**Establishing Decision Thresholds for the Analysis Phase of ACE-V:**

**Development of a Utility Metric**

*Mémoire de thèse de doctorat*

**Heidi Eldridge**

**April 2014**

# Table of Contents

[Table of Contents 2](#_Toc386433889)

[Introduction 3](#_Toc386433890)

[State of the Art 6](#_Toc386433891)

[Decision-making is most vulnerable to variability, bias and error near quality thresholds; however, these thresholds have not been defined. 6](#_Toc386433892)

[It is not understood how determinations of value are made or how weight is assigned during feature selection. 8](#_Toc386433893)

[A major contributing factor to utility decisions may be the rarity of the features observed, but this has not been studied. 12](#_Toc386433894)

[Clusters of minutiae may also have a significant influence on weighting in Analysis decisions. 16](#_Toc386433895)

[Analysis documentation provides transparency and may increase accuracy and reliability. 17](#_Toc386433896)

[Once a utility metric has been constructed and validated, it can be utilized for the benefit of the community in a variety of ways. 18](#_Toc386433897)

[Quality thresholds may be established for limiting Identifications, requiring additional QA measures, and reducing QA measures. 20](#_Toc386433898)

[The information used to render a conclusion should be conveyed to a jury in a way that is comprehensible and gives an idea of the significance of the finding, along with its limitations. 24](#_Toc386433899)

[Objectives and Work Plan 26](#_Toc386433900)

[Task One: The Pilot Experiment 26](#_Toc386433901)

[Task Two: The Large Scale Experiment and Development of the Utility Metric 33](#_Toc386433902)

[Task Three: The Documentation Experiment 36](#_Toc386433903)

[Task Four: Validating the Metric Short Form 38](#_Toc386433904)

[Task Five: Validating the Metric for Predicting Comparison Value Thresholds 40](#_Toc386433905)

[Originality of the Project 46](#_Toc386433906)

[Conclusion 47](#_Toc386433907)

[Timeline 48](#_Toc386433908)

[Bibliography 50](#_Toc386433909)

[Appendices 53](#_Toc386433910)

[Appendix A – Terms and Definitions 53](#_Toc386433911)

[Appendix B – Abbreviations and Acronyms 55](#_Toc386433912)

[Appendix C – Utility Score Continuum 56](#_Toc386433913)

[Appendix D – Utility Metric Questionnaire 57](#_Toc386433914)

# Introduction

The ACE-V process[[1]](#footnote-1) is the accepted and most widely-used method for conducting comparisons of friction ridge detail and rendering conclusions regarding the source of an unknown mark. During each phase of this process, numerous decisions are made and some level of subjectivity is unavoidably introduced.

During the Analysis phase, the examiner must analyze the impression, selecting features to be used in comparison and assigning them expected weight. This selection of features and assignment of weight is determined through interpretation of distortion, noise, relative rarity of features, presence of focal areas and searchable clusters of minutiae (known as target groups), and determination of anatomical source and orientation, if possible. In the Comparison phase, the mark (or unknown impression) is compared to a print (or known impression) and decisions are made about the degree of similarity, or dissimilarity, in the features of each. In the Evaluation phase, the examiner makes a decision regarding the source of the mark, based upon the objective data observed during the first two phases, and interpreted using her training, knowledge, and experience. The Verification phase is a repetition of these first three phases by another qualified examiner, who will make his own determinations.

Because these processes take place in the mind of the examiner they are, by definition, subjective. However, “subjective” should not be confused with “arbitrary”. These decisions are not made on a whim, but are the result of careful study of the data and application of knowledge gained through in-depth and lengthy training and experience.

Nonetheless, because the decisions are, in fact, subjective, they are subject to a level of inconsistency between examiners, and even within-examiner inconsistency at different viewings has been documented (see, e.g. (Dror *et al.* 2011; Langenburg 2012)). This potential for inconsistency can reflect negatively on fingerprint comparison science in a variety of ways. First, consistency (repeatability and reproducibility) is a component of the reliability of a method (Langenburg 2009). Without demonstrated reliability of results and conclusions, fingerprint examinations may not meet the threshold for admissibility in a court of law. Second, inconsistency leads to professional disagreements. If two experts cannot agree on their conclusions after reviewing the same data using the same methodology, both the methodology itself and the expertise of the experts are called into question. Third, without consistency in methods and terminology, experts are not speaking the same language, and thus cannot communicate effectively, or conduct and report meaningful research. Without consistency of methods and terminology, there will always be some level of confusion regarding *which* version of the method was being used, or *what* level of difficulty was tested, or *how* the conclusion was reached, to name only a few. In fact, this very concern was expressed by the Habers, two of the discipline’s harshest critics, in a recent paper (Haber and Haber).

Many of the studies on the ACE-V process to date have treated the examiner as a “black box”, testing only the outcomes, or conclusions, of the ACE-V process without examining in detail the ways in which these conclusions were reached (see, e.g. (Langenburg 2009; Ulery *et al.* 2011)). However, there has been recent movement toward opening the black box, as it were, attempting to elucidate specific decision-making criteria at various points throughout the process (see, e.g. (Langenburg *et al.* 2012)).

This work will attempt to open a piece of that black box and examine examiner decision-making during the Analysis phase, specifically in the areas of weight given to features due to rarity and target groups, and decision thresholds for different levels of quality. As a tool to aid in understanding these processes, and to improve the speed and consistency of their application, a new quality metric will be developed and validated.

The concept of a quality metric has been addressed before and will be discussed in greater detail below. However, briefly, current quality metrics seem to focus on only one dimension of quality, typically either clarity or minutiae count, while ignoring factors such as rarity of features and their groupings that may greatly impact the overall usefulness of a mark. They also tend to focus on a single threshold: value versus no value; or low, medium, or high quality.

This research will attempt to develop a more nuanced and flexible metric that is able to give information on a variety of Analysis decision thresholds, to include: value v. no value; complex v. non-complex; AFIS quality v. AFIS quality, but proceed with caution[[2]](#footnote-2); standard quality v. high quality[[3]](#footnote-3); appropriate for a definitive conclusion v. definitive conclusion not warranted[[4]](#footnote-4); and low, medium, and high difficulty.[[5]](#footnote-5) For this reason, it will be referred to throughout this research as a “utility metric” as it will measure the utility of the mark at a number of levels, and not be limited to a single aspect of “quality”.

Finally, in light of these findings, recommendations will be made on the development of policies and training will be developed.

# State of the Art

## *Decision-making is most vulnerable to variability, bias and error near quality thresholds; however, these thresholds have not been defined.*

One of the subjective determinations made during the Analysis phase of the ACE-V methodology is the determination of “value”. Each examiner has a personal threshold for where value is achieved, and many also differ on their definition of value. Value can variably refer to a mark that contains sufficient information to be identified to a source, to a mark that contains sufficient information only to definitively exclude a subject, or simply a mark that contains sufficient information to provide some useful information to inform an investigation.

Even once value has been established, there can be the issue of complexity – the of-value mark represents a wide range in quality of marks from the barely-of-value mark to the mark that is so clear and complete, it may exceed the corresponding standard in quality.

Many of the existing reports discussing bias, error rate, and variability of sufficiency decisions (see, e.g. (Dror *et al.* 2005; Langenburg 2012; Ulery *et al.* 2012)) have determined that the greatest fluctuation occurs on “complex”, “difficult”, or “borderline” marks, yet these terms are not defined. Most often in studies, the difficulty of the test mark is determined by a consensus of some number of experts, but no objective measure of how those experts came to their determination is given, nor is discourse typically had on the treatment of outliers. Likewise, one of the most prevalent criticisms of commercial proficiency tests, such as that offered by Collaborative Testing Services (CTS), is that the difficulty of the marks given on the test is not representative of casework, or is not challenging enough. However, once again, no measurement has been offered of the marks on these tests, or of typical difficulty of casework, for that matter. Terms are being thrown around that have no real meaning. What is a “difficult mark”? What is “representative of normal casework”? Without objective ways to define and standardize these terms, meaningful communication on the issues surrounding them seems a hopeless task.

Given the subjective nature of fingerprint examination, it seems clear that these marks of various qualities cannot be treated the same way. More complex marks should be afforded more time for consideration, be analyzed with more care, be documented more thoroughly, and be reviewed more closely (Ashbaugh 1999; Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST) 2012). It may be reasonable to require additional quality assurance (QA) policies or to require examiners to stop short of an Identification decision in the case of increasingly complex marks.

However, since the determination of “value” and “complexity” varies from examiner to examiner, it is difficult to construct policies to govern the examination of a complex mark, when two different people may give two different opinions on whether or not the mark is complex. Likewise, it is difficult to set up testing situations, be they for research or for proficiency testing, in which varying difficulties of marks are to be presented, when there is no standard metric of fingerprint difficulty. A study meant to test conditions at the threshold of an identifiable mark, or a proficiency test meant to be a challenge, has little meaning if the marks being used do not represent the level of difficulty that is sought.

One recommendation of the NAS Report (National Research Council 2009) is to require certification of forensic scientists in any discipline in which they practice. To fulfill this goal, every latent print examiner would have to take, and pass, some sort of standardized test demonstrating their continued proficiency at fingerprint comparisons. However, the utility of such a proficiency test would be severely limited if all of the marks on the test were very easy. Is the purpose of the certification intended to be a mark of minimal competence, standard competence, or achievement? Wouldn’t it be useful for quality managers to know which of their employees were competent at all levels of difficulty, and which had particular weaknesses that might require further training?

Mnookin suggests (2010) that testimony might be offered in conjunction with proficiency test results. She calls for the development of a metric for use in designing proficiency tests. She specifically recommends that this metric of difficulty be validated and be applied to proficiency tests in such a way that the level of proficiency of an examiner may be ascertained and compared to the level of difficulty of the prints being presented in a particular case. In this way, a jury may feel comfortable that the expert testifying to a conclusion is qualified to do so, based upon having demonstrated proficiency at the same level of difficulty as the marks present in the case at hand.

Current proficiency and certification tests have not been standardized to any objective metric and are not capable of testing differing shades of competence.

## *It is not understood how determinations of value are made or how weight is assigned during feature selection.*

While ACE-V prescribes broad steps for the friction ridge comparison process, the interpretation of how to apply those steps varies widely among agencies, and even among examiners within the same agency. Every decision made during the process is, in effect, a black box. This leads to a lack of consistency in decisions, particularly at the Analysis phase (see, e.g. (Neumann *et al.* 2013; Ulery *et al.* 2013)).

Adding to the difficulty of studying Analysis phenomena is the fact that, when it comes to Analysis, there really is no “ground truth”. While the Evaluation decision can be checked for correctness in a controlled study (i.e. there is a known source, and only one correct source, of each mark), the true “correct” Analysis judgment can only be based upon opinion – it is a value determination, not a statement of fact.

How then does one evaluate the quality of an Analysis model? Typically, ground truth in Analysis must be determined by proxy, through use of an expert consensus. The performance of the model can then be tested by seeing how well it is able to predict the consensus judgments provided by the experts.

There is a great need for studies that examine each decision in a controlled environment and with appropriate documentation of the decision-making process such that thresholds for decision-making may be determined and variability between examinations can be reduced to the extent possible.

It is not understood what specific factors go into determinations of value, complexity, and assignment of weight. Ashbaugh suggests factors that should be considered during Analysis (substrate, matrix, red flags, etc) (Ashbaugh 1999), but their interpretation and weighting is left up to the individual examiner. It is not understood what the interaction is between these factors. For instance, what is the interaction between amount of distortion versus number of minutiae in a value decision? How many minutiae are needed to make up for a mark that lacks clarity? Or, what is the weight given to a rare feature versus a lack of clarity? How rare must the feature be to tip the balance? An understanding of how weight is assigned to information in the mind of the examiner is needed to begin to understand the interplay between the various factors considered at the many decision checkpoints in the process. It is possible that some types of information have a greater ability to outweigh, or trump, other types.

While the Analysis phase has been defined as an information-gathering phase where the quality of the mark is assessed, and the areas to consider during Analysis have been well laid-out (e.g. Levels One, Two, and Three detail, distortion factors, anatomical source, orientation, size of the impression, etc), nobody has broken down the steps, or phases, of Analysis itself.

This work proposes that there are three distinct steps, or tasks, that occur during the information-gathering of Analysis:

* **The Observation Task – What do you see?**

The first task of Analysis is to catalog the observations that are made of the mark. This may be done sub-consciously, or explicitly; mentally, or with extensive documentation. Regardless of how it is done or how it is documented, the first step is always the same – look at the mark and determine what information is visible.

* **The Assigning Weight Task – Is it useful?**

During this task, the features that were noted are assessed for their value to the mark, and ultimately, to a comparison. The assignment of weight to each piece of observed data is determined by answering (again, whether sub-consciously or explicitly) two questions:

* + - How distinctive is it?

This concept is often expressed by use of the term “selectivity”, which is meant to convey the discriminating power of the feature. However, selectivity is a statistical concept that is not often clear or linguistically accessible to jurors, along with being frequently misused by latent print examiners themselves. Hence, clearer terminology is desirable.

