's very important that we do assessments of the accuracy of that instrument. So, one challenge is validation of the accuracy and performance characteristics of human judgment. If we're going to do an algorithmic approach, then obviously we're going to have to validate that as well. So, I'd say regardless of how the examination is done, the greatest challenge for the examination is validation. Then, on the reporting side, the challenge is how to present the findings in a way that takes into account both the strengths and limitations of the performance of the method as revealed by the validation studies, and if we're in an area where validation has not been done or is incomplete, how to acknowledge that in a forthright manner. ... I [also] think it would be a mistake to assume you can do a single black box study and [assume] you're done. The "one and done" approach that I've heard a lot of people take is clearly

not realistic. … Knowing what the limits of the human instrument are in terms of accuracy is really important for the overall operation of the system. From my point of view, validation and performance testing should be a continuing part of the job. It should be incorporated into lab work. If we incorporated that kind of routine empirical testing into the way we do casework, it would make us more of a scientific discipline (O#2).

The greatest challenge that I've observed is actually resources. … I have had a chance to see the conditions that real forensic scientists work under. They're not the conditions that Hollywood tells the public about. The real conditions are often overworked people [and] under-resourced people with no time to get the results out. I mean, that's the real world. To me, that's the greatest challenge to forensic science, to convince our society to put in the resources so that people can do the best job, so that this intuitive expertise that I [believe forensic scientists have], is actually allowed to work without having the pressure that can induce errors (O#3).

16.6 Appendix E-6

PARTICIPANTS' ELABORATED RESPONSES RELATED TO THE USE OF ALGORITHMS

Participants' responses related to the use of algorithms in court and the benefits and risks/limitations of them:

Laboratory Managers

I think that's an excellent thing to assist in better understanding why you came up with this opinion. But the danger is that people then rely too much on the number and it's really for the expert to frame that, to help you with understanding how much weight there is on what I'm saying. The algorithm is not testifying today. I am, and this is what it means. I think [an algorithm] just really is going to help them. If framed and used correctly, recognizing that much of the time we're not testifying, it's a report that has to fly and we just don't want to throw up the number without all the basis for it, the strengths, the limitations, just so people have a crystal-clear understanding because if they're going to base their decision, plea bargain, [or] whatever, it needs to be as clear as possible for them—on both sides (LM#1).

Yes. I think ultimately they have a very real and very large role. I think the greatest benefit on the algorithms is the relative consistency of the result case over case. The ability to engineer the system for accommodating the fact that these things [our heads] are biased engines. That's not bad thing, [but] there's a lot of advantage in the compliment of that bias engine with a [algorithm] that's going to do the same thing every time, [it] could get you much closer to reliable results. … There's going to be less variance because the person's kid is having trouble in school, or they're not feeling well, or, they're grossly underpaid, or it's cold, or they haven't slept. It's going to even some of that stuff out. ...

[Further,] they are a force multiplier for analysts. No way, is anybody ever going to cough up enough resources for there to be actually enough analysts to do what everybody thinks is getting done. So, somehow algorithmic compute-based tools are going to have to amplify what analysts we do have. So, I think there is [also] an enormous role in simply from [a perspective of] building capacity. ... [However,] I think the biggest risk is becoming overly reliant and we just exchange the categorical certain answer from the spectacle nerd for now, an infallible algorithm. ... The pitfalls are the desire to want it to replace the analysts—the desire to view it as a cost savings thing that lets me get away from having all these analysts. ... And, I would hope that a lot of these facial recognition things are a cautionary tale for everyone that depending on what those algorithms are originally trained on, builds biases into the algorithm, and those are biases that you can't wash them back out (LM#2).

I agree with using algorithms. I think it's something we need to do and should do. I do not agree with that being the only thing we do. I want to use algorithms and I want the expert with their experience and so forth—both hand-in-hand going to court. ... The benefit of using the algorithms is [that] it's a little bit more standardized methodology across the industry so the rules are a little bit clearer for everybody on what to use [for examination and interpretation purposes]. ... The disadvantage is that we stop there and don't use the other expertise and training and methodologies that the person has, which is huge, as human beings are so good at looking at things without breaking down to individual pieces, looking at things holistically and saying there's something wrong or something different. That skill, while it may not be objective and something you can attach a number to in a computer, is very, very valuable. So, the negatives here are doing one without the other. They both have their strengths and they both have their weaknesses—[particularly] when they're done exclusive to the other (LM#3).

<u>Prosecutors</u>

For the pattern disciplines, based on what I know at this point, which I know very little about what's available or what the underpinnings of them would be. I don't think that it's necessary. I think it would overly complicate things and I would not be in favor of it at this point. Again, my mind is open to be changed as if I was presented with evidence that said, okay, here's how we can compute how rare this particular fingerprint impression is. I would be open to hearing that. ... Math is hard. People don't like math. ... I think that it can be confusing and, again, if there's an accurate way to report something that's less confusing, that would always be my preference across the board. But I do think with DNA, it is necessary (P#1).

Algorithms can add value to the case in so far as giving weight to whatever the conclusion is. So, in that way, it's helpful. ... The more data that exists that underlies whatever the algorithm is, I think the more likely it is that there's going to be buy-in from all stakeholders. ... As long as it's based on appropriate data, I have no problem with it. ... [But] the risks go back to whether there's a sufficient body of data that supports whatever the conclusion is. ... And in context of the case, does it depend on the robustness of the database which is being. Let's say it's fingerprints, are we looking at the likelihood ratio

in relation to [my entire city], just [my] county, [my entire] state, country-wide, worldwide? What is the database? And then, how strong is that conclusion and how appropriate is that likelihood ratio based on the evidence (P#2).

[Algorithms] allow the scientists to do computations in seconds that would be undoable in a human timeframe, and so it gives you way more information and helps you weigh the evidence. I think more information is always good, so, yeah, I'm totally in favor of that … if it were valid. I think it's working very well with the DNA [but] I do not see how we establish the numbers or the levels of confidence in pattern matching, because quite frankly, unless you are completely confident, it's an exclusion, it's not a match. … Again, I don't see how we come up with one that's valid [but if we did] the challenges I think are explaining it to a jury. … [Ultimately,] anything that increases the accuracy of the forensic science, I think is useful. The bottom line is if you, as a forensic scientist, have confidence, then I have confidence in the result. So, whatever makes you more confident, I think works for me and works for the court system (P#3).

Defense Attorneys

As long as we have transparency. You can't go into criminal court and just say, the black box told me this, and so therefore it's so. I spent a lot of time litigating around that issue. I'm for progress, but not by proprietary companies seeking to make money and not being transparent about the data and allowing opposing experts access to that data to examine the basis for these opinions. ... The greatest benefit would be is that you move away from unsupportable categorical claims into something that has some empirical basis to it and that you would actually have a number that's based on a valid statistical database, a population frequency database that is transparent and known. That's progress. That's something where there's real scientific efforts to be more accurate, more precise about what it is you're saying when you're "matching" a latent to a person. [But,] I'm never not going to be concerned about proprietary software being used in these circumstances. ... [Overall,] I think that there's a place for [algorithms]. I don't want to say anything that would retard progress, but there are some things that look like progress and sound science-y that aren't, and I think it's particularly dangerous to introduce new technology or new ways of expressing conclusions that are no more grounded in empirical data than they were previously. They may sound more modest, but in many ways they're even worse, because now we've got some algorithm, now we have a machine, and so therefore it's better, it's smarter. We, as human beings, the faith that we place in technology is [significant]. We don't need any more evidence than you and I are talking right now—the computer said so (D#1).