The author prefers the use of the term “distinctive”. This is a word that is in common usage in the English language and will be more familiar to the average juror. Its definitions include “markedly individual”, “notable”, and “serving to differentiate”. Thus, it neatly encapsulates the real crux of the concept of selectivity – when a feature is being selected, the examiner is considering its rarity (how markedly individual it is); whether it is notable (if there is something quirky or unusual about the feature that makes it stand out and grab one’s notice - that feeling that the examiner would be certain to recognize it if they saw it again); and, overall, its capacity to differentiate (whether it can help to discriminate *this particular mark* from other marks that may have a similar appearance).

While formal training on rarity is not typically a part of an agency’s training program, examiners do learn through experience to recognize pattern types, minutiae, and level 3 features that are less frequently encountered. These less common data are given relatively more weight both during Analysis and Comparison, such that a few unusual features may be enough to tip the scales between the decision to keep a mark or not, or the decision to identify or not. When one adds in the extra element of some features having that extra bit of “personality” that makes them really stand out in the memory (this may properly be termed as Level 3 detail), it can contribute even further to the weight the examiner is willing to assign to the data.

* + - How confident am I?

Because every touch differs slightly in pressure, deposition matrix, substrate, etc, no two impressions can ever be exactly alike. Thus, every time a latent print examiner looks at a mark, there is some amount of distortion present and the examiner is doing some level of interpretation of the differences between two impressions, and assigning tolerances for how much dissimilarity is acceptable.

While they may not be aware of it, when the examiner is interpreting distortion and assigning tolerances, what they are really doing is assessing the *risk of error* associated with each feature. How certain are they that something is there? How certain are they of the identity of the thing they see (e.g. is it a ridge ending, or a bifurcation? Is that pattern a loop, or could it be a whorl?). What is the chance the examiner has misinterpreted what they think they see? What is the penalty for such a misinterpretation? The murkier the image, the more difficult it is to answer these questions, and thus, the higher the chance of making an incorrect determination.

On top of this, there are cultural, policy-driven aspects to be considered. Some agencies (for example, the Dutch experts discussed by Langenburg (2012)) put a premium on only utilizing minutiae that are clear and unambiguous. Others may have additional documentation requirements for marks that have low clarity, or are deemed “complex”. For an examiner operating in such an agency culture, a minutia in a smudged or low-contrast area would carry considerably higher risk of negative consequences. Thus, they must decide for each feature that they *think* they see, how high is the risk associated with using this feature? Is the risk worth the potential benefit (i.e. having the use of that feature during the Comparison phase, or having sufficient features to even proceed to the Comparison phase)?

Of course, this concept of *risk* has a flipside with which most examiners are more familiar – that of *confidence*. Risk and confidence have an inverse relationship – the lower the risk, the higher the confidence. Thus, when an examiner assigns high confidence to a feature, they have *de facto* determined that the risk of an error (or the risk of an unacceptably negative outcome, should there be an error) is low. Conversely, if the examiner assigns low confidence to a feature, they are indicating that there is a high chance the feature is not what they think it is, or is not present at all, and that there is a higher risk of error (or negative consequences) associated with its use.

In the assigning weight task, the two questions work in concert to settle on an overall, appropriately balanced, opinion of the value, or usefulness, of the mark. Does this mark contain sufficient discriminating power, or weight, to be worth carrying forward into a comparison? This brings us to the third task of Analysis.

* **The Decision Task – What decision do you make?**

In this final step, the actual Analysis decision is made. The examiner considers the data that were observed, the cumulative weight of those data, along with any associated concerns about the reliability of the data, and makes a determination about the anticipated usefulness of the mark. Depending upon the particular agency policies, the examiner may consider whether the mark can be used to identify; whether it is of sufficient quality to exclude only[[6]](#footnote-6); or whether it can merely be used to inform an investigation, even in the absence of the ability to render a definitive conclusion. This consideration results in the determination of “value” or “no value” for the mark.

## *A major contributing factor to utility decisions may be the rarity of the features observed, but this has not been studied.*

To date, several attempts have been made to understand the nature of quality assessments and to create an automated metric of quality. Hicklin et al (2011) conducted a survey of examiners aimed at creating objective metrics. However, they explicitly stated that to them, quality was equated with clarity (“…determined in terms of the confidence that the presence, absence, and details of features can be precisely detected”) only and no other factors affecting the quality of a mark (such as specificity or quantity of features) were considered. Furthermore, they explicitly instructed their participants to ignore any agency notions of utility, saying:

“The participants were instructed to base their assessments on their fundamental understanding of friction ridge impressions with no operational goals or legal consequences, not to invoke any agency practices or policies for the analysis of a latent print, and not to consider whether they would testify in court to their assessments”.

2012 saw two published reports on the development of automated metrics for fingerprint quality. Both were focused on a lights-out approach for use in an AFIS environment. The study by Yoon et al (2012) focused on average ridge clarity and minutiae count as the main factors driving quality. Rarity of features was not considered and the focus of the research was on looking at the probability of getting a hit in the top-100 candidate list. The other study of 2012 focused on a quality metric was a Final Technical Report for the National Institute of Justice (NIJ) presented by Murch et al (2012). This report focused on the development of software for automated feature extraction for use in a lights-out AFIS system. While this study did consider “rare features”, they defined these not as minutiae types that were less frequent in the population, but as rare configurations of multiple minutiae that were designated by creating triangles between minutiae and measuring the ridge counts between minutiae pairs to come up with statistically unusual compound features.

In 2013, joint research by the Federal Bureau of Investigation (FBI) and Noblis, Inc. was published that presented Quality Mapping software that utilized a color-coding scheme for local clarity assessments (Hicklin *et al.* 2013). This research, as with their previous 2011 work, was at pains to distinguish between quality and clarity, and emphasized that this software focused on only the *clarity* of the mark, not its overall quality (a label that they define to include the quantity and distinctiveness of features). Local clarity scores were aggregated into an overall clarity score for the mark and these results were compared to subjective results from an examiner survey with good correlation. However, this study did not take rarity of features or minutiae counts into consideration. It did suggest that the quality maps could be used for designating complex comparisons and for determining areas of a mark that may not be given weight due to deficiencies in the corresponding areas of the print. Langenburg (2012) suggested this tool could be used for quantifying the number of pixels in each quality category versus the total area of the image to calculate an overall quality score for the image.

Additionally, in 2013, a group at Pennsylvania State University published on a quality grading system that utilized three readily-available software programs to produce a measurement of the quality of a mark (Pulsifer *et al.* 2013). This process utilized clarity, identifiable minutiae, and percentage of area with identifiable minutiae to calculate the grade of the mark. As with all the others, this metric focused on the *quality* of the mark (i.e. its clarity and minutiae count), not on its overall *usefulness* (i.e. what can reasonably be *done* with the mark), nor does it provide quality thresholds for different decisions.

While each of these approaches is a step in the right direction, none account for the variable of rarity of the features present in the impression. Each of these tools focuses on the image quality of the mark, without considering the specificity of the information contained therein, which is the cornerstone of *discriminability*. As discriminability is what allows source attributions to be made in the first place, neglecting to consider it in a sufficiency decision would seem to fly in the face of the discipline’s claim to be able to associate a mark back to a specific source.

Examiners tend to give additional weight to what they perceive as rare features. Simply put, every examiner knows that a plain ridge ending by itself is more common than a compound feature, such as a spur or an enclosure. Most examiners would even agree that some compound features, such as enclosures, are more common than others, such as trifurcations. Thus, when a rare compound feature is seen during comparison, it is given relatively more weight than a common feature would be given. Trained examiners do this instinctively, but the accuracies of their weight attributions are not known. A brief investigation by Osterburg (1964) indicated a lack of consistency among examiners in assigning a subjective rank to the frequencies of ten minutiae types.

Various models[[7]](#footnote-7) have been constructed in an attempt to demonstrate the strength of associations made between marks and prints or to support the uniqueness of fingerprints. Many of these studies have included a survey of the frequencies of specific features within a sample population[[8]](#footnote-8). However, even in the more modern studies, these frequencies were typically taken in limited areas (e.g. distal portions of thumbs (Stoney and Thornton 1987)) and often from limited pattern types and fingers (e.g. ulnar loops and whorls from index and middle fingers (Champod and Margot 1996)).

A survey of the Spanish population (Gutiérrez-Redomero *et al.* 2011) and a follow-up study incorporating Argentinian data (Gutiérrez-Redomero *et al.* 2012) took a more comprehensive approach and established frequencies of many types of minutiae in a large area of each finger and across all ten fingers. This data provides an excellent starting point for establishing usable population frequencies of minutiae type. However, the NIST Expert Working Group on Human Factors in Latent Print Analysis (hereafter “NIST Human Factors Group”) calls for still more studies to establish this important information for use by the friction ridge examination community (Expert Working Group on Human Factors in Latent Print Analysis 2012). Gutiérrez-Redomero et al also call for their research to be repeated in additional populations to establish the distributions of minutiae type more broadly (Gutiérrez-Redomero *et al.* 2012).

Once population frequencies of various minutiae types have been established, examiners can then be tested, and if necessary, trained, to calibrate the weights they assign to different rarity types to reflect real world values, rather than a nebulous hunch based on experience. While most examiners have developed a feel for relative frequency through years of experience looking at marks and prints, no formalized training exists on the actual frequencies of these features, or on the appropriate assignment of weight. Anecdotally, the author is also not aware of any examiners who factor the finger position or location of the minutiae on the finger into their assessment of feature rarity; however, the Gutiérrez-Redomero et al data have demonstrated that both factors affect the frequency of minutiae. Therefore, a particular minutia should be given more or less weight, dependent on its location. These are areas of training that are sorely lacking and need to be addressed.

With the advent of the NAS Report, court challenges to fingerprint admissibility are plumbing the depths of the “hows” behind what latent print examiners do. It is no longer sufficient to testify that some features are given more weight than others. The opposing counsel wants to know how we *know* the feature is rare, and how much rarer it is than others[[9]](#footnote-9).

The rarity of features should, in fact, be incorporated into decisions at both the Analysis and Evaluation steps of the process. Neumann argues that in order to achieve a consistent logical framework throughout the ACE-V process, the emphasis should be on the discriminatory power of the *mark*, not the level of agreement between the mark and the print (Neumann 2012). However, this cannot be accomplished when examiners are only evaluating specificity of the mark in a sub-conscious (and likely inaccurate) way, and when no standard measure of the specificity of a mark is being applied.

Neumann further states unequivocally that “…the number of features itself is not a sufficient indicator of the specificity of the marks” (Neumann 2012) and his data clearly support that a mark with a low number of highly specific minutiae can contain more discriminating power than a mark with a high number of minutiae of low specificity. This contradicts the most recent publication from Ulery et al, the first paper discussing the so-called “white box study” (Ulery *et al.* 2013), which finds that minutiae counts alone were the single biggest factor driving sufficiency decisions. However, the Ulery group did note in their paper that clarity and minutiae count were correlated in their study marks such that the two effects could not be separated. Additionally, specificity of features was not accounted for in this study.

While minutiae count is a significant factor in the sufficiency decision, Neumann’s work has shown that it cannot be the only factor under consideration. The reasoning behind the 1995 Ne’urim Declaration that “[n]o scientific basis exists for requiring that a pre-determined minimum number of friction ridge features must be present in two impressions in order to establish a positive identification” (Almog and Springer 1996) was that not all minutiae are equal and thus cannot be given equal weight. That maxim is as true today as it was in 1995. Thus it seems clear that not only notions of minutiae count and clarity, but also rarity of features must be taken into account when designing a metric of the utility of a mark.

## *Clusters of minutiae may also have a significant influence on weighting in Analysis decisions.*

As with rarity of features, another criterion of quality that has been hitherto overlooked in the design of quality metrics has been the presence of data clusters, or target groups. Conventional wisdom and friction ridge comparison training have long held that the best way to search a mark is to start by locating an easily-recognizable cluster of minutiae, known as a target group, and memorizing this image to use in the search.

Because the presence of an easily-recognizable target cluster of data aids in the searching task, their presence will often figure into the Analysis decision. While a mark may be located and identified without any clear target clusters of data by using a tedious, laborious, and time-intensive method of ridge-by-ridge searching, the time-saving benefits of a target group are obvious.

While the presence of a target group may make the searching task easier and faster, it is less clear if there is any intrinsic benefit to the target group itself. Does the target group provide weight in the ultimate Evaluation decision? Or is it merely a location tool? If the target group does provide weight to the Evaluation decision, it must be due to the rarity of the features within the target group. Groupings of highly discriminating features carry more weight than groupings of low specificity features much in the same way that a cluster of two or more simple minutiae in extremely close proximity may form a more discriminating complex feature[[10]](#footnote-10). Thus, the value of a target group is closely linked with the concept of rarity.