I don't think the use of algorithms is inherently good or bad in forensic science … there are some obvious and undeniable benefits [such as] speed and expense. Algorithms can handle things much faster than humans and on a much bigger scale … but, I think one thing that's fairly inherent in the criminal justice system and the use of the algorithms is that they are often used before we know how good they are, before we know the strengths and weaknesses, before we know how to judge whether it's operating well in this case or any case … [and] police, prosecutors, and judges accept the evidence because it is

computer based, believing that because computer code is involved, it must work. ... Experts who testify about algorithms have no idea about the human imposed parameters of those algorithms, so they can't even begin to explain any decisions that went into how that algorithm operates and any limitations or weaknesses and how those limitations or weaknesses might impact this case and this testimony and this evidence. ... In my experience, it's way more difficult to figure out when these systems fail in the criminal justice system than elsewhere. Ground truth is so murky. It can be really hard to figure out when these systems work and don't work in the criminal justice system. [That said,] I think algorithms should have a role [in forensic science]. I think, when algorithms replicate the ability of human examiners in their interpretation, I'm much more comfortable with that use of an algorithm. And if I'm comfortable that proper validation has been done, that there has been meaningful oversight of that validation by people not impacted by its implementation and that examiners give proper caveats about the outputs of those algorithms, then I say go for it. ... [However, I am concerned that] inevitably they will be used in the criminal justice system in a role that far exceeds what I'm calling for (D#2).

That's a complicated question. I think in the long run it is something that will improve forensics. In the short term, I think the problem is there needs to be significantly greater transparency. The fact that there's objectivity is what is great about algorithms ... [but,] the limitations of algorithms are all the assumptions that go into the creation of the algorithms and the ability of the person presenting the results, as well as the end users understanding how the algorithms work, and what the limitations are of the information that's being presented. I think that's challenging. ... There needs to be scientists from a variety of different backgrounds involved, weighing in, and there also needs to be some significant oversight. ... If they're used, there needs to be some type of accreditation and enforcement. ... [Overall,] I think there's a role [for algorithms], but I think it's just really important for everybody to understand globally that the machines are only as good as the information that they consider, and there are real limitations. I think we are just moving way too fast and we need to take a pause and really start to understand and accept that machines have all of these same limitations that humans have and they are not the answer to all of world's problems (D#3).

<u>Judges</u>

I think algorithms can be helpful, to a degree, if they are totally transparent. When you say, "oh, here's an algorithmic formula," you give the impression that, "oh, this is something like calculus." But, in fact, it involves all sorts of choices on the part of the person who puts the program together. So, that has to be fully transparent and open to peer review so that you know whether it's a good algorithm, a bad algorithm, an algorithm that has a high error rate, an algorithm that has a low error rate, and so forth. ... I think really good algorithms could reduce the subjective portion of the analysis. ... [However,] some companies are obscuring inquiry through trade secrecy laws, but even where that doesn't operate it's very hard for even defense counsel [to review]. ... Even in those states where the trade secrecy law objection is overruled, they have to hire an expert. You can't expect a defense counsel to be an expert in these algorithms. In many states, there's no

money available to hire that kind of expert. Many states, even where you can hire an expert on the other side, the expert gets very limited disclosure (J#1).

I don't necessarily have a problem with using algorithms. I think there are likely reasons that [they are] beneficial in doing some analyses that we just aren't capable of doing without the use of computers. For me, the biggest thing is transparency. I think if you're going to utilize algorithms that the algorithms that are being utilized need to be transparent as far as access to their source code [and] access to the assumptions that are being placed into that algorithm, making sure that there's equal access, for everyone, to be able to either utilize the software in whatever way it is that they want to utilize it for answering their own questions or at a minimum that there's access for purposes of doing appropriate research. ... The best way I can say is transparency. For me, that's what it's mostly about. ... If they're being utilized with assumptions baked into them, that aren't known, I think there is risk for the potential for misuse. ... [Further,] if you have something that you're not transparent with, [then] the assumption is you're hiding something. This is very big in the culture, I think, in criminal law in particular, and a lot with defense lawyers [who] assume that if they aren't given access something, it's because you're hiding something. I think the community is getting more and more to the point where if you aren't being transparent with something, they think you're hiding something. So, from [that] standpoint, if we are utilizing algorithms that we don't understand and that we haven't provided enough transparency around how it is that they're actually doing what it is that they're doing, I think that we erode the confidence in the analysis as well as potentially in the system itself, and that's where I think it becomes really concerning (J#2).

I think that algorithms are here to stay. ... There's a great potential [with algorithms], [if] done correctly, to create criminal justice reform to a degree that we've never seen before. There is extraordinary potential for that, because there will be an ability, if these tools are designed correctly and they're validated correctly, and they have the right degree of trustworthiness, including [this concept of] fairness, they have an ability to take out some of the human biases that have plagued the criminal justice system. So, I think there's great potential, ... but there are certain risks. ... What we need is a national conversation on what that means and how to create trustworthy and reliable algorithms that can be used for individual liberty determinations. That's where the rubber meets the road. ... The greatest risk is that we allow complex design and complex tools to just snow us a little bit ... [and] that we don't have these conversations as to what fairness means and what fair design is and what trustworthiness is in time (J#3).

Other (Academic Scholars)

It depends on the algorithm. ... [In general, the benefit] is performance [and the ability] to program algorithms to do things that humans can't do. ... I think that using computational algorithms that the reporting scientist understands the basis of and is able to explain is a really good thing. Using, let's say at the other end of the spectrum, algorithms based on machine learning, which have come from, let's say a manufacturer who won't disclose the training set, and that the reporting scientist doesn't understand [or] can't explain and come talk about any biases, for example, or any limitations [of it], is

deeply problematic. ... [Further,] I'm not convinced that there is a legal basis on which to introduce that evidence, because who is the expert is the question that I then come to, is it the algorithm? It can't be the algorithm because you can't cross examine an algorithm, or is it the expert who's giving the results of the algorithm in which case that expert has to be an expert [of the algorithm] and has to understand what they're talking about. So, algorithms, yes. Algorithms with a lack of understanding, even by the manufacturer [where] nobody knows what they're really doing and the basis of their decision-making, I think is really problematic. ... I think that would need a whole set of legal safeguards around it that is different from the legal safeguards that already exist around expert evidence. So, I'm not saying it could never be done, but I'm saying that I think legal scholars need to think very carefully about the safeguards that would need to be in place for machine learning algorithms to be accepted on their own, without explanation as evidence. ... [In those types of algorithms,] I think it'd be quite difficult to make sure that the validation was sufficiently comprehensive [such] that you never questioned the output because that's what it would be coming to, you don't question the output because there's no one who can answer the question. So, the extent of validation that you would need to do that you would never need to question the output would be phenomenal, I suspect (O#1).

I certainly think it's possible to use algorithms for court purposes. ... It all depends upon the success of the algorithm and whether it's been validated and is appropriate. I think algorithms may well be preferable to human examiners giving opinions based upon experience because the use of the algorithm reduces the chances for bias and it may allow better estimation and calibration of that strength of the evidence. ... [However,] these models tend to be very complicated and difficult to assess. Algorithms have advantages, but it's going to require a whole new realm of expertise to evaluate them. ... One area where I'm a little bit worried is [whether] the practitioners have enough expertise to be able to assess whether it's working properly in a given instance, in other words, case specific evaluation of the appropriateness of the algorithm. I think it's important to have somebody knowledgeable look at it and say, you know, is the machine in doing this analysis making assumptions that seem plausible in light of what we know about the data? Or is the machine going off the rails? ... Has it done something inappropriate? ... There is a risk that practitioners will use the models without fully understanding them… (O#2).