Whether the target group is noted during Analysis for the searching benefit it may provide, or for the value in specificity offered by its components, it seems clear that the presence of one or more target groups is a criterion that contributes to the overall usefulness of the mark, as assessed during Analysis. If this is the case, this variable should also be accounted for in a utility metric. A mark that has one or more easily-recognizable target groups should be easier to locate and identify, and likewise may be identified with greater confidence, than a mark that has a spattering of solo minutiae spread throughout the mark without any neighbors to anchor them.

## *Analysis documentation provides transparency and may increase accuracy and reliability.*

The amount of documentation required to demonstrate the steps taken during the ACE-V process varies from agency to agency. SWGFAST provides a standard on documentation (Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST) 2012), but even this provides only a broad minimum guideline and its interpretation and implementation vary widely between agencies.

Without documentation of the data observed and the thought process behind ACE-V, both transparency and consistency are very difficult to achieve.

It is possible, and has been suggested ((Expert Working Group on Human Factors in Latent Print Analysis 2012; Langenburg 2012)), that the very act of documenting data and thought process increases accuracy and reliability of the process, particularly in complex cases. This seems probable for two reasons. First, the act of slowing down and carefully analyzing each bit of data in order to record it in the notes may in and of itself promote greater care and accuracy. Second, by taking notes, an examiner’s accountability is increased, for she knows that someone else may come behind her and scrutinize her work. This knowledge of accountability may cause her to exercise greater care in feature selection and other decision-making and may lead to greater conservatism, which as Langenburg (2012) demonstrated, leads to greater consensus, although it may also lead to a decrease in sensitivity.

At a minimum, increased documentation allows us to study the thought process behind decision-making in ACE-V, which can lead to the development of better practices to increase the reliability of the process.

Several schemes have been proposed for quick and easy documentation of the ACE-V process. One of the best known is the GYRO system (Langenburg and Champod 2011). In GYRO, individual features are assigned a color value based upon the examiner’s confidence in the existence and identity of the feature. A separate color (orange) is reserved for features that were noted after Comparison was begun and the exemplar was seen, thus making any potential bias introduced by the exemplar transparent.

The colors used by the examiner with GYRO presumably give some indication of the weight the examiner assigned to each feature based upon his confidence in the feature (i.e. a green, or high confidence, feature would be assigned more weight than a red, or low-confidence, feature). However, as with the current quality metrics, GYRO does not account for rarity of features or feature groupings. Langenburg (2012) states that there are two components to weight – examiner confidence in the presence and identity of the feature, and the rarity, or discriminating power, of the feature. GYRO only accounts for one of these components. A modification to document the perceived rarity of the feature would give a more complete picture of the decision-making that went into the Analysis of the mark.

## *Once a utility metric has been constructed and validated, it can be utilized for the benefit of the community in a variety of ways.*

Operationally, it would be cumbersome to apply a quality metric to every piece of ridge detail that is examined in a crime laboratory, nor would it be necessary. For the vast majority of marks examined, value decisions are straightforward, and complexity is self-evident. However, there are five distinct arenas where an objective measure of mark quality could be of use:

1. Quality Assurance

Many laboratories are striving to improve their quality assurance measures and documentation, in order to provide additional safeguards against errors. One example would be to require additional documentation or review of complex, difficult, or borderline marks. However, it is difficult to write policy that is predicated on terminology that has not been defined – how does one require under written policy that an additional verification be done on all complex marks, for example, when nobody has defined what constitutes “complex”? On the other hand, marks of very high quality could be subjected to fewer QA policies, which could save laboratories time that could be better allocated to more difficult marks.

1. Proficiency Testing

Companies such as CTS sell proficiency tests to laboratories throughout the United States. These test results are often invoked in court as proof of an examiner’s continued competence. However, the test can fluctuate drastically from year to year. In some years, the tests are so simple as to be insulting to the skill of the examiner. In other years, there are marks on the test that are so poor in quality, many examiners would not even have called them of value for Identification were they encountered in actual casework.

There is no stated standard of how many easy, medium, or difficult marks are given on the test. Without this information, it is unclear what level of proficiency is being tested. In order for the results of these tests to be meaningful, proficiency test companies should be employing a standard metric when designing their test to ensure that each year, a pre-determined number of marks from each category of difficulty is given to test-takers. Only by doing this can there be certainty that the test was of sufficient difficulty each year to warrant a claim of “proficiency” by its successful completion.

It may also be useful to design tests of differing difficulty, or to award proficiency at differing levels based upon the level of difficulty of the marks that were successfully completed on the test. This would give both managers and juries a clearer idea of the skill level of individual examiners. For managers, this information could be used to identify examiners who would benefit from additional training, or to select the most skilled examiners for particularly difficult or high-profile cases. For jurors, it could give a sense of how concerned they need be about the likelihood of an error made in this particular case, by this particular examiner, rather than trying to extrapolate from some general error rate for the field.

1. Certification

As with proficiency testing, certification tests should be a carefully standardized test of some level of competence. Whether the goal of certification is to ensure minimal competence, or to be a mark of achievement, the difficulty of marks correctly identified to reach each rank must be defined by an agreed-upon measure, or the distinction is without meaning.

1. Research Design

Since the release of the NAS Report in 2009, the one thing that fingerprint examiners and critics alike seem to agree upon is the need for more research. The Report demanded answers to a variety of questions related to fingerprint comparison science, including questions regarding error rate, standards, proficiency, bias, and many others. Researchers are racing to develop experiments that will provide the data to answer these questions and, hopefully, will strengthen the foundation upon which friction ridge comparison science rests.

Most, if not all, of these research projects will hinge upon test subjects (typically examiners in the field) examining some number of known and unknown impressions and reaching some conclusions about them under a variety of circumstances. In every case, the results of these enquiries could be better standardized, and their conclusions more meaningful, if the difficulty of the marks used to conduct the research was standardized. There must be consistency from researcher to researcher on what constitutes a “difficult” or “borderline” or “ambiguous” mark.

1. Testimony

Currently all Identifications are treated as equivalent in value and certainty. However, it has been shown that more errors occur in the comparison of marks that are of marginal quality (Langenburg 2012). Furthermore, not all examiners possess equal skill in comparing complex marks. Therefore, it may be helpful to juries to be given information regarding both the level of difficulty of the specific mark in a given case, and the level of skill of the examiner testifying to it.

The utility metric can be useful on both counts. By applying the metric to any mark on which one is going to testify, the examiner can provide the jury with information regarding the objective difficulty of the mark. And if annual proficiency tests have been designed to test examiners at differing levels of difficulty, as suggested above, the examiner can furthermore reassure the jury that he has passed a proficiency test at the requisite difficulty level to be competent to have examined the mark at issue in a particular case.

## *Quality thresholds may be established for limiting Identifications, requiring additional QA measures, and reducing QA measures.*

Historically, the results of friction ridge examinations have fallen into one of two bins: Identification and Exclusion. An Identification indicates that the mark in question was made by the individual compared, while an Exclusion indicates that it was not. The third category, Inconclusive, was reserved for cases where a determination could not be made. This is the way things have been done at least since 1979, when the International Association for Identification (IAI) passed Resolution VII[[11]](#footnote-11), expressly forbidding its members to testify to probabilistic conclusions (Champod 1995). All identifications were, in the words of the Resolution, “positive” and to the exclusion of all others.

During this time, it was widely held within the discipline that the only valid reason for an Inconclusive decision was that better standards were needed. It was thought that if one had appropriate exemplars, the competent examiner should be able to conclusively determine whether the mark did, or did not, originate from a particular individual. If the problem was with the quality of the mark, then that mark was clearly not suitable and should not have been designated as such in the first place.

This philosophy created a culture in which all identifications were treated as equal. A mark of very high clarity that was identified was given the same label (Identification) as a mark of marginal quality. Once the “Identification” threshold was crossed, all Identifications carried the same amount of weight, were reported the same way, and were presented in court the same way. This practice created two phenomena: First, examiners were forced to claim the same amount of confidence in the Identification to the poor quality mark as they had in the high quality mark – a situation that was very uncomfortable for many and that seemed to defy logic and common sense. Second, examiners were often compelled to identify marks that they probably should not have been identifying. Poor quality marks where some detail could be found in common with the corresponding print were being identified because there was institutionalized pressure to reach a definitive conclusion and the mark could not be excluded.

Now, after the release of the NAS Report, with the advent of probabilistic models, and with the IAI rescinding their prohibition on probabilistic conclusions, there is a growing trend toward creating more bins of conclusions, or shades of grey.[[12]](#footnote-12)

Not all identifications are created equal. Some carry more information than others, and therefore, more confidence. Similarly, there is sometimes not enough information present that an Identification decision is warranted, yet the fact that information was found in common is still probative. Some agencies are beginning to embrace this distinction, creating sub-categories of Inconclusive. For instance, the Las Vegas Metropolitan Police Department (LVMPD) has an allowable conclusion of “Cannot Exclude”, which indicates that detail was found in agreement between the mark and the print, but not enough to rise to the threshold of an Identification. This conclusion is in keeping with the third prong of Locard’s 1914 recommendation (Champod 1995) that:

1. if more than 12 concurring points are present and the fingerprint is sharp, the certainty of identity is beyond debate
2. if 8 to 12 concurring points are involved, then the case is borderline and the certainty of identity will depend on:
3. the sharpness of the fingerprint
4. the rarity of its type
5. the presence of the center of the figure and the triangle in the exploitable part of the print
6. the presence of pores
7. the perfect and obvious identity regarding the width of the papillary ridges and valleys, the direction of the lines, and the angular value of the bifurcations
8. if a limited number of characteristic points are present, the fingerprints cannot provide certainty for an identification, but only a presumption proportional to the number of points available and their clarity.

While a conclusion such as “Cannot Exclude” does not incorporate a numerical proportion[[13]](#footnote-13), it does at least convey some presumptive information to the jury. This information is typically lost entirely under the currently used practice, where the results of such comparisons would either be reported as Inconclusive, and would likely never make it into a courtroom for further explanation; or would be reported as an Identification, which would be overstating the strength of the evidence.

With the creation and validation of a utility metric, these categories can have an objective measure behind them. A threshold of mark quality may be determined below which an Identification decision is not warranted, but a softer conclusion, such as Cannot Exclude, may reasonably be offered.

In a similar vein, Identifications of candidates located through the use of an Automated Fingerprint Identification System (AFIS) are treated exactly the same as those made to named suspects developed through investigation. Dror et al (2012) have demonstrated that a risk of erroneous conclusions due to bias created by list position may exist with the use of AFIS; while Dror and Mnookin (2010) have outlined reasons to exercise caution when making identifications using AFIS; and numerous sources (see, e.g. (United States Department of Justice and Office of the Inspector General - Oversight and Review Division 2006; Langenburg 2012; Neumann 2012; Lennard 2013)) have pointed out the statistical dangers of making an Identification decision based upon selecting a candidate out of the large pool provided by an AFIS.

In light of these concerns, it may be prudent to establish a quality threshold below which Identifications made using AFIS must be subjected to additional QA or documentation requirements. However, no experimentation has currently been done to establish if a greater number of errors are made due to the use of AFIS, or whether additional QA or documentation procedures would ameliorate the problem, if one exists.

On the other end of the spectrum, there are marks of exceptional clarity that do not require such close scrutiny. It may be reasonable, then, to set a threshold of high quality, beyond which abbreviated documentation procedures are followed. This could save examiners time in applying uniform documentation policies across the board, even on marks that clearly do not require it. Other QA policies may be relaxed in the case of these high quality marks, as well. Furthermore, the level of scrutiny applied to these high quality marks in court may be reasonably expected to be different than with poor quality marks that score lower on the quality metric scale.

It is recognized, of course, that none of these decisions exist within a vacuum. While there are general ideas about what “should” be enough to reach a definitive conclusion, or to use AFIS without additional QA measures, or even to keep a mark for comparison at all, in reality these decisions are colored by the agency culture in which the examiner is operating. So while it is true that “not all identifications are created equal” in the sense that some have higher quantity and quality of information, it may also be said that “not all agencies are created equal”. What may be an acceptable practice in one agency culture may be considered shockingly reckless by another.

Put back into an Analysis perspective, these cultural differences may contribute heavily to observed differences in value decisions. A large agency that deals predominantly with violent crimes, such as the FBI, may have a significantly different threshold for value decisions than a small local agency that sees a high volume of property crimes, which may again differ from a culture such as that of the Dutch experts reported by Langenburg (2012), who place a premium on consistency of minutiae selection, but consequently allow many marks to go uncompared that would be considered by other agencies.

Even within the same agency, there will be some differences in perspective that will allow two examiners (or the same examiner at two different times) to reach different value decisions on the same mark (and thus, the same available information). For example, an examiner who has recently had a false positive identification discovered in his work may become for a time much more conservative, and let many marks go that he would have consistently called “of value” prior to the error being discovered.

These ideas can be properly explored using a decision theory framework, as described by Biedermann et al (2008). A chapter of the ensuing work will be dedicated to putting decision theory in the proper context in the development of the utility metric.