It depends on what the algorithms are applied to. ... The greatest benefit [of algorithms] is actually to let humans do what they do best. ... The way that I think of the use of algorithms in forensic science is [that] there are things that I think can be done more efficiently by algorithms, and then that frees up the human expertise to deal with the more difficult things. That's the kind of deployment that I'd like to see happen. So, that's the greatest benefit because that gives the people who have the expertise more time. ... To me, the application of algorithmic techniques that are not tied to demographic factors, I'm in favor of that. DNA is an example of that. [However,] algorithms that are tied to large characterizations of populations—that's where I think it becomes dangerous. ... I'm terribly worried [about that], and we already see this. There are algorithms that police departments use, predictive algorithms, about crime, for example, and there have been a number of studies that show that, as one might expect, African-American communities don't fare well for these algorithms. ... So, the biggest danger [is] that [people] will use

algorithms inappropriately where bias can come into their views without even knowing (O#3).

Participants' responses related to how algorithms can be trusted for use in court, including issues concerning the disclosure of source code:

Laboratory Managers

That goes to validation and I'm a super big believer in validation. Everything we do should be tested [using] mock samples within the range and scope of anything you're actually going to apply it to, … [and validation] should be well within the understanding of the expert to be able to answer those questions and [explain] that it's fit for purpose before we used it. I understand the concerns [of trust], but that just means we've got to do our job in showing these tools are valid before we actually apply them to the case. … I do believe that having appropriate validation data and showing that you don't have to see in the black box to see that it's reliable. … I think largely revealing source codes is just a tactic. Nobody's going to take that source code and go "Aha!" … I'm not going to say it's useless, but to spend the kind of money to make it of any use, in practical terms [it's] virtually never going to happen. That side [requesting source-code] is just going to ask for something, to eventually get the answer of "no," so, they in turn, they have something go "well, if only we had that thing, then we would have been able to show something." … That said, even though I respect this strategy, if we're being a hundred percent transparent, [if] you want the thing, knock yourself out, here [it is]. ... But I do respect that that is competitive [so] bind it to that particular case only and throw on some significant penalties if it's leaked out, because that is the person's livelihood. ... Frankly, as I said, it's a tactic, and if you give it to them, it's not going to be of much use. ... It's a waste of time, but you know what, knock yourself out, here it is as long as it's protected. … [Ultimately,] my goal is to maximize the value of evidence. … If somebody isn't willing to turn [the source code] over and there's a percentage of the time that [the evidence] is compromised in court, I'm going to have to [take that into consideration] … I'm duty bound to pick the one that's a better product, [and source code disclosure] is a feature. … I don't have to agree with it or disagree with it. We know what happens, so I'm going to have to choose what's best for our cases, which is the one that gives more value, which is going to be the one that is okay with [disclosure]. I think [the vendors] just have to get past those legal hurdles and [realize] it's just part of the reality of the environment in which they're doing business (LM#1).

[The issue of trust comes down to] how you set up your framework as part of validation and before you start assessing an actual case of what valid data means and looks like, and essentially, how you put into the datasets, the traceable control. What does control mean? ... You need to think through what is basically both that positive and negative control that can be put in as an internal standard within the data collected as part of validation. But then not only is this part of [initial] validation, but it is routinely inserted into all cases so there is an internal standard in everything you do. These algorithms [can] change over time, they [can] learn, they are not [necessarily] static. ... What that internal standard is for a latent print, I don't know, but there needs to be something that's in there that is an

assessment of [whether] the algorithm is behaving as I expect on that particular application. … [That said,] by and large the source code is specious—it's a red herring. These pieces of software are fantastically complex. ... Who else on the planet is going to be able to actually assess looking at the raw source code what anything means there? It's nonsensical. What is sensical is the internal validation of that result. That's the part that should be there. ... It's the standard that is put into those data so that every case, every instance can be self-validated (LM#2).

I think the more open the models and peer discussions are about how these things are done [will improve trust] … so [that] you're actually able to do peer review and testing, and people could talk about the limitations and benefits. As long as that's done, I think we can advance a lot quicker and to everybody's benefit. But when it's done in a proprietary fashion ... you [have to] feed a lot of unknown "black" data into the system and you get a result at the end. I can look at the results and see if it's good or bad, ... but I can't get a fully good understanding of what's under the hood. ... There's intellectual property, I get that, but the more open we are, the better [understanding] we're going to have about limitations. ... The problem with validation is I don't have a perfect world [and] validation is subject to some limitations based on what I fed it. ... In a perfect world, what I would like is for the analyst to know the algorithms that are used so manually I could pick apart an algorithm. ... All that guesswork is gone. So now a computer can do all the grunt work for me, and I can actually do it manually and say, this is why I'm attaching a number. That's what I would like to see in a perfect world. It doesn't mean the validations are not important. They are, but they are only black box validations. I don't know what's in the box. … [That said,] I'm a big proponent of intellectual property, but that's not necessarily for courtroom use. … [In] the perfect world, if you're dealing with people's lives in the courtroom, knowing everything about how decisions are made is a better approach. So, where do we go from there? We, in [our jurisdiction], have chosen, in DNA mixture interpretations, we've chosen software where they disclosed their algorithms, and we did that for a reason. I think the best way to do this is under protective order, so for purposes of that case, you disclose it, but it's not open to the general public. ... [However,] whether it's absolutely necessary [to disclose source code], I'm not at the point right now that I would say it is. I think you can validate [the algorithm] copiously to the point where you can get reasonably good inferences about its efficacy. It's limited and it's not perfect, but I think it's still usable (LM#3).

Prosecutors

[Trust is] a valid concern and maybe another reason to just not go there unless we need it, which I don't think we do, at least based on what I've seen so far. ... I don't think [source code] is something that necessarily should have to be disclosed in the first instance. However, if the defense wants it, then I think that they should have access to it. I think that steps can be taken to protect any proprietary interests in the source code. … I think that when you create software that is going to be used as part of the criminal justice system you have to realize that constitutional rights are going to come into play. But I also don't think it's something that automatically has to be disclosed, at least under the rules in my

jurisdiction, and I'm guessing most. But if requested and if a good cause is shown, then it's something that could be disclosed (P#1).

[Trust] just goes back to whether there's a sufficient body of data that supports whatever the conclusion is. That's really my only question. Again, if it's scientifically valid and the scientific community is saying this is good science, then as a prosecutor, I'm behind it. Honestly, my opinion is who am I as a prosecutor to stand in the way of scientists saying this is legitimate science, you know, and we agree by and large that this is what should be offered. To me, what drives my decisions here is what is legitimate science and what are the scientists saying? Not as much of what are the lawyers saying about it? What are the scientists saying about it (P#2)?

We have discovery and defense experts in cross examination to settle those questions and explore those questions, and I'm all in favor of giving the defense every tool that they need to investigate the algorithm. ... I trust it because I understand the process of validation. I understand developmental validation and I understand the validation that the lab does to test the limits of the software or the technology in their lab. ... So, if my lab has a great deal of confidence in it, then I have a great deal of confidence in it (P#3).