## *The information used to render a conclusion should be conveyed to a jury in a way that is comprehensible and gives an idea of the significance of the finding, along with its limitations.*

Under the prevalent two-decision model of friction ridge conclusions (Identification v. Exclusion), testifying to one’s conclusion is relatively straightforward, if somewhat lacking in transparency. However, with the emergence of more nuanced decision-making, friction ridge examiners will need to re-think the way they present their findings to a jury (Cole 2011; Cole 2014).

With increased documentation, it will be easier to recall and report the thought processes that went into the decision, as well as to increase transparency about the process. However, it is unknown what affect this information will have on a jury. Will this additional information help to clarify matters for them, or only serve to confuse them?[[14]](#footnote-14) Furthermore, is the same level of detail in testimony necessary for all marks, or can an abbreviated explanation suffice in cases where the mark is of high quality?

The creation and validation of a suitable metric of mark quality can once again aid in establishing thresholds. Above a certain level, the abbreviated explanations may be more appropriate. Furthermore, the metric itself may be of assistance to the jury. If an examiner puts any mark on which he is going to testify through the metric, then the score may be presented to the jury as part of the testimony. The examiner may give the jury an idea of how much concern or care is warranted in the interpretation of the results, based upon the quality of the mark. This may aid the jury greatly in determining how much weight to attribute to the evidence presented by the friction ridge examiner. In the case of a poor quality mark, the examiner may be able to demonstrate why he reached his conclusion sufficiently to allay any concerns the jury may have had – but by presenting the quality information to them, they will at least know that they ought to have had concerns to begin with.

With the recent push for greater transparency in testimony and greater modesty in conclusions, it seems desirable to present the jury with as much information as possible regarding how conclusions were arrived at, how strong the evidence is, and to give conclusions that do not overstate the significance of the evidence. However, research is needed to establish appropriate language to achieve these goals without simply bogging the jury down with unneeded and confusing technical information. As Lennard puts it (2013):

“We justifiably strive for scientifically defensible means of presenting our evidence, but we arguably need an approach that better meets the needs of the court, including the jury, the judge, and the other legal practitioners involved in the process.”

In other words, just because greater technical accuracy satisfies good scientific principle, it doesn’t necessarily follow that it makes the information more understandable to a jury. We must remember that explaining our findings to the jury in a manner that aids them to understand is the ultimate goal of testimony, and that most of them are not scientists themselves (Eldridge 2012).

As it is currently unknown if information regarding quality scores, weight attributed to rarity, significance of findings, and even explanations of the limitations of those findings will tend to enlighten or confuse the jury, research in this area is clearly needed (Langenburg 2012). Happily, some research on juror perceptions (Holmgren and Fordham 2011) tends to show that jurors appreciate experts’ attempts at humility, admitting errors, and appearing human, which seems to indicate that the inclusion of greater transparency and more modest claims would be welcomed and would not lead jurors to lose faith in the experts’ expertise, as some practitioners have feared.

# Objectives and Work Plan

It is the objective of this research to develop and validate a standardized metric for the utility of marks that will incorporate both the rarity of features and the presence of useful target groups, as appropriate. This metric will be used as a tool to explore decision-making in the Analysis phase of the ACE-V process and to develop reasonable thresholds for several categories of mark complexity. Once established, these thresholds may be used to inform policy, create more standardized and nuanced data sets for proficiency tests and research, and offer more transparent testimony in a court of law. Training programs will also be developed and offered at national educational conferences to demonstrate the benefits and use of the metric.

In order to meet this objective, a series of research experiments will be carried out to categorize and appropriately weight the various components that go into the Analysis decision. Once the metric has been designed, further experiments will be conducted to test the performance of the metric against inter-examiner variability; against other existing quality metrics; and to test its ability to accurately predict the level of comparison value associated with a mark, by testing against ultimate outcomes. These experiments and validations will be outlined below.

## *Task One: The Pilot Experiment*

Prior to the development of the metric, data must be collected from a small group of examiners to gauge how the decision-making process during Analysis is done so that the metric can be “taught” to model this process. While the number of examiners used during this phase may be small, the amount of data collected from each will be enormous. In order to determine *which* pieces of data figure most prominently in the Analysis decision, examiners must be asked to focus explicitly on all potential areas so that it can be determined which contribute most (or at all) to the overall Analysis decision.

Since the goal of the research is to develop an objective tool for evaluating the utility, or usefulness, of a mark, it is desirable to draw data from a large sample of marks with which to teach the model how to behave. For this reason, rather than selecting only a few marks and having every participant Analyze them, this experiment will make use of a sampling scheme more similar to that employed by the FBI/Noblis study (Ulery *et al.* 2011) in which a great many marks were utilized, but only a small sub-set of examiners looked at each one. This allows for data to be collected across a greater breadth of marks of different qualities, which is more helpful in designing this type of tool.

To meet these ends, the pilot study will consist of ten hand-picked examiners who are known personally, or by reputation, to the author. These individuals will be selected for their skill as examiners, their ability to articulate their thought processes, and their likelihood to actually complete the tasks being requested of them. Because the Analysis phase is inherently lacking in “ground truth” (that is, there is no objective right or wrong answer to the value of a mark), these ten Pilot Experiment participants will be utilized to define “ground truth” for each mark via their consensus opinion on the marks with which they are presented. For this reason, it is vital that all ten participants fully complete the experimental tasks, and they will not be selected to participate without a firm commitment that they are able to do so.

In order to keep the experimental task manageable for these participants, but still maximize the amount of data collected, the experiment will be structured as follows. One hundred marks will be selected for use in the experiment. These marks will be roughly pre-categorized by the investigators into one of three categories: low, medium, and high difficulty. These three categories will be represented in the proportion 10:35:55. Each participant will be asked to Analyze, using the procedure described below, 30 marks. If 15-16 minutes are spent on each mark, this will represent an approximately 8-hour time commitment from each participant. The marks will be randomly assigned to each participant such that each participant sees marks of varying difficulty in the proportion 3:5+:10+ where the + indicates that *at least* that number of marks from the category will be viewed and such that each of the 100 marks in the experiment will be viewed by exactly 3 participants.

As a common problem in previous examiner-driven experimentation seems to have been a lack of consistency in how the participants follow the directions, understand the concepts, and interpret terminology, an important component of the pilot experiment will be training. A User Manual will be developed for key areas and given to the participants *prior to* the participants being asked to embark on the experimental tasks. It is hoped that this manual will allow for the greatest possible degree of consistency between participants. Throughout the Experiment One description, reference will be made to concepts that will be included in the User Manual.

In the pilot experiment, a customized Pianos software platform will be used to capture, as much as possible, the participants’ thought processes as they move through the Analysis process. Once the User Manual has been read, participants will begin the Analysis task on marks selected for varying degrees of difficulty, clarity, number of target groups, presence of rare compound minutiae, etc.

As the participant is shown each experimental mark, she will be complete her Analysis process using Pianos and will be required to answer specific questions before proceeding to the next mark.

Level One

The participants will be asked to consider Level 1 detail. They will be asked to select a pattern type (from among a list of choices provided by Pianos). They will also have the opportunity to select more than one pattern type, if they cannot be certain of only one. Once a pattern type is selected, there are, as noted previously, two questions of interest regarding the choice:

* How distinctive is it?
  + The participant will be given a checkbox that can be checked if they feel that there is something distinctive about the pattern that makes it stand out for them. The meaning and interpretation of this question, along with image-and-text examples, will be provided within the User Manual.
* How confident am I?
  + If, during completion of the “Quality of level 1, 2 and 3 details” question in the Conclusion section of Pianos, the participant selects the “Indistinct” radio button, he will be tacitly indicating that he is lacking some confidence regarding pattern type.

Level Two

The information of interest to collect regarding Level 2 detail will center on minutiae selection and target groups.

*Minutiae*

As the participant proceeds with the Analysis of the mark, he will be requested to mark each minutia that is visible to him in the mark. Following the symbols established in previous versions of Pianos, participants will be able to mark each minutia using one of four symbols: circle, square, triangle, and diamond.

For three of the symbols, the previous meanings will be attached (i.e. circle for ridge ending, square for bifurcation, and triangle to indicate uncertainty about minutia type). The fourth symbol, the diamond, will be chosen to indicate that the participant is less than certain that the minutia truly exists at all. This would correspond to a Red, or possibly Yellow (depending on the participant’s particular level of confidence) designation under the GYRO color scheme (Langenburg and Champod 2011). The definitions of all four symbols will be explicitly laid out in the User Manual provided to all participants. In this way, information about the *confidence* of the examiner in each minutia will be captured.

Beyond confidence, it is also useful to understand the *weight* that the examiner has assigned to minutiae. [[15]](#footnote-15) While minutiae that are assigned low weight are of little consequence since they are unlikely to contribute to the conclusion (and thus, unlikely to contribute to an error), minutiae that are assigned a high weight could be instrumental in the final outcome of a comparison. As weights should properly be determined during the Analysis phase, participants should be able to articulate which, if any, minutiae they are assigning additional weight to (typically due to extraordinary clarity, or high specificity).

Thus, in the Conclusion section of the Pianos interface, participants will be asked if they gave additional weight to any particular minutiae in their Analysis. If they select “Yes”, they will be asked to list any minutia(e) to which they assigned extra weight. They will respond by giving the number(s) of the minutia(e) of interest.

*Combined Minutiae*

In addition to selecting each minutia, participants will be requested to indicate any combined minutiae that they note. Combined minutiae will be covered with the User Manual and instruction will be given on how to decide if a grouping of minutiae qualifies as a “combined minutia”. A new tool in Pianos will be developed that allows for the grouping of multiple minutiae into a combined minutia by a series of clicks. Instruction on this tool will also be given in the User Manual. Each single minutia and combined minutia will be assigned a sequential integer by Pianos for ease of reference.

Participants will not be asked to name the combined minutiae (i.e. to declare it as a “lake” or “enclosure”, etc). Prior efforts at categorizing combined minutiae have demonstrated that examiners are not using standardized terminology for combined minutiae, and attempts to force them to do so seem to just create more confusion. For the purposes of this study, it is far more important to know that there *were* combined minutiae noted than to concern ourselves with what the participants wanted to call them.

*Target Groups*

Participants will also be requested to consider any target groups that they note within the mark and that they would use in the Analysis decision, or that they would anticipate using during a Comparison. A discussion on what constitutes a target group will be included in the User Manual. A new tool in Pianos will be utilized to group individual minutiae (or combined minutiae) into target groups by a series of clicks. It is recognized that in some cases, a combined minutia and a selected target group may be one and the same. Each designated target group will be assigned a color by Pianos for easy visual reference.

While a distinctive combined minutia may often serve as an excellent target group on its own, there will be times when an examiner may combine a combined minutia with other minutiae to form their target group. There are also times when a lack of good combined minutiae (or good combined minutiae in a good area of the mark) will force an examiner to select a target group that does not contain any combined minutiae. For instance, in a murky or low specificity mark, an initial target group may simply be a ridge ending three ridges off the core. For this reason, it is important to track combined minutiae and target groups separately, although the two may frequently overlap.

As with Level One Detail, there are fundamental questions that are of interest regarding the selection of Level Two Details:

* How distinctive is it?
  + This question will be tacitly dealt with by the weight question asked in the Conclusion section. Any minutiae that were noted as being given extra weight will be assumed to be in some way distinctive.
* How confident am I?
  + This has already been captured by the use of one of the four symbols (circle, square, triangle, diamond) during feature selection. If the participant felt less than confident in a feature, they would have used the triangle or diamond symbols.

Level Three

Previous research has demonstrated that examiners are terribly inconsistent, not only in their use of Level 3 detail, but even in their definitions of what constitutes a Level 3 detail (Anthonioz *et al.* 2008). SWGFAST defines Level 3 detail as “ridge structures (edge shapes and pores), and their relative arrangements” and further notes that “[c]reases, scars, warts, incipient ridges, and other features may be reflected in all three levels of details” (Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST) 2013), yet many examiners would classify creases, scars, and incipient ridges as Level 3 detail. Ashbaugh defines it as “…small shapes on the ridge, the relative location of pores, and the small details contained in accidental damage to the friction ridges.” He gives as specific examples, “the alignment or misalignment of individual ridge units, ridge unit shape, ridge unit thickness or thinness, and relative pore location […] twisting or puckering of the ridge units caused by a scar.” He also shows a figure to demonstrate Level 3 detail (Figure 4.26 in the reference book) in which the descriptive text notes “[t]he ridges fluctuate in thickness, from fat ridge units to thin. Some are misaligned while others flow smoothly.” (Ashbaugh 1996). There are features (e.g. pores and edge shapes) that most examiners would agree are Level 3 detail, yet these are rarely used in actual casework and research on the reproducibility of the recording of pores and edge details has shown that they are not particularly robust to consistent recording (Richmond 2004; Anthonioz *et al.* 2011).

For these reasons, this research will concern itself less with how the participants define Level 3 Detail, and more with whether or not they recognized and utilized it at all.

As the participants complete the “Quality of level 1, 2, and 3 details” question in the Conclusion section of Pianos, they will indicate how much Level 3 detail they noted/considered. If they select “Abundant”, they will be further prompted to check boxes corresponding to types of Level 3 detail that they noted *and assigned weight to* during their analysis (for example, if they saw pores, but do not routinely use pores in their Analysis or Comparison of a mark – so the pores really didn’t figure into their decision – they should not select “Pores”). The checkboxes of types of Level 3 detail will be populated with types of Level 3 detail that were covered in detail in the User Manual.

Clarity

Participants will be queried on their interpretation of the clarity in the mark. They will be asked to annotate the quality of the mark using the Quality tool in Pianos (which allows for polygon-based annotation of the high, medium, and low areas of clarity). They will then be asked in the Conclusion section:

* Does the clarity of this mark impact your Analysis decision?
  + The clarity of this mark adds weight to my desire to keep the mark
  + The clarity of this mark adds weight to my desire to discard the mark
  + The clarity of this mark does not impact my Analysis decision either way

Distortion

Participants will also be queried on their interpretation of the distortion present in the mark. As with minutiae nomenclature and Level 3 Detail interpretation, it seems that examiners lack consistency in naming and interpreting distortion factors. For this reason, they will not be asked to detail what specific distortion factors they saw, but only to give a general sense of how any distortion present impacted their Analysis of the mark. They will be asked in the Conclusion section:

* Overall, how much has distortion impacted the mark?
  + High Distortion – There are major distortion factors that severely impact the interpretation of the mark
  + Medium Distortion – There are areas of significant distortion, but also areas of low distortion with usable features
  + Low Distortion – There are minor distortion factors that are easily explained or worked through
  + No Distortion – There are no distortion factors that impact the interpretation of the mark

Analysis Decision

In the Conclusion section, participants will be asked for their final Analysis decisions. They will be presented with groups of Analysis thresholds and will be asked to make a judgment in each group. Because the value, or usefulness, of a mark may be given many different labels depending on context, this research will attempt to capture the participants’ feelings about each of the different value groups. It is hoped that the eventual metric that will be designed will be capable of sorting any mark into each of the categories, which may be overlapping, not mutually exclusive. In this way, the metric will provide a flexibility that allows it to evaluate the value of a mark for several different situations. Information on each category and the meaning of its response options will be given in the User Manual.

The categories will be as follows:

* Value Decision
  + Of Value
  + Not of Value
* Complexity
  + Complex
  + Non-Complex
* AFIS Quality
  + AFIS Quality
  + Not AFIS Quality
  + AFIS Quality, but Additional Caution Recommended
* Difficulty
  + Easy
  + Medium
  + Difficult
* Appropriateness for a Conclusive Decision
  + A conclusive decision (Identification, Exclusion) would be appropriate
  + There is enough information here to do a comparison and possibly find information in common (or not in common), so the mark has some probative or investigative value, but there is not enough here to warrant a definitive conclusion[[16]](#footnote-16)
  + This mark lacks sufficient information to proceed with a comparison

## *Task Two: The Large Scale Experiment and Development of the Utility Metric*

As discussed above, current attempts to develop a quality metric have focused mainly on clarity of the mark, or on number of minutiae. None have incorporated rarity of features selected, or presence of searchable target groups. This research will attempt to develop such a metric, and validate it against users such that it can be used to achieve repeatable and reproducible results. Furthermore, most quality metrics currently in development seem to focus solely on the question of value. This metric will attempt to establish utility thresholds for numerous uses, to include: value v. no value; complex v. non-complex; AFIS quality v. AFIS quality, but proceed with caution; standard quality v. high quality; and value for Definitive Conclusion v. Definitive Conclusion not warranted. A graphic representing the possible relationship between some of these categories is presented in Appendix C.

This development and validation will be a multiple-step process, beginning with a survey of subjective, examiner-driven Analysis of marks, and moving to a more automated, computer-driven analysis based upon data gleaned from the early stages of the research.

1. Phase One – Using the Pilot Experiment Data to Design the Metric

As described above, the Pilot Experiment will be used to collect data from a group of examiners that may be used to inform the design of the metric. During the Pilot Experiment, participants will have been asked to evaluate numerous marks of varying quality and answer a number of questions about each that describe the information visible in the mark and how the participant sees the value of that information. The participants will also have been asked to provide an Analysis decision in a number of subjective decision threshold categories.

The results of the Pilot Experiment will be examined for multiple purposes. First, the subjective categories will be examined to test for consensus between examiners on the overall quality of the marks. Pianos will also be used to compute consensus quality maps and distance from each examiner to the consensus on clarity and minutiae for each mark, as described in (Neumann *et al.* 2013). Second, the individual observations made will be examined to determine which contribute more heavily to the examiners’ overall decisions. Using Bayesnet software, such as Hugin, the data from the Pilot Experiment will be put into the model, which will learn the factors of the examiners’ Analyses that carry the greatest weight. Each category of observation will thus be assigned mathematical weight, such that use of the metric can produce a numerical score. Also factoring into the score assigned to each mark will be objective rarity data on the features contained in the marks, taken from previously conducted feature distribution studies (e.g. (Champod and Margot 1996; Gutiérrez-Redomero *et al.* 2011; Gutiérrez-Redomero *et al.* 2012)). The distributions of scores will then be examined to see if they fall out at reasonable thresholds that match up against the utility categories described in the Pilot Experiment.

*AFIS Suitability*

As a subset of Phase One, the development of the AFIS Quality/Not AFIS Quality/AFIS Quality, but Proceed with Caution categories may require a separate analysis. Given the particular pitfalls of AFIS searching (namely, the high risk of a coincidental match in a large database), it seems that there are four situations in which the risk of a false identification made using AFIS is greater:

1. A low minutiae count was entered

This can be determined by looking at the actual number of minutiae annotated by the participants and determining where the threshold for “low” lies. This then becomes a simple binary response – either the minutiae count was “low” or it was not.

1. Minutiae entered were at areas of lower specificity, such as deltas and the outflows of loops

This may need to be determined manually for each mark by looking at the areas where the participant annotated minutiae and flagging those marks for which the statement is true.

1. No rare features or target groups are selected

The data collected by Pianos will capture any target groups that were annotated. Likewise, Pianos will already be utilizing minutiae frequency data to determine if rare features were selected.

1. Clarity of the mark is low

This could be calculated by using the data provided by participants using the quality mapping tool in Pianos. Much as proposed by Langenburg (2012), an overall quality score for the image can be calculated by looking at the relative number of clear pixels versus poor pixels as annotated by participants. To take it a step further, one could see how many minutiae were marked in the good areas to appropriately weight the clarity of the image (if there is a large area with high clarity, yet most of the selected minutiae lie within a small area of poor clarity, the high clarity area is not contributing materially to the quality of the mark for AFIS purposes).

These risk situations can be calculated based upon data collected by Pianos, but their weighting may be different from the analysis being done for the other categories. Here, we are less concerned with the weight being put on particular information *by the examiner* and more concerned with weighting the inherent risk that is brought to the AFIS search by each contributing factor. Once calculated, these four risk factors can be weighted using Bayesnet to understand how each one contributes to the risk of a coincidental match in AFIS. This will inform the metric’s assignment of a mark to one of the three AFIS categories.

1. Phase Two – The Large-Scale Experiment (LSX)

The basic structure of the Pilot Experiment will now be repeated in the Large-Scale Experiment (LSX). The specific criteria for the LSX will be adjusted in light of data collected during the Pilot Experiment, or any flaws in experimental design that became apparent during the execution of the Pilot Experiment. The LSX will utilize as many participants as can be enticed to participate in order to represent the largest cross-section of the latent print examiner population possible.

Once the LSX has been completed, the data will be analyzed, as described in Phase One. The Bayesnet will be further taught in light of the new data, as well as comparing the data from the LSX against that from the Pilot Experiment to see if they are roughly in line. The newly-adjusted model developed using the data from the LSX will be carried forward into the remaining Tasks.

## *Task Three: The Documentation Experiment*

While increased documentation is often called for as an aid to transparency, accountability, and memory (see, e.g, (Langenburg and Champod 2011; Expert Working Group on Human Factors in Latent Print Analysis 2012)), it is seldom investigated as a tool to increase accuracy. As previously discussed, the very act of creating documentation may lead to an increase in accuracy. It is this question that will be explored in this experiment, making use of the utility metric to provide marks at varying levels of difficulty.

As relates to documentation of Analysis, there are three questions of interest:

1. What is being considered during Analysis and how is it weighted?
2. Does the act of documentation actually contribute to better outcomes (in this case, greater consistency in Analysis decisions)?
3. Some prior research efforts (see, e.g. (Langenburg *et al.* 2012)) have explored documentation, but it is not clear if the documentation was done contemporaneously with the observations to actually provide aid to a decision, or whether it was added on after the decisions were already made.

To try to gain insight into these three questions, participants will be broken into two experimental groups. Each group will be given marks of varying difficulty, selected from four categories (each group will be presented with marks from each of the four categories). The four categories represented will be:

1. Marks of high quality with low specificity
2. Marks of low quality with high specificity
3. Marks of low quality *and* low specificity
4. Marks of high quality and high specificity (these marks will be used as a control and few of them will be utilized)

The purpose of using marks from these four categories is to explore the interplay between clarity and rarity. If one feature is very clear and definitely present, but has very low specificity (e.g. a ridge ending) while another feature is in a murky, distorted area of the mark, but is highly distinctive, which will be assigned higher weight? Can one compensate for the other? In other words, can a low clarity mark meet the value threshold if it contains highly discriminating minutiae? Likewise, will a high clarity mark fail to meet the value threshold if it contains only common, low specificity minutiae without any “anchors” (i.e. target groups or focal points) to assist them?

The two experimental groups into which the participants will be separated will be:

1. Control Group (no annotation)

The control group will be shown marks and will be asked to consider features in several categories (which will have been determined in accordance with the results of Experiments One and Two as categories of greatest impact to Analysis). This group will work with high-quality photographic images on paper outside of Pianos (i.e. they will be given *no tools* to annotate the mark). They will further be requested not to annotate the mark in any way as they consider it. They will simply look at the mark. At the end of their Analysis, they will be asked to give their Analysis decisions in the analysis decision threshold categories that have been previously used in Experiments One and Two.

1. Experimental Group (full annotation)

Experimental Group One will be shown marks and will have all the same tools (and all the same questions to answer) as in Experiment Two.

Once the annotation of one mark is complete, the participant will be shown the next mark. Only after the annotation of all marks is complete will Analysis decisions be made. The participants will then be shown all the same marks again, in random order. Their annotations will all be visible to them. As each mark is presented, the participant will be asked to give their Analysis decisions in the analysis decision threshold categories that have been previously used in Experiments One and Two.

The separation of annotation and Analysis decision is necessary because it allows examination of the information of interest in Question Three, outlined above. If the participants are forced to complete their documentation *prior to* making an Analysis decision, and then are separated from the mark for a time and come back to the mark at a later time to make the Analysis decision based upon the annotations they previously made, it should be possible to see the effect of the annotations, if any, upon the ultimate Analysis decisions, in comparison to the control group, which did not annotate and did not have the separation between initial observations and Analysis decisions. This should help to separate the effects of documenting prior to a decision versus making the decision and simply justifying it through documentation after the fact.

This experiment will be on a much smaller scale than the LSX; only around 20 marks will be used, representing a much smaller time commitment, and a smaller number of participants will be solicited.

Results will be examined for an increase of consensus on Analysis decisions with increased documentation prior to the Analysis decision. Results will also be analyzed to look for differing effects of documentation upon marks of variable quality.

## *Task Four: Validating the Metric Short Form*

By this point, it will already have been determined what factors are the most important in Analysis, and their relative weights. The metric will be capable of categorizing marks into a number of categories through the use of numerical scores generated by the model after the collection of the subjective observations of an examiner. It will no longer be necessary to quiz examiners exhaustively on every aspect of each mark. A shorthand form will suffice to encapsulate their observations in key categories, which the metric will then translate into the aforementioned scores. This shorter form will be the practical tool that can be used by laboratories, researchers, and test designers to quickly collect the basic information needed to objectively sort marks into appropriate categories.

In order to use this metric short form for various applications, it must first be tested against the decisions of real examiners to verify that it is accurately predicting Analysis decisions across the various thresholds of interest based on the information those examiners have provided.

1. Phase One – Creation of A Short Form

Using the information obtained in Experiments One and Two, a software interface will be developed that allows users to select a description of each category of data visible in the mark. The weighting for each category of data will be incorporated into the software, which will automatically generate a score based upon the information provided. This score will be translated into the appropriate utility category(ies)[[17]](#footnote-17).

The information required by the short form[[18]](#footnote-18) will comprise ten categories of visual cues to include: Anatomical Source[[19]](#footnote-19); Orientation Information[[20]](#footnote-20); Size of the Mark; Clarity; Levels One, Two, and Three Detail; Distortion; Existence of Target Groups; and Rarity of Features.