Defense Attorneys

Transparency is number one. ... The source code has to be turned over to an independent software engineer for the defense to examine and to test on the evidence at issue, you have to have full access. ... There's no counterbalance at all. Trade secrets is absurd. It's absurd that we're even having this conversation as relates to criminal justice period, full stop. ... There is no other, including commercial litigation, where [source-code] wouldn't be turned over and examined. It's good enough for Apple versus Sony and it's not good against the people v. Smith? That concern will never go away. ... They should not get in the business if they don't want to turn it over. This is not Kmart (D#1).

Trust is hard, but really robust validation is a part of that; however, it's not the full answer because what we know is that, for instance in DNA, there are so many variables that can affect the reliability of an outcome that we never test for all of them. We never know the combination of those that might, at some point, affect reliability. Validation can be a big part of [trust], but validation is never without holes. ... We [also] know there've been studies that show that criminal courts are very bad gatekeepers of forensic evidence in a way that they're not in civil court, that they failed time and time again to assess forensic evidence in criminal cases with the same kind of eye that they do in civil cases. [Further,] we know that in civil cases judges have never really disallowed one side in to get access to source code ... and trade secrets is a non-issue because usually they get it in the context of a protective order. It's never been questioned [in civil litigation], but for many years it has been questioned in the criminal justice system because of all the biases that are part of defending somebody who's charged with a violent and maybe vile act. Judges treat criminal cases differently than civil cases, [there's] just really no doubt about that, and because of that, defendants in criminal cases for many, many years have been denied access to source code. Although more recently that trend has turned. ... What would I need

to be comfortable with widespread use and acceptance of an algorithm in the criminal justice system? First, I would need source code. ... Developers should not work in any forensic space where the results of their algorithm operation are intended as evidence unless they are willing to publicly disclose their code. ... Second, I would need some kind of oversight board—a team of neutral academic experts—provided with the time and resources to analyze the code, stress test it, and publish understandable reports about the assumptions underlying the code, the limits of operation based on stress testing, recommendations for improvement, and recommendations for testimony caveats based on their work. I wouldn't accept that work with open arms from either the developer or from the forensic science community in general. ... I think the forensic science community has proven time and time again that they are incapable of describing the caveats that should accompany forensic opinions. Third, a pilot period of years, during which a limited deployment in casework is constantly reviewed by the neutral academic team to make sure that the system is being used as intended and that experts do not misstate the value of the evidence in court (D#2).

The most important thing is transparency. The algorithms and the software have to be made available and they have to be able to be used by experts from both sides. They should be tested to see what the limits are, to push them to their limits to see when they fail. I think that's the whole premise of testing algorithms—you try to make them fail. That's the whole premise behind you doing any kind of validation study at all. So I just think availability [and] transparency is probably the most important issue. ... [Further,] I think giving access to the source code into the software to experts who are working for both sides is important, and giving them the time they need [to have] the ability to see when the software works and when it doesn't work. That is really, really important. ... If prosecutors are going to offer this service, then they should be prepared to turn over the discovery, and the discovery that I'm talking about in this context is the access to source code and the software, as well as all validation information and et cetera. ... Source codes are turned over every day in civil litigation with protective orders. ... The measures that criminal judges have taken to prevent defense attorneys from getting access to the source code are not seen anywhere else. There is no reason why defense experts can't have decent access to source codes. The trade secrets argument doesn't fly (D#3).

<u>Judges</u>

I think [source code] absolutely should be disclosed in every case. I don't see how you can tell the judge, let alone the defense lawyer, [they] can evaluate whether it's a good algorithmic approach or not if you don't know how what went into the source code and what its components were, how they were arrived at it, and so forth. And, give me a break about trades secrets. I appreciate that companies like to make money, but we're talking about human liberty here, and that has to trump any concerns over trade secrets (J#1).

I personally think that it should be open source codes, period. ... I respect the fact that there's intellectual property issues and so forth that's around that, but I think that we have mechanisms to assist in protecting that (J#2).

I think that what it means to be trustworthy is very close to what it means to be reliable, but I think it incorporates something else. Reliability is simply, "does the tool work as it is intended to work?" And it's almost like, "does it calculate in the correct way," is it reliable in that way. Trustworthy certainly incorporates that, but it [also] incorporates something else, which is a concept of fairness, and that has got a subjective component [and] sort of normative component as well. What I would say is that we have got to determine first, what is our standard for that form of fairness that we're aiming towards? ... A trustworthy [algorithmic] tool would achieve both a reliability in terms of functioning as the tool is intended to function, so it has, for instance, an output score [or] an outcome that is expected, but also achieves a level of fairness that I think is quite a complex question, but it's both of those things. ... I think that source code is important because it goes not only to understanding reliability, but you can tell reliability with output, but source code tells you something else about the selection of the inputs and the weighting. ... So, to sort of reduce the importance of source code down to a memorable phrase, I would say "the means to the end matter." The source code matters because in the criminal justice area, we are in a unique area in the American system where we have through our Constitution set out a framework for liberties, where there is due process and due process on an individual level. When we're dealing with due process and equal protection under the United States constitution, we are now in a world where "the means to the end" matter, the means are contained within the source code. ... In my view, if an algorithm is going to be used for a liberty-based decision, a criminal defendant is entitled to have access to the source code, and I would say for an adequate defense, just as a criminal defendant is entitled to the experts that he or she can demonstrate are needed to put on an adequate defense, that same individual is entitled to an expert who can then help them analyze the algorithm. ... [On the issue of trade secrets,] the one thing that courts know how to deal with are trade secrets, because it is frequently the case that there is information disclosed every single day in courts all over this country that is top secret and is under protective orders, even a highly confidential, or attorney's eyes only, kind of a restriction on the protective order. ... The reality is that there is not a reason for a court to deny access to source code based upon competitive issues. That's not what these defendants are interested in doing. They're not going to go running out and open a competing business (J#3).

Other (Academic Scholars)

I think that we really need to be as transparent as we can be for legal purposes, and this is where I do continue to separate machine learning algorithms from straight programmed algorithms. On the programmed algorithm side, we need to make sure that we are validating them in a careful, risk-based manner and understanding the risk points of the processes, the weak elements, [and] the combinations of circumstances that would make their use more problematic. [Further,] we need to be taking it as a whole system—not just the algorithm, but the data that we're feeding it, the person who is making decisions, ... [and] the person who's explaining it at the other end. So, rather than just concentrating on the algorithm, I think that validation needs to concentrate on the end-to-end system and really understand those weak points. ... For the machine learning algorithms, I think that there really has to be some explicit thinking about where the safeguards are going to be and how and when those machine learning type algorithms would be admissible before

anyone tries to put them through as evidence. ... [As for source code disclosure,] I don't think there's a real place for secret science in the criminal justice system. I don't think the basis of an algorithm should be kept secret. I think that the models that [the algorithm] uses and how it works should be in the public domain. ... However, I just feel that the circumstances in which [source code] would be required are pretty rare. ... [Now,] if it happened that in a particular case, the functioning of the algorithm was so central to the actual issue of relevance in the case that it required disclosure of the source code, then yes, disclose it. ... [The problem is] if you were to disclose [the source code]. Who's going to make sense of it (O#1)?