Also incorporated into this computer-based metric will be rarity data taken from previous minutiae frequency studies (Champod and Margot 1996; Gutiérrez-Redomero *et al.* 2011; Gutiérrez-Redomero *et al.* 2012). Users may be prompted to indicate specific minutiae observed, and this information then used to provide weight in the score calculation based upon the discriminating power of the features present. This data may also be handled categorically, with users prompted only to select a *bin* of rarity, based upon the quantity of features noted in various rarity categories, which the metric would then weight appropriately and take into consideration in the scoring of the mark.

1. Phase Two – Testing the Metric Short Form Against Examiners

The metric short form, once developed, will then be tested against the decisions of expert examiners. Since assessments in each value category and each category of visual cues will have already been provided by the examiners participating in Experiment Two (the LSX), there is no need to apply to a further batch of participants; the previous data can be entered into the short form for each mark by the investigators, and the outcomes provided by the tool compared to the consensus decisions for each mark, as provided by the respondents in Experiment Two.

A mock-up of the questionnaire is provided in Appendix D to give an idea of the categories and response options that may be provided and may be adjusted prior to use.

Results will be examined to determine if the metric is successfully predicting the utility judgments of skilled examiners.

1. Phase Three – Testing the Metric Against Other Quality Tools

As mentioned previously in this work, there are existing quality tools that perform analysis of marks in a lights-out fashion, giving an overall assessment of the quality of the mark. These tools focus predominantly on minutiae counts and clarity. While the tool proposed in this research will be more comprehensive, taking more data into account, and, it is hoped, will be more flexible, it is not known if the ultimate decisions provided by this tool will differ significantly from the decisions provided by the existing tools.

In this phase of the research, a subset of the test marks used thus far in the research will be tested in available quality models (see, e.g. (Yoon, Liu, & Jain, 2012) ) to see the assessments provided by those tools. Lights out assessment results will be compared to the hybrid examiner-metric results provided by this system to see if there is a significant difference in overall outcomes, in flexibility, and in the nuances of the provided categories.[[21]](#footnote-21)

## *Task Five: Validating the Metric for Predicting Comparison Value Thresholds*

Because the metric being proposed is, in essence, an Analysis tool, the bulk of the experimentation in this research has been focused on the Analysis phase of the ACE-V process. However, Analysis is only the first step. While it is certainly desirable that a metric for assigning value to marks should be consistent, and in line with examiner assessments, if it does not accurately predict the value of marks for ultimate outcomes, it is of little practical use. For example, if the metric is labeling marks as “of value” or “low difficulty” and then examiners are consistently making erroneous identifications on those marks, then clearly the metric has mis-labeled them.

Task Five looks at the metric in the context of the entire ACE process – that is, it utilizes the metric to Analyze marks, then tests those Analysis decisions against the ultimate outcomes of comparison decisions in various categories the metric is designed to measure.

The marks selected for use in Task Five will be Analyzed using the metric, and will be assigned to categories based upon their scores. Marks will be selected for use along the full spectrum of possible Analysis designations. Using a 50:50 mix of mated and non-mated pairings in each category, participants will be asked to render conclusion decisions, which will then be checked to see if they are in synch with the predictions made by the metric. Since this experiment is testing the performance of the metric, marks will be pre-assigned by the metric to value categories and participants will not be asked to provide an Analysis decision; they will proceed directly to the comparison exercise. In order to keep the participants from being biased in their Evaluation decisions based upon the categories assigned by the metric, they will be kept blind to this information (i.e. participants will be told only that the marks are “of value”; they will not be given any information on the various categories to which each mark has been assigned by the metric).

A large AFIS will be used to locate close non-matches for the non-mated pair sets. If the metric is performing correctly, it is expected that error rates will be significantly higher in categories such as: Complex; AFIS Quality, but Additional Caution Recommended; Difficult; and Not Appropriate for a Conclusive Decision.

Here it is necessary to digress briefly. While the experiment outlined above represents an ideal for this point in the project, there may be some practical limitations to its execution. First, it may well prove to be the case that the experimentation is not able to elucidate how well the metric predicts outcomes in relation to ground truth for two reasons: first, because typically there are so few false positive errors made there is not a large data set from which to glean any useful information; and second, because the cause of an error in the Evaluation decision may not always be attributable to the mark. If the problem lies with the print, or with the correspondence between the two (e.g. the mark and the print both have relatively high quantity/quality of features, but the specific area needed to correspond between the two is unclear or incompletely recorded in one of the two), then the predictions of the metric will have little, if anything, to do with the accuracy of the ultimate outcome.

One way to address the first limitation could be to focus on the erroneous exclusion decisions, rather than the erroneous identification decisions. These tend to be much more plentiful and would provide a much larger data set; however, they still have the advantage of testing the metric because it is designed to distinguish marks that should not be used for *any* definitive conclusion, not only an identification.

Another way to address the first limitation could be by using examiner agreement as a proxy for ground truth. In other words, if the metric is intended to be a sort of red flag, or early warning system, that a mark is on shaky ground and should not be relied upon for certain uses or certain types of conclusions, it is not so much the rate of errors one is concerned with as the *chance* that an error may occur. Naturally, it is expected that there is a much higher chance of errors occurring using marks that have a higher variability between examiner conclusions. It may be said that in any given panel of examiners analyzing a mark, if they all agree, the mark is more reliable than if there is a split in decisions. Thus, if there is disagreement between examiners, then the mark is *de facto* suspect and should be treated with additional caution.

The second practical limitation of this phase of the experimentation is that, after a great deal of time and examiner participation spent on the early experiments, it requires that once again, an experimental design be launched, participants solicited, and a special Pianos interface be designed. This represents another monumental undertaking late in the research. One possible solution to save resources could be to utilize the data that has already been collected by Neumann et al (2013). In this research, which was conducted using Pianos, 15 mark and print pairings have already been presented to approximately 160 examiners and evaluation decisions have been made.

It would be possible to use the short form on each of the marks from the NIJ sufficiency project, get a score from the metric, and then test the categorizations provided by the metric against the ultimate outcomes recorded by the participants in the NIJ sufficiency study.

This approach has the advantage of not requiring new participants and new experimentation. However, it has limitations of its own. First, the NIJ pairings were not equally split; there were 12 true matches and 3 non-matched pairs. Thus, there were few opportunities for false positive errors. Second, all of the marks in the study were chosen to be borderline, or difficult marks. This will test the metric nicely along that range of the quality continuum, but leaves the other portions of the spectrum untested.

As discussed above, some of these limitations could be addressed by the measures proposed, but the narrowness of the range of quality would remain.

This portion of the research, as previously mentioned, occurs late in the project and a great deal may have changed, or been learned, by that point that will help to guide the direction this experiment may take. In the meantime, the author is open to guidance and recommendations concerning the best way to direct this phase of the validation of the metric.

If the full range of the categorizations provided by the metric is tested, the following describes the expectations in different categories, if the metric is successfully assigning utility categories for the marks.

*Difficulty Thresholds*

The difficulty threshold category is broken into three levels (easy, medium, and difficult). In the “easy” category, it would be expected that both false positive and false negative rates would be zero, or near-zero (or that examiner conclusions would be unanimous or near-unanimous). Naturally, it would be expected that the error rates (and examiner variability) would be somewhat elevated at the medium difficulty level, and further elevated at the difficult difficulty level. It would be interesting to see if the false positive rate for either of these categories is roughly in line with the error rate established by the Noblis/FBI black box study (Ulery *et al.* 2011).

*Complexity*

As with the difficulty levels, it is expected that marks designated by the metric as “complex” will show a higher false positive and false negative rate than their “non-complex” counterparts. If there is a clear demarcation between the two, it will justify the existence of a “complex” label that requires additional QA measures, such as additional documentation, additional review, etc.

*Appropriateness for a Conclusive Decision*

Under the current two-conclusion model, the only acceptable conclusions for a comparison of mark and print are the definitive Identification, or Exclusion. Any comparison that cannot result in one of these two conclusions is placed into the all-encompassing bin of “Inconclusive”.

However, there are no gradations in the inconclusive bin. Therefore, cases in which detail is found in agreement, but is not enough to rise to the threshold of an identification, are relegated to the catch-all “Inconclusive” and their potential probative value is lost. On the other hand, marks that are of poor quality may be receiving the “Identification” label when some information is found in agreement with a print, although their quality truly does not support such a definitive statement and a more temperate conclusion would be more appropriate. [[22]](#footnote-22)

Interestingly, the same argument can be made at the other end of the spectrum, near the exclusion decision. There are comparisons where some apparent dissimilarities may be noted, but the clarity of the mark is such that there is not sufficient confidence in the nature of the dissimilarities to warrant a conclusion of exclusion. The author is not aware of any labs that have designated a label for this situation, probably because no labs are comfortable reporting this situation. Following thinking such as the One Dissimilarity Doctrine (Thornton 1977), many labs are more comfortable with reporting an exclusion when an apparent discrepancy is noted. Given the high levels of false exclusions reported in all existing research on error rates (see, e.g. (Langenburg 2009; Ulery *et al.* 2011; Langenburg *et al.* 2012)), this is a practice that could likely benefit from scrutiny.

Gradually, more pressure is coming to bear toward more transparency in conclusions and testimony and more modesty in claims, which would include both presenting probative information that may not rise to the Identification threshold, and also stopping short of that threshold when the strength of the evidence does not support it. The same can be said of the exclusion end of the spectrum – there should be a movement toward stopping short of an exclusion decision when there is not sufficient strength of evidence to support it.

While it could be argued that this is a comparison decision that should be made during the Evaluation stage of the process, this would present a logical inconsistency. If an examiner is unable to render a definitive conclusion (i.e. identification or exclusion), there can be only two causes: a problem with the mark, or a problem with the print. If the problem is with the print, the examiner may report an Incomplete result (a subset of the Inconclusive decision), and request new exemplars that record the area that is lacking, whether in clarity or presence.

If, however, the problem is with the mark (that is, the mark is of sufficient quality for comparison, meaning that it can be searched upon and some information may be found in agreement or disagreement with a set of known prints, such that a weak association or dissociation may be made, but not of the strength that would warrant a definitive conclusion), this is information that should be known to the examiner during the Analysis phase, prior to the prints ever being seen. It may be, due to Level One information, that some particular prints could be excluded using these marks, but that there is not sufficient data contained in the marks to be able to exclude *all* marks that one may have occasion to compare. This does not change the underlying logic of the designation – the mark is of some probative value, but may not be sufficient to completely exclude a subject.

Under this logic, the decision should be made, upon the strength of the mark, and prior to beginning comparison, whether the mark is sufficient for a Conclusive Decision (whether that decision be Identification, or Exclusion).

There are two areas of data analysis that would be interesting for this question:

1. When identifications are made to marks from the Not Appropriate for a Conclusive Decision category, are a significantly higher number of false identifications being made, as compared to marks that are deemed appropriate for a conclusive decision? Likewise, when exclusions are made to marks in this category, are there a significantly higher number of false exclusions? Looked at in the context of examiner agreement, is there more variability in examiner decisions for the marks from the Not Appropriate for a Conclusive Decision category?
2. Would people make use of a softer conclusion in these cases, if one was offered to them? If participants are presented with conclusion options of “Insufficient Agreement” and “Insufficient Disagreement”, along with a description of the their use, would they make use of it? In particular, would their use coincide appropriately with the marks that belong to the Not Appropriate for a Conclusive Decision category, or would participants begin using these softer conclusions on marks that were perfectly appropriate for a conclusive decision? Of course, this question could not be examined, if the NIJ sufficiency data were used, rather than collecting data from new participants.

A further area of interest regarding the designation of Not Appropriate for a Conclusive Decision is whether these marks can be distinguished from marks that are simply “Difficult”. The first step is to see whether the metric makes a distinction between the two. While there may be overlap between the categories, does the metric assign some marks to each category exclusively? The second step is to examine the comparison outcomes surrounding marks in these two categories. Will a significantly greater number of the “Difficult” marks result in a definitive conclusion, and a *correct* definitive conclusion, than for those in the Not Appropriate category? It will be important to determine whether there is truly a distinction between these two categories. If there is not, it would suggest that once a mark has been designated as “Difficult”, it really should not be used to render definitive conclusions. On the other hand, a clear difference between the categories will support that difficult marks may be used for definitive conclusions, with appropriate QA measures in place.

*AFIS*

As noted above, the current two-conclusion model of Identification and Exclusion may often pressure examiners into declaring an Identification where the strength of the evidence does not warrant it. This problem is exacerbated in the case of AFIS, where the large database size makes the probability of making a false identification due to similarity of characteristics much higher.

As with the “Identification” bin, “AFIS Quality” is actually a continuum, a range of quality that is acceptable for entry into the AFIS system. While marks of high quality may be easily and safely identified to candidates produced via an AFIS search, marks on the other end of the spectrum should be treated with additional caution, and may warrant the use of additional QA measures, possibly even including an interdiction on identifying marks below a certain quality threshold when the corresponding print was found via AFIS search. These marks may be excellent candidates for the Insufficient Agreement conclusion option mentioned above. It may be that these marks may be used to develop a potential suspect, but should not be used for positive identification, or may only be used in conjunction with additional QA requirements.

It will be of interest to note if a significant number of errors (or significant disagreement between examiners) are seen with marks belonging to the “AFIS Quality, but Additional Caution Recommended” category as compared to the “AFIS Quality” category. If so, there may be support for the recommendation of a “caution threshold” for AFIS, below which additional QA measures should be required.[[23]](#footnote-23) Additional future research would be needed to test recommended QA measures for practical efficacy at lowering this error rate.

For instance, since AFIS only searches based upon the minutiae entered, any candidates returned have been selected on the basis of a similarity to those minutiae that were entered. Therefore, any minutiae that were *not* entered into the search are independent of the candidate list that was returned. One possible QA measure to reduce the risk of false identifications in AFIS would be to require that before any mark that falls into a demonstrated risk category may be identified, additional data that was not entered into AFIS must be found in common between the mark and the print. By requiring additional comparison beyond the data that were used to create the candidate list, independent corroboration of a putative match could be obtained.

# Originality of the Project

The ACE-V process has long operated as a black box, with researchers only recently beginning to investigate error rates and other outcome-based inquiries. It is even more recent that researchers have begun to open the black box and examine the actual decision-making processes that go into ACE‑V’s ultimate conclusions.

This research builds upon early work that has been done in investigating ACE-V decision-making (Langenburg 2012); development of transparent documentation of Analysis (Langenburg and Champod 2011); study of population frequencies of minutiae types (Gutiérrez-Redomero *et al.* 2011); and development of a measure of mark quality (Hicklin *et al.* 2011; Murch *et al.* 2012; Yoon *et al.* 2012; Hicklin *et al.* 2013; Pulsifer *et al.* 2013; Ulery *et al.* 2013).

However, none of the previous research in the literature has examined the rarity of features in the assessment of the quality of a mark. While it is acknowledged that the weight attributed to a feature has two components – certainty of presence and identity, and rarity or discriminating power – only the first component, typically measured by clarity, has been given attention in the literature. This work is the first of its kind that attempts to include information on the discriminating power of features into measurements of mark quality, and to set multiple thresholds of suitability and difficulty level based upon the entire picture of the quality, or utility, of the mark.

While many commentators have criticized the use of a dichotomous conclusion structure (i.e. Identification v. Exclusion) for friction ridge examination, there is also no existing research supporting the use of a more nuanced reporting scheme, in which a more modest claim, such as “Insufficient Agreement” may be offered. This research will pave the way in exploring this option and attempt to establish appropriate, objectively-based, thresholds for its use.

Finally, while research has suggested that list order of AFIS candidates may lead to bias errors (Dror *et al.* 2012), no research has been done to identify the marks that are more susceptible to erroneous identification through AFIS, or to suggest ways to mitigate this risk. This research will attempt to do just that.

The tools that will be developed during the course of this research, namely the utility metric, and new proposed conclusion category, are novel and represent the natural next step in the evolution of the ACE-V process.

# Conclusion

The ACE-V methodology has been the predominant process used for friction ridge comparisons for nearly two decades, but it still operates largely as a black box. Different agencies and different examiners perform the steps differently, and consequently, different conclusions may be reached.

Efforts are now being made to increase consistency, reliability, and transparency in the method. This research will contribute to this effort by examining the role of rarity of features in making sufficiency determinations with a view to developing a utility metric that can be used to set objective thresholds of quality at a number of levels. It is hoped that this metric will be embraced as a tool within the community to set QA standards and policies, determine difficulty levels of marks for proficiency testing and research, and provide more transparent information in the courtroom.

The National Institute of Justice has identified a list of High-Priority Needs for research (National Institute of Justice 2013), of which this research falls into the following subset:

* + Tools to expand the utility of or provide a quantitative measure/statistical evaluation and interpretation of forensic comparisons:
    - Pattern and impression evidence
    - Population frequencies of phenotypic traits used in identification comparisons

As a sub-objective, this research will propose the introduction of a new sub-category of the Inconclusive conclusion that will allow probative information regarding comparisons that do not rise to the threshold of Identification or Exclusion to still be presented in the courtroom, as well as encouraging examiners not to “push the envelope” on marks of poor quality for which a definitive conclusion should not be rendered. The efficacy of this conclusion will also be tested and training on its use developed.

It is hoped that the introduction of these tools will contribute materially to transparency, consistency, and objectivity in the application of the ACE-V process.

# Timeline

Timeline is approximate and may be adjusted as the needs of the research develop. Findings will be written up and submitted for publication in peer-reviewed scientific journals along the way as appropriate.

***2013-2014***

Complete mémoire.

***Summer 2014***

Visit Lausanne to defend mémoire and prepare for research.

***2014-2015***

Collect suitable marks for use in experimentation.

Recruit participants for Pilot Experiment.

Finalize Pianos format for Pilot Experiment.

Update Pianos User Manual to incorporate all necessary information.

Begin recruitment process for participants for Experiment Two (LSX).

Run Pilot Experiment.

***Summer 2015***

Visit Lausanne to work on Pilot Experiment data and begin development of metric.

Attend IAI Conference to present and recruit.

***2015-2016***

Develop first iteration of the metric from Pilot Experiment data.

Adjust Pianos as needed.

Complete recruitment of participants for Experiment Two.

Run Experiment Two.

***Summer 2016***

Visit Lausanne to work on Experiment Two data and begin adjustment of metric.

Attend IAI Conference to present and recruit.

***2016-2017***

Adjust metric as needed to incorporate data from Experiment Two.

Finalize Pianos format for Experiment Three.

Recruit participants for Experiment Three.

Run Experiment Three.

Begin development of Short Form.

***Summer 2017***

Attend IAI Conference to present.

***2017-2018***

Complete and validate Short Form (Task Four).

Validate the metric for predicting comparison value thresholds (Task Five).

***Summer 2018***

Attend IAI Conference to present.

***2018-2019***

Write up dissertation (US or Lausanne).

Defend dissertation (Lausanne).

# Bibliography

**State of Washington v. Danavian Hunter, pp. Superior Court of the State of Washington in the Count of King.**

**2010 Oregon v. Stephen Angius, pp. State of Oregon Circuit Court, Lane County.**

**Almog, J., and E. Springer (Editors), 1996 *Proceedings of the International Symposium on Fingerprint Detection and Identification, June 26-30, 1995, Ne'urim, Israel*. Israel National Police, Jerusalem.**

**Anthonioz, A., N. Egli, C. Champod, C. Neumann, R. Puch-Solis *et al.*, 2008 Level 3 Details and Their Role in Fingerprint Identification: A Survey Among Practitioners. Journal of Forensic Identification 58: 562-589.**

**Anthonioz, A., N. Egli, C. Champod, C. Neumann, R. Puch-Solis *et al.*, 2011 Investigation of the Reproductibility of Third-Level Characteristics. Journal of Forensic Identification 61: 171-192.**

**Ashbaugh, D. R., 1996 *Qualitative - Quantitative Friction Ridge Analysis*. Ridgeology Consulting Services, Hope (British Columbia Canada).**

**Ashbaugh, D. R., 1999 *Qualitative-Quantitative Friction Ridge Analysis – An Introduction to Basic and Advanced Ridgeology*. CRC Press, Boca Raton.**

**Biedermann, A., S. Bozza and F. Taroni, 2008 Decision theoretic properties of forensic identification: Underlying logic and argumentative implications. Forensic Science International 177: 120-132.**

**Champod, C., 1995 Locard, Numerical Standards and «Probable» Identification. Journal of Forensic Identification 45: 132-159.**

**Champod, C., and P. A. Margot, 1996 Computer Assisted Analysis of Minutiæ Occurrences on Fingerprints, pp. 305-318 in *Proceedings of the International Symposium on Fingerprint Detection and Identification*, edited by J. Almog and E. Springer. Israel National Police, Ne’urim, Israel, June 26-30, 1995.**

**Cole, S. A., 2011 Splitting Hairs? Evaluating ‘Split Testimony’ as an Approach to the Problem of Forensic Expert Evidence. Sydney Law Review 33: 459-485.**

**Cole, S. A., 2014 Individualization is dead, long live individualization! Reforms of reporting practices for fingerprint analysis in the United States. Law, Probability and Risk.**

**Davis, J., 1979 Thoughts on Resolution VII. Identification News: 5-6.**

**Dror, I. E., C. Champod, G. Langenburg, D. Charlton, H. Hunt *et al.*, 2011 Cognitive issues in fingerprint analysis: Inter- and intra-expert consistency and the effect of a 'target' comparison. Forensic Science International 208: 10-17.**

**Dror, I. E., and J. L. Mnookin, 2010 The use of technology in human expert domains: challenges and risks arising from the use of automated fingerprint identification systems in forensic science. Law Probability and Risk 9: 47-67.**

**Dror, I. E., A. Péron, S.-L. Hind and D. Charlton, 2005 When Emotions Get to the Better of us: The Effect of Contextual Top-Down Processing on Matching Fingerprints. Applied Cognitive Psychology 19: 799-809.**

**Dror, I. E., K. Wertheim, P. Fraser-Mackenzie and J. Walajtys, 2012 The Impact of Human–Technology Cooperation and Distributed Cognition in Forensic Science: Biasing Effects of AFIS Contextual Information on Human Experts. Journal of Forensic Sciences 57: 343-352.**

**Eldridge, H., 2012 "I am 100% certain of my conclusion" (But should the jury be certain?). Evidence Technology Magazine: 8.**

**Expert Working Group on Human Factors in Latent Print Analysis, 2012 *Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach*. U.S. Department of Commerce, National Institute of Standards and Technology, Washington DC.**

**Gutiérrez-Redomero, E., C. Alonso-Rodríguez, L. E. Hernández-Hurtado and J. L. Rodríguez-Villalba, 2011 Distribution of the Minutiae in the Fingerprints of a Sample of the Spanish Population. Forensic Science International 208: 79-90.**

**Gutiérrez-Redomero, E., N. Rivalderia, C. Alonso-Rodriguez, L. M. Martin, J. E. Dipierri *et al.*, 2012 Are there population differences in minutiae frequencies? A comparative study of two Argentinian population samples and one Spanish sample. Forensic Science International 222: 266-276.**

**Haber, R. N., and L. Haber, Experimental results of fingerprint comparison validity and reliability: A review and critical analysis. Science & Justice.**

**Hicklin, R. A., J. Buscaglia and M. A. Roberts, 2013 Assessing the Clarity of Friction Ridge Impressions. Forensic Science International 226: 106-117.**

**Hicklin, R. A., J. Buscaglia, M. A. Roberts, S. B. Meagher, W. Fellner *et al.*, 2011 Latent Fingerprint Quality: A Survey of Examiners. Journal of Forensic Identification 61: 385-418.**

**Holmgren, J. A., and J. Fordham, 2011 The CSI Effect and the Canadian and the Australian Jury. Journal of Forensic Sciences 56: S63-S71.**

**Langenburg, G., 2009 A Method Performance Pilot Study: Testing the Accuracy, Precision, Repeatability, Reproducibility, and Biasability of the ACE-V Process. Journal of Forensic Identification 59: 219-257.**

**Langenburg, G., 2012 A Critical Analysis and Study of the ACE-V Process, pp. 355 in *Institute of Forensic Science / School of Criminal Justice*. University of Lausanne, Lausanne.**

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

**Langenburg, G., C. Champod and T. Genessay, 2012 Informing the judgments of fingerprint analysts using quality metric and statistical assessment tools. Forensic Science International 219: 183-198.**

**Lennard, C. J., 2013 Fingerprint identification: how far have we come? Australian Journal of Forensic Sciences to appear.**

**McQuiston-Surrett, D., and M. J. Saks, 2009 The Testimony of Forensic Identiﬁcation Science: What Expert Witnesses Say and What Factﬁnders Hear Law and Human Behavior 33: 436-453.**

**Mnookin, J. L., 2010 The Courts, the NAS, and the Future of Forensic Science. Brooklyn Law Review 75: 1209-1275.**

**Moenssens, A., 1979 Resolution VII - A Monkey On Our Back. Identification News: 3-5.**

**Murch, R. S., A. L. Abbott, E. A. Fox, M. S. Hsiao and B. Budowle, 2012 Establishing the Quantitative Basis for Sufficiency Thresholds and Metrics for Friction Ridge Pattern Detail and the Foundation for a Standard, pp. U.S. Department of Justice, Washington DC.**

**National Institute of Justice, 2013 Forensic and Investigative Sciences Research Priorities, pp.**

**National Research Council, 2009 *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press, Washington, D.C.**

**Neumann, C., 2012 Statistics and Probabilities as a Means to Support Fingerprint Examination, pp. 407-452 in *Lee and Gaensslen's Advances in Fingerprint Technology*, edited by R. S. Ramotowski. CRC Press, Boca Raton.**

**Neumann, C., C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz *et al.*, 2007 Computation of Likelihood Ratios in Fingerprint Identification for Configurations of Any Number of Minutiae. Journal of Forensic Sciences 52: 54-64.**