There has to be empirical studies with known source samples that are carefully chosen to match the kind of samples that are processed in casework. And then the proof is in the pudding, you have to see how the method performs under circumstances where you know what the right answer is. You have to do that to assure that it works well for the typical case, and then you have to continue doing it to explore limiting conditions [and] to push the system until it breaks. You need to know the breaking point. Your worry is people are going to be over-confident or there's a problem or error in those borderline cases, and you can get into trouble working with borderline or inappropriate cases if you don't know where the borderline is. ... [As for source code disclosure,] I think framing it as a balance of countervailing risks or issues is the right way to put it. ... The benefit is that the parties, by reviewing the code, might find some issues or some problems that otherwise wouldn't be found if they weren't exposed to outside critical scrutiny. ... Review of the source code by defense experts has in fact uncovered several instances of problems. Although specific examples haven't necessarily been terribly consequential, at least there's the claim that "reviewing of source code is utterly useless," which I've heard people make, has been disproved. ... On the other side, the risk of course is that the source code is intellectual property, and people who have invested thousands or millions of dollars into developing it don't want it to be stolen. ... Has intellectual property theft occurred as a result of defense disclosure? As far as I know, there's like zero, none. ... So, from my point of view, if you balance it, should courts be allowing this disclosure? Yeah, I think it should be done (O#2).

Well, two things: transparency and performance testing. Transparency, because if I were in some sort of legal situation where an algorithm played a role in determining my freedom or even more consequentially my life, I would want my attorney(s) to have the ability to bring their experts to look at the algorithm [and] to make sure that I wasn't a victim of bias. So, transparency is for me, the first thing. The second thing is [that] I want these algorithms tested on a regular basis, looking for failure modes. I want the reliability testing as part of the use of it. ... [As for source code disclosure,] I'm very much aware of the issue of proprietary software and I also pay deference to it. So, if source code is exposed to examination, it should be done so under the conditions that those who are doing the examination are legally, with severe penalties, required not to disseminate that information. In other words, you want to build a fence around your experts so that if they disclose outside of the boundary case, they pay a heavy penalty, which can mean jail, as far as I'm concerned. If you can't get that kind of trusted system in place, then you fall back on reliability testing. Again, you should have a right to test the reliability of the codes as they produce probative evidence (O#3).

Participants' responses related to the use of algorithms based on AI/ML methods:

Laboratory Managers

> *I can test the black box and show it's fit for purpose. … Here's my acceptance criteria. I do my testing. It meets the criteria. It works. It's fit for purpose. Now that I determine it's fit for purpose, the better you can make that, if it's self-learning, it [has a] competitive advantage. So now I've got these two [options], this one is static, and [the other] one is self-learning. The fact that this one can get better [distinguishes it] from the other one. I'm going to choose the better [option]. ... So, you can't turn over source code, [well] I didn't really see that as being a real problem before. … If it provides a better value of results, which I should show through my validation, my ongoing testing, I should always be picking the one that's better (LM#1).*

> *I think that it is appropriate to use them. You need to have appropriate both positive and negative controlling that are inherent in the data set every time, to demonstrate every time that the result has come out as appropriate. ... [Not knowing the full limits of a black box system] is a concern, and part of validation needs to press as hard as you practically can at where the limits are. I am a big fan of saying test to the point of failure. How do you really know where you need to back off if you can't find a point where it breaks? So, provided you worked hard to try and make the stuff fail [during] validation, [then,] yes, my concern is mitigated on the continuing control and setting what that internal standard is, where you try and bracket your expected results. The perfect circumstance would be, I've got controls that bracket my expected result line so I can demonstrate either side and really everything is an interpolation between those controls, not an extrapolation beyond the limits of those controls (LM#2).*

> *I don't have a problem using them, but again, I think the uncertainties and the fuzziness needs to be fully understood. I think the people that are advocating for using it need to be the first people that talk about the limitations of the methodology. ... A lot of this will come out in validation studies. ... I don't think using it is a bad thing, as long as you know the limitations. If we don't know those limitations, taking it to court then could cause more damage than good, and that's a problem. Those limitations have to be understood before it's actually used (LM#3).*

Prosecutors

> *Who am I going to call as a witness at a [admissibility] hearing to explain how this system works that I'm trying to show meets the admissibility standard for my jurisdiction? … I'm quite certain that maybe comparable types of evidence are admitted in certain types of civil litigation, so there must be a way to do it—they have the same rules of evidence that we do, we just have additional constitutional limitations … [Overall,] I wonder if they could be admitted under current Legal rules regarding evidence and admissibility and then constitutional requirements such as due process and confrontation (P#1).*

I would want to see the same type of data to ensure that there's reliability behind any type of machine learning as I would with any type of evidence that is being offered. I have to be able to take my own test drive and understand what it is before I would offer anything. … Prosecutors just can't blindly offer evidence. We have to know that it's reliable. So, whether it's a source code or something behind the source code, I still need to be comfortable. ... [However,] I think that part of what goes into training the algorithm [is where] the science is. Once it's been appropriately trained and then the data supports the accuracy of it doing what it's called to do, then I have less pause for using a machine learning type algorithm. … I think it all depends on if it's good information that's going into the software, then I'm hoping that there's going to be good information coming out of it (P#2).

I would think that you would test that kind of algorithm the same way you do any other technology by using known samples. I know what the findings are, and what kind of answer does the machine give me, and that's how you validate it. So, if it were properly validated, I don't see what the problem is. ... I can see the confrontation issue. I don't see a due process issue, but I can see the argument that would be made. Except I think I might have problems with the concept of something that is completely a black box, because you could put the developer on the witness stand to explain how they came up with it, what goes into it, what the considerations are, what the factors are, what the settings are, what the parameters are. So, maybe I'm having problem with the concept of something that's so completely a black box that nobody understands and nobody can explain, because if it's truly that opaque, then I don't see how I don't see how it's useful. I have trouble with the idea that something is so completely opaque that there's no explanation at all (P#3).

Defense Attorneys

No better to have a machine in there speculating. Probably worse than it is a human being. … You can't have somebody who just turns on the machine and you're coming in and testifying. If we don't know exactly how the machine works, why it works, what its error rates are, how it was developed and why, then it should never be used in criminal court. … It is, in my view, a sixth amendment violation, no matter what—if you were denied your right to confrontation, you were denied due process of law (D#1).

I think that the understanding of machine learning, even for the most highly trained computer scientists, is really [limited]—there's still a lot for us to know. So, implementing machine learning in forensic science is a scary prospect when it's still not that well understood in the broader scientific community. ... It's a tough question [if I would be comfortable at all]. It might in part go back to my answer about some independent board who not only assesses the algorithm itself, but who has a say so in the training data and how that system is trained so that the decisions made on training data makes sense. I'd be even more concerned about AI system unless we had some of those procedures in place to assess whether these algorithms are being developed and validated in a scientifically defensible way (D#2).

I'm not going to say that it should never be used in a criminal context, but it just seems to me, again, that's a bigger argument for more transparency and for a greater testing by people who are independent so at least when these things come out, they come out well before they're ever used in a criminal case so this is all known and it has all played out long before anybody tries to admit this as evidence against someone. ... I think [admissibility] would have to be on a case-by-case basis. ... I think the complication comes in when we try to find out what's behind the black box. It again speaks to the importance of giving lots of people access to the algorithm and to the source code. ... I do think there are valid confrontation clause concerns that courts are really going to have to grapple with. But that's true regardless of AI. That's true when you're talking about just any kind of algorithm even if there is no artificial intelligence involved. ... For example, if the forensic examiner is testifying to [algorithmic] results in court and we ask them questions about how the machine is doing X, Y, and Z and how it's deciding to do this versus that, they can't answer that question. We can't ask the machine those questions so there is a level of hearsay and a confrontation issue. So, I do think there are valid confrontation clause issues that really haven't been litigated as robustly as they need to be (D#3).

Judges

At a minimum you need to know what the error rate is. ... But, also, I'm a little suspicious about any notion in the legal system where we say, "we don't know why X causes Y, but we know it does." You know, if you were in a toxic tort case and you said, "this drug causes cancer, but we can't tell you why. It doesn't fit any of our plausible knowledge about how the human body operates. We just know it does." I think a lot of scientists, a lot of lawyers, would be very skeptical about the use of that because ultimately the law depends on reason, not on assumptions. ... So, I am skeptical of the black box approach (J#1).

They fascinate me and scare me all at the same time. I can't say that access to the source code is the "be all and end all" of anything. I do think that there are some black boxes that we may not know, but I don't even know how to begin to assess that stuff. I think it's got all sorts of potential, potentially good applications, but it's [a] pretty open question. It's kind of scary. ... [Whether the use of these algorithms could be an infringement on Constitutional Rights, such as Due Process or Confrontation,] I think that's part of where it has [been] raised thus far in the litigation that's begun around the country—mostly with the component of open discovery laws and confrontation. I don't think that in the end, it's going to be an absolute barrier. There's lots and lots of courts around the country that have already approved the use of a number of different black box type models in DNA without requiring source code and without requiring things beyond sort of validation studies and so forth, either by the lab or by the industry that created it in the first place. ... But, I really think that if we're going to start using them, that we need to figure out what it is that we do need for purposes of making sure that there's essentially buy-in from everybody, that this is why this is working and that we can have some check on the fact that it is working in the way that we believe that it's working (J#2).

It is true that the source code can be a black box, and the source code can be perhaps incomprehensible to mere mortals. ... What are we going to see if this stuff is nearly

incomprehensible? Well, first of all, there is a design information that's behind the source code. You can find a variety of instructions that are actually in the source code, and you can also talk to the designers of the source code. The "source code" we use is a shorthand—it can also be access to the design of the instrument. So, let's not limit ourselves just to getting a drive with a source code [file]. Let's think about it as access to the design of the instrument. ... Understanding how the instrument was designed is absolutely critical to understanding the calibration of the instrument and the choices. ... They all had a human progenitor at one point in time who designed the objective of the tool, who designed the initial manner in which the tool was going to work. ... So, there's a whole bunch that goes into a source code. What are we going to learn? We don't know until we look. Are we going to reach a point where the source code is no longer informative? Maybe, but I would suggest to you that either we're not there yet, or we don't have to be there yet. There are ways in which we can ask the tool itself to give us information on what it's using as inputs, what it's using as its weightings. We can review those and determine whether they correspond to our sense of fairness. ... If you were in Europe and you were under a GDPR framework, you would be required to make the logic of the output understandable to mere mortals. It is doable. So, I don't believe, and I'm not ready to accept right now that the black box of source code means that we back away from it and say, "ah, it's too complex," because then we have given up extraordinarily important constitutional principles to this black box algorithm. We can't do that. We don't have the right to do that. There is no principle under the American justice system that allows us to do that. ... [Ultimately,] I think there are serious due process issues with a defendant being denied access to understanding information that underlies a tool being used for liberty decision. ... We should not just assume away the importance and the benefits of cross examination [of the algorithm through the expert] because of the complexity of the tool. ... [If the expert is unable to adequately describe the details underlying the tool,] I think at that point in time, the defense counsel could argue that there should be an exclusion of the evidence, without access to that tool, because they're unable to explain what's underneath it, and so we have no idea, we're unable to test it. These tools are unregulated right now. ... What we're doing is we are making assumptions based upon an unregulated set of design criteria that the tool has been made in the right way. I say, it's too important a decision to either leave it completely unregulated or not allow at least examination into the underpinnings of the tool (J#3).

Other (Academic Scholars)

I think if [the algorithm] is not understood to the developers and it's a total black box, then I struggle to see on what basis that there is fair transparency in the [legal] proceeding. ... You can validate black box, [but] you're going to be limited because, to me, validation is based on risk and risk is based on understanding. If you don't know where the weak points are, it's very difficult to do a validation that is sufficiently comprehensive, that it will pick out all the weak points when you've had nothing to inform that. ... You could validate forever and not get to the end of the set of circumstances [for which the algorithm could be applied]. So, how do you find the weaknesses (O#1)?

That's a little scary ... for those [types of algorithms], maybe we just need to rely more on the validation on known source samples. ... I think we should take full advantage of the AI and other approaches to improve our accuracy, but it's particularly important in those cases that if we don't actually understand how it's doing what it's doing, then we may not fully understand how it could break down and where the limits are. So, I would say for those kinds of algorithms, it's even more important to have testing that explores the limitations and where they break down (O#2).

I am not supportive at all in that case, because with modern machine learning, even the people who develop the AI don't know what it is doing. So, in those circumstances, reliability testing is the only thing I know that you can bring to bear. … [Although] they'll be black boxes, … you can give them these large data sets and you can watch how they perform so that you can quantify [and] measure how inputs and outputs are related. ... Theoretically, it could be acceptable to use these systems if we have reliability testing. [The problem is], the testing has to be large and broad because you don't know where the failure modes are and therefore have to do the equivalent of stress testing—you apply the most severe scenarios that you can to test the reliability since, in the case of ML, that's the best you can do. ... [That said,] is this type of reliability testing practical? What I've talked about is the ideal. I don't think the idea is actually practical [and] realizable. I don't think you could actually implement it (O#3).

Participants' responses related to whether algorithms should be regulated and, if so, by whom and how:

Laboratory Managers

I feel that a weakness of our forensic science enterprise is that we don't have a cohesive, guidance mechanism as much as I think maybe we should. I do respect that people don't want to have big brother saying that this is the only way to do it. Yet at the same time, there's danger to our disciplines if we fragment. ... Now that is evolving. We've got OSACs and we're moving in a better direction, but there really is no office of forensic science. There is no central coordination. People can still disregard right now. ... [There is also] a tremendous amount of duplication of effort [across the enterprise]. ... I think [full regulation] would probably be considered by many as an overreach, but the court system in a way should be self-regulating to a point. ... I think it's been fairly reasonable so far and I think the defense community is pretty well interconnected that when [issues] come out, they're on top of it and that information diffuses. So, I think there is some fairly successful self-policing. I just would like to see a little better organization and cohesion to do that (LM#1).

I'm not sure I've got a good answer for that. ... It's not a broad practice, but there's at least the concept of licensing analysts [and] accrediting laboratories. What is really different about a piece of software? … Why wouldn't you license a piece of software? It gives a framework for some audit and accountability. It's certainly going to increase the cost of everything. ... [But] having some kind of framework of audit and accountability probably

has some merit. Who? That was a tough one to say too. The logic would be something that is more [on] the national level, [such as] NIST or [an entity] like that. However, you're dealing with adjudication of these laws being a state's right. How does that quite work? You basically have to have a nationally licensed tool for something that is a state's right. I could see an awful lot of push back. ... I'd love to think [that an oversight regulatory body] was an advantage, but I've seen a lot of places where it gets to be a hindrance really quick. ... [As opposed to a regulatory oversight body that approves specific algorithms or algorithmic tools], maybe it's more about the structure of the requirement for the system of internal standards and controls and the demonstration time over time over time, every case, is really more of the way of doing it (LM#2).

I do think they should be standardized. Regulated, I don't know. I don't know if I have an answer to that. I'm a proponent of ultimate standardization and the industry deciding what's best practice. After that's done, if regulation would help implement, fine, but I've seen too many things have been regulated that shouldn't have been regulated. So, I would rather that the industry itself develop best practice like we're doing with OSAC and like we do in academics before the government actually steps in. Every court case is different. I think the attorneys and the judges should be able to have the flexibility to use the information appropriately (LM#3).

Prosecutors

Certainly, it should not be regulated in any way by the legal side, the legal system. I think that would be a question that would fall within the relevant scientific community. ... I don't think it should be the lawyers at all. I think what you're going to do and how you're going to do it is something that falls on the scientists because you all are the scientists. And then, if we can use it and how we can use it, that falls on us (P#1).

I think [algorithms can be regulated] in the same way that forensic science is already being regulated. It's being regulated through best practice committees and through the court system, and I think that those are putting sufficient limitations around forensic science in general, and that would apply the same with algorithms (P#2).

I think that regulation in a reasonable way gives everybody confidence in the science. ... [However,] I'm not sure what that regulation would look like, and I'm not sure how, for lack of a better word, political, as opposed to scientific, that regulation would be. I've seen things become political very fast, and so I don't know how you stop that from happening. If the regulatory group or body becomes political then it becomes useless as far as I'm concerned. ... I feel like they get hijacked by non-scientists who have a very definite agenda. That does not work very well because the science question has been left way behind and the argument is all about something else (P#3).

Defense Attorneys

Yes [algorithms should be regulated, and] I think it should be an independent scientific entity or something like the Food and Drug Administration, or it should be housed in NIST

[(National Institute of Standards and Technology)], a scientific body that's a measurement-based science. That takes the pressure off of particular practitioners or particular cases or prosecutors. [This way] you can understand the limitations of a forensic technique and [ensure it is] validated outside of criminal court. We do that for every other consumer product, but we don't do it forensics (D#1).

There should be independent bodies to assess their function, their validation, how they operate, who should be able to review training data, who should be able to require the appropriate caveats during testimony, who should be able to require that proper standards are used to develop [the algorithms], whether it's IEEE standards or others. … [The notion that the legal system could regulate algorithms is] really a laughable position. The criminal justice system has proven to be an utter failure as gatekeepers of forensic evidence. We've opened the door to bite mark evidence that wasn't validated. We've opened the door to bullet lead analysis that didn't make sense. We allowed decades worth of misstatements of hair comparison, evidence that overstated the value of it. Fingerprint evidence had overstated the value of it. DNA mixture interpretation that overstated the value of it. And never, even when presented with the other side that raised questions about that, almost never did the criminal justice system have the capacity to properly assess that. That capacity will be even less when it comes to computer-based systems. Judges will continue as they have with human-based systems to utterly fail as meaningful gatekeepers of forensic evidence (D#2).

Yes, and by IEEE or something along those lines. It seems like there should be a minimum [set of] requirements. … I think it has to be an approval authority [and] they also have to regulate how the algorithms are used (D#3).

Judges

Yes, [but] not just algorithms. I think there is a real need for an Institute of Forensic Science staffed by a high-level scientists who could tell us with the neutrality that we deserve, this is good forensic science, this is bad forensic science, this is possible forensic science but it has to be improved and here's how to go about improving it. That was essentially the recommendation of the 2009 National Academy of Sciences report, and I'm very disappointed that it's never developed much traction. ... [While some stakeholders might think the legal system is an appropriate means of regulating forensic science], I think it has proven to be defective. [With] all these cases where there was forensic science introduced and then the guy turned out to be innocent, I don't think that can be brushed off. These are human beings who are being sent to prison, often for very long terms, and [it's] not just one or two, although that would be bad enough, but hundreds because of defective forensic science. Now, can a legal system make that less likely? I think Daubert was a step in the right direction. I'm very disappointed that it hasn't worked in the criminal context the way it has worked more successfully in the civil context, but I don't think the legal system, ultimately, is well positioned to regulate forensic science. Judges know beans about science. Lawyers know beans about science. The natural thing when you have that kind of problem is to turn it over to the people who do know about science, the scientists. So, I think that would be a better approach (J#1).

Yes, [but] the by whom and how is a much harder question. ... [Whether the legal system is an appropriate means of regulating forensic science,] no, [but] I will also say I'm not sure the federal government is the place to regulate it either (J#2).

In my view, there should be a form of regulation that is for any liberty-based decision. It's a broad question in terms of algorithms and any kind of forensic science, … [but] if it's going to be used for a liberty-based decision for a human being, then they need to meet the constitutional standards, so they should be regulated. But, they need to be regulated in a very careful way, by people who are in the field and who are responsible for upholding the constitutional standards in the criminal justice area. ... I think there does need to be some form of regulation. The, how, I think, is extraordinarily complicated, but I don't accept that it can't be done (J#3).

Other (Academic Scholars)

It's not the algorithms that need to be regulated, it's the methods, and the methods include the people, the algorithms, the data, and everything else. ... I think that if we regulate an algorithm, we're not regulating inappropriate use of the algorithm. So, we're much better [off by] regulating the method … because [that] enables us to make sure that people aren't just putting any old [junk] in and that they validated it, that they are trained and competent people that are able to interpret it, [that they] are able to explain in court the conceptual basis of it, and so on. That's what I think that needs regulating rather than just that little bit in the middle, [just the] algorithm, because otherwise [junk] in, [junk] out (O#1).

Yes, I still think it would be nice if we had a national institute of forensic sciences contemplated by the NAS report in 2009. I think the OSAC approach to creating standards is beneficial, but I think that the OSAC approach does not do well for, um, assuring rigor of rigorous validation. ... Right now, we're stuck with the regulatory authority being exercised by judges who, for the most part, have not shown a willingness to apply rigorous quality control with regard to validation of forensic science. ... Years ago, when I first started out interested in this field, I thought the path forward was going to be through litigation. I thought that litigation under Daubert and Frye was what was going to establish a quality control for forensic science, and it just needed people like me to come in and explain to these judges why they needed to set rigorous standards. That didn't work. I spent a number of years litigating cases and I ended up feeling I was getting nowhere. I mean, judges are not well-positioned to evaluate science and they're not competent to do it. They really want to see the evidence admitted [and] they don't want to hold up criminal prosecutions because of uncertainties about some nuance of evidence. ... I ended up thinking that litigation is the worst possible way to try to resolve scientific dispute about validation—it polarizes everybody into opposing camps. So, I'd like to see more federal involvement with agencies that have the ability to make some scientific assessment and set regulations on their own. I think that would be appropriate (O#2).

My intuition is, yes, but I don't know how you could do that. I don't know what a regulatory machine looks like, so I don't really feel qualified to answer this question. ... But [if you

ask me] as a citizen and potential member of a jury, then I don't want them in the court. … If the algorithms are based on machine learning that are total black boxes, I don't want them in the court (O#3).

Participants' responses related to what they would describe as the greatest challenges facing the operational use of computational algorithms in forensic science for court purposes:

Laboratory Managers

Resources. To stay on top of how quick things are developing, it's taking more and more resources. We all have backlogs and we're focusing on those. To take people off of [casework] to train them, then get these new things up to speed and implement them and then change people's minds [takes resources]. ... How can we do a job in a technological field without the resources to bring in these new things? Not only are algorithms coming, they're already here. It's allowing us to do a better and better job. But it takes resources to do that (LM#1).

Resources. Because software itself is expensive, even more so though, is the training and implementation arc of getting people to accept it and understand it, to be able to use it and use it correctly. That's an expensive effort. And let's face it, labs are underwater already. ... Trying to get a group that is underwater, desperately overwhelmed, that can't catch their breath between [cases], to have enough bandwidth to even be able to accept a new tool and not see it as just, "oh my God, you have one more thing." That's going to take time. And, even we don't have bandwidth in there [despite being a relatively well funded laboratory compared to others]. ... That's what's going to face all of these algorithms. ... It's not that people don't see the advantage of them or see the potential benefit, but how do we get from here to there when everybody is madly trying to decide which horrible, awful crime they're going to put first and which horrible, awful crime goes second. So that's what's under that trivial answer of resource. Then, you also think of all the rest of the infrastructure that goes with being able to effectively use these algorithms—the compute, the storage, the data management—where do we put all of these results? How do we store all of these results? How do we maintain that output, which has probably got some proprietary aspect of the outputs in such a fashion that 20 years from now I can still access those results and be able to explain it? Again, it comes back to a resource issue of all of the infrastructure that goes around the use of that algorithm (LM#2).

It's the difficulty in actually developing and implementing [the algorithms]. Getting public data [to support the development], because we have privacy issues and so forth, and then the resources that are needed for the practitioners to begin integrating into their day to day (LM#3).

Prosecutors

We want science to evolve, so we're happy to embrace new things as things get better. But then assuming we got past those [admissibility] hurdles, I would just say making it more

complicated—taking evidence that right now I don't consider to be that complicated and making it more complicated would probably be the biggest challenge for me. ... It would make the presentation of scientific evidence more difficult in trials. ... The more complicated you make that, the harder it gets for the scientists to communicate with the people that they need to communicate with—be it the jury, the attorneys, the cops—to explain what their findings are (P#1).

I think it's getting stakeholders to understand. ... I think [algorithms are] very foreign to people in the entire forensic science community. You're going to get pushback from current forensic scientists [and] you're going to get pushback from all types of lawyers. Judges are not going to understand it. It's just not something that we're used to. So, I think we really want to see the data and understand it. I think that's really the issue, is understanding and ensuring that it's reliable. I just want it to be something that is scientifically valid and clear for our presentation (P#2).

I think training the scientists within the labs, to validate it, and to understand it and have confidence in it. I'm not the scientist. I'm using the science and what I want is reliable science that is easy to understand and easy to explain to lay people. So, if the scientists from my public labs, if they're well-trained, they understand it, and they have the confidence in it—that's the challenge, to make sure that happens because under those circumstances, I think the entire system can have confidence in the output and confidence in the results, even though some parties in the courtroom may not be happy with the results (P#3).

Defense Attorneys

Probably the practitioners themselves. I think that every time [watch presentations] the presenter will get booed off the stage if they don't make it really, really clear that we still need the examiners and they still have to do exactly what they've always done and nobody's losing their jobs. ... It will [also] be money and education, because if you don't have the educational requirements that would be necessary to have people that are engaged in higher level math that it's going to require, then it won't happen. [It will depend on] the amount of money that we're willing to invest in forensic sciences to get them more scientific and also the amount of money that we're willing to invest in the education of our forensic analysts. ... You shouldn't have to fire all the latent fingerprint experts and go to Stanford and hire a bunch of Stanford grads. You have to be willing to invest in the training and get those examiners where they need to be, to understand how to use this machine. ... Status quo is an incredibly powerful force. You could just say that the status quo is going to be what's preventing it. You know, bite mark evidence is still admissible in all 50 states… (D#1).

I think [the greatest challenge is] the people occupying positions of judges without the interest or competency to understand. There's a long-documented history of criminal judges not understanding scientific evidence ... and they don't really have an interest in it. I've been in front of judges who have said on the court record, "it's not really my job to second guess scientists. That's not what I do. I am not qualified to second guess the

scientist." Even though that is the very role of a judge, that it is inherent in the job, in the role of a judge, to do exactly that. ... They will continue to be the biggest hindrance to the assessment of complex scientific evidence in criminal cases. ... It is hard to imagine a day when this reality improves enough that judges have a positive effect on forensic science rather than the effect they mostly have today, which is maintaining the status quo in favor of police and prosecutors (D#2).

[The greatest challenge] is these non-scientists understanding what this machine is doing and the limitations of what the machine [and] results are. [Further,] having a forensic examiner, very few of which have a background in computational . . . anything, explaining accurately to these lay people what this machine is doing and the limitations of what this machine is doing (D#3).

Judges

N/A

Other (Academic Scholars)

Shared understanding is one of them. ... I think we need to be careful that we're all talking about the same thing and that where there are differences, say between programmed and AI strict machine learning [type algorithms], and even within that category that we recognize differences between supervised and unsupervised learning and those such things. Another challenge is that there needs to be a scientific, and not an emotive, debate about the issues. So, rather than there being a whole series of high profile, court cases where the use of a particular algorithm is at the center of the case, I think we're much better to plan and set requirements ahead of time, and really think about if we're going to use algorithms, this is how we're going to use them—these are the validation standards that you have to meet, this is how you have to make sure that the end-to-end process works and not just the bit in the middle, and here is the legal framework within which they work. ... [Finally,] I think we need to really work on education of practitioners and our legal colleagues in terms of fundamentals of probabilistic [concepts], in terms of what it means to be transparent and to disclose limitations, and how we work with these kinds of new technologies (O#1).

Challenge number one is, can we come up with good algorithms? ... The technical capabilities of modern statistical techniques are impressive and there's tremendous potential to come up with machine-based approaches that have the potential to approve on human judgment. So, first challenge is let's come up with really good, robust algorithms and then evaluate them carefully to show that they work well. The second challenge will be implementing them by forensic science practitioners who don't have the training and background to fully understand how these things work. Most people who become forensic science practitioners are not very sophisticated about statistics and in fact find these things kind of frightening. I think the DNA people have been able to adopt and incorporate the probabilistic genotype, but I think it's been kind of wrenching for them. I think it's been really challenging for some of the analysts to come to grips with it. I see a lot of evidence

that, even among practitioners who think they understand these things, maybe they don't understand them as well as they think they do. ... [Additionally,] doing the training needed to operate these things in an effective way and make them understandable in court is another huge challenge. ... I don't think it's a reason not to develop algorithms. We need to be realistic about how easy it is to implement them. ... I think it's the future. I think it's the path forward. I think it will address and resolve a lot of problems that we're facing with human beings making these judgements so I think there are great prospects and making greater use of algorithms will improve forensic science, but it's not an easy fix. It's going to be challenging. It's going to require a lot of training. I think we need to think seriously about, given our movement toward these algorithms, the way we train forensic scientists and select them. So, picking people who have higher levels of mathematical and statistical aptitude training might be really important. At the same time, I think we need to be sensitive to current practitioners who are math phobic and, kind of ease them in and select more of those practitioners who have degrees in math and statistics, or the harder physical sciences and, thus, may be capable of moving into the new world with a greater degree of facility than we may see from the typical pattern matching person (O#2).

The greatest challenge is not to be caught by the race to the bottom, because if you're going to be doing this through commercial entities, it's going to invariably become the race to the bottom in terms of cutting costs. To me, that's the greatest challenge, is how do you do this without people pursuing pathways that cut corners, and therefore we risk the reliability of the systems. I mean, how do we stop that from happening? To me, that's the greatest challenge (O#3).