**Neumann, C., C. Champod, R. Puch-Solis, D. Meuwly, N. Egli *et al.*, 2006 Computation of Likelihood Ratios in Fingerprint Identification for Configurations of Three Minutiae. Journal of Forensic Sciences 51: 1255-1266.**

**Neumann, C., C. Champod, M. Yoo, T. Genessay and G. Langenburg, 2013 Improving the Understanding and the Reliability of the Concept of "Sufficiency" in Friction Ridge Examination, pp. National Institute of Justice.**

**Neumann, C., I. W. Evett and J. Skerrett, 2012 Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm. Journal of the Royal Statistical Society 175: 371-415 (with discussion).**

**Osterburg, J. W., and S. A. Bloomington, 1964 An Inquiry Into the Nature of Proof "The Identity of Fingerprints". Journal of Forensic Sciences 9: 413-427.**

**Pulsifer, D. P., S. A. Muhlberger, S. F. Williams, R. C. Shaler and A. Lakhtakia, 2013 An objective fingerprint quality-grading system. Forensic Science International 231: 204-207.**

**Richmond, S., 2004 Do fingerprint ridges and characteristics within ridges change with pressure?, pp. Australian Federal Police.**

**Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST), 2012 Standard for the Documentation of Analysis, Comparison, Evaluation and Verification (ACE-V) (Latent), ver. 2.0.**

**Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST), 2013 Standard for Examining Friction Ridge Impresions and Resulting Conclusions, ver. 2.0.**

**Stoney, D. A., and J. I. Thornton, 1986 A Critical Analysis of Quantitative Fingerprint Individuality Models. Journal of Forensic Sciences 31: 1187-1216.**

**Stoney, D. A., and J. I. Thornton, 1987 A Systematic Study of Epidermal Ridge Minutiæ. Journal of Forensic Sciences 32: 1182-1203.**

**Thornton, J. I., 1977 The One-Dissimilarity Doctrine in Fingerprint Identification. International Criminal Police Review 32: 89-95.**

**Ulery, B. T., R. A. Hicklin, J. Buscaglia and M. A. Roberts, 2011 Accuracy and Reliability of Forensic Latent Fingerprint Decisions. Proceedings of the National Academy of Sciences of the United States of America 108: 7733-7738.**

**Ulery, B. T., R. A. Hicklin, J. Buscaglia and M. A. Roberts, 2012 Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners. PLoS ONE 7: e32800.**

**Ulery, B. T., R. A. Hicklin, G. I. Kiebuzinski, M. A. Roberts and J. Buscaglia, 2013 Understanding the Sufficiency of Information for Latent Fingerprint Value Determinations. Forensic Science International 230: 99-106.**

**United States Department of Justice, and Office of the Inspector General - Oversight and Review Division, 2006 *A Review of the FBI's Handling of the Brandon Mayfield Case*, Washington DC.**

**Yoon, S., E. liu and A. K. Jain, 2012 On Latent Fingerprint Image Quality, pp. in *Proceedings of the 5th International Workshop on Computational Forensics*, Tsukuba, Japan.**

# Appendices

## *Appendix A – Terms and Definitions*

**Cannot Exclude** – A possible conclusion with the LVMPD defined as “…when the latent print has detail in agreement with the exemplar prints; however, it is insufficient to identify the source due to the quality or quantity of the *latent print*. There is data to support the conclusion that the latent print and the exemplar print were made by the same source; however, the selectivity of the available corresponding data is not strong enough to disregard the possibility that another source could have left the print.”

**Exemplar** – See Print.

**Focal Area** – A delta or a core.

**Insufficient Agreement** – A conclusion proposed by this research that indicates that detail was found in agreement between a mark and a print, but the amount of agreement is insufficient to meet the threshold required for an Identification decision. This conclusion is intended to provide a means for presenting information on the results of comparisons that might otherwise be reported as Inconclusive and also to provide an acceptable conclusion that is less than Identification in the case of marks of low quality that should not be Identified, but are nonetheless probative.

**Mark** – A friction ridge impression of unknown origin; known in the United States as a “latent print”.

**Print** – A friction ridge impression of known origin, typically taken with ink under controlled conditions.

**Quality Metric** – A tool used for objectively measuring the information available in a mark to determine its value.

**Resolution VII** – A resolution passed by the IAI in 1979, then reworded and passed again in 1980, that stated: “[The delegates of the IAI] state unanimously that friction ridge identifications are positive, and officially oppose any testimony or reporting of possible, probably or likely friction ridge identifications found on hands and feet…“ (Champod 1995). This resolution was rescinded in 2010 by Resolution 2010-18 of the IAI.

**Specificity (of minutiae)** – A reference to the discriminating power of a particular minutia, typically based upon its rarity, or perceived rarity, in the population at large. Less common minutiae are said to have greater specificity.

**Target Group** – An easily recognizable grouping of minutiae, typically used as a search image during Comparison.

**Utility Metric** – A type of quality metric proposed in this research that uses data about a mark in multiple categories to calculate a score that can be used to measure the utility of the mark for a number of uses to include: value v. no value; complex v. non-complex; AFIS quality v. AFIS quality, but proceed with caution; standard quality v. high quality; appropriate for a definitive conclusion v. definitive conclusion not warranted; and low, medium, and high difficulty.

**Voir Dire** – A process by which an expert recites his qualifications to satisfy the Court of his expert status.

## *Appendix B – Abbreviations and Acronyms*

ACE-V Analysis-Comparison-Evaluation-Verification

AFIS Automated Fingerprint Identification System

CLPE Certified Latent Print Examiner (a distinction conferred by the IAI after successful completion of an exam)

CTS Collaborative Testing Services

GYRO Green-Yellow-Red-Orange Annotation System

IAI International Association for Identification

LVMPD Las Vegas Metropolitan Police Department

LSX Large Scale Experiment (Experiment Two)

NAS National Academy of Sciences

NIJ National Institute of Justice

NIST National Institute of Standards and Technology

NRC National Research Council

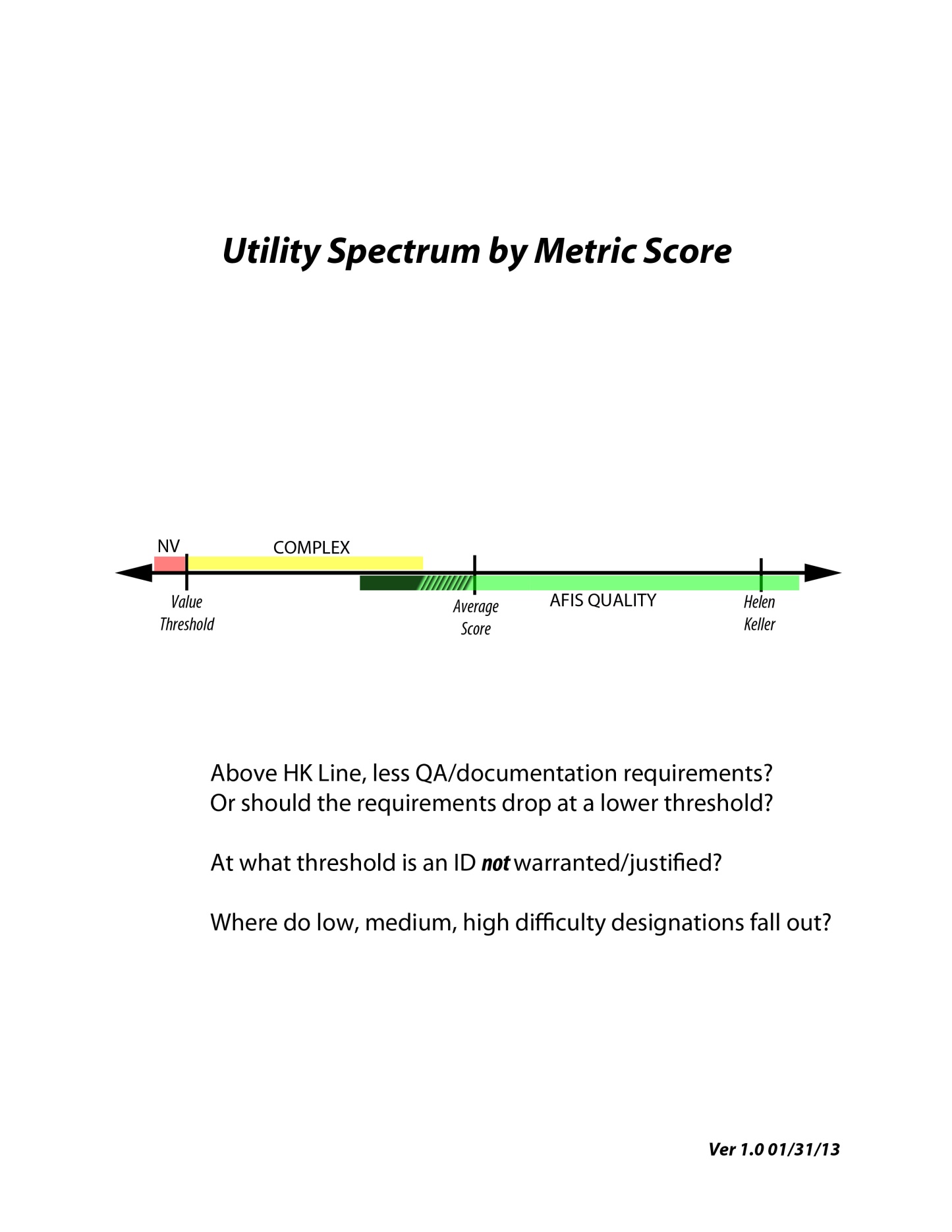
OIG Office of the Inspector General

QA Quality Assurance

SWGFAST Scientific Working Group on Friction Ridge Analysis, Study and Technology

ULW Universal Latent Workstation

## *Appendix C – Utility Score Continuum*

****

## *Appendix D – Utility Metric Questionnaire*

Utility Metric

**Anatomical Source Score: \_\_\_\_\_\_\_**

* Anatomical source is obvious and certain
* Anatomical source can be determined within one reference
* Anatomical source cannot be determined

**Orientation Information Score: \_\_\_\_\_\_\_**

* Orientation is clear
* Orientation cues exist to orient within one reference
* Orientation is unknown

**Size or Area Score: \_\_\_\_\_\_\_**

* Plain impression size
* At least half plain impression size
* At least a quarter of plain impression size
* Less than a quarter of plain impression size

**Clarity Score: \_\_\_\_\_\_\_**

* High clarity – features clearly visible throughout print
* Medium clarity – features clearly visible in approximately half the visible area
* Low clarity – features clearly visible in only a limited area
* Very poor clarity – features difficult to make out

**Level One Score: \_\_\_\_\_\_\_**

* Pattern type is clear
* Likely pattern type is known, but a reference may be necessary
* Some pattern cues present
* No pattern cues present

**Level Two Score: \_\_\_\_\_\_\_**

* 20 or more high confidence Galton points visible
* Between 8 and 19 high confidence Galton points visible
* Between 5 and 7 high confidence Galton points visible
* Fewer than 5 high confidence Galton points visible (lower-confidence points may be present)

**Level Three Score: \_\_\_\_\_\_\_**

* Clear level 3 detail visible in more than half the print
* Clear level 3 detail visible in less than half the print
* Some amount of level 3 detail visible, but clarity of details is marginal
* No reliable level 3 detail

**Rarity of Features Score: \_\_\_\_\_\_\_**

* One or more groupings of highly discriminative features
* One or more features of extreme rarity (e.g. trifurcation, em)
* One or more features of higher rarity (e.g. bridge, overlap)
* One or more features of moderate rarity (e.g. dot, short ridge)
* No compound minutiae (e.g. all ridge endings and bifurcations)

**Grouping of Features Score: \_\_\_\_\_\_\_**

* Most features are arranged in easily recognizable target groups; some features isolated
* One or more easily recognizable target groups, but many features isolated
* Features are visible, but are spread out and do not form easily searchable target groups

**Distortion Score: \_\_\_\_\_\_\_**

* No distortion impacting interpretation of the print
* Minor distortions that are easily explained/worked through
* Areas of significant distortion, but also areas of low distortion with usable features
* Major distortions that severely impact the interpretation of the print

**Subjective Bins**

* No value
* Of value – high difficulty
* Of value – medium difficulty
* Of value – low difficulty

*If print is of value:*

**Subjective Complexity**

* Complex
* Not Complex

**AFIS Determination**

* AFIS Quality
* Non AFIS Quality
* AFIS Quality, but Additional QA Measures Warranted

**TOTAL SCORE: \_\_\_\_\_\_\_**

**BIN: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

1. An acronym consisting of four phases: Analysis, Comparison, Evaluation, and Verification, used to describe the process undertaken in the comparison of two areas of friction ridge detail. [↑](#footnote-ref-1)
2. As will be described further on in the mémoire, there are marks that have sufficient quality to be entered into the AFIS system, yet for which the risk of a coincidental match is higher due to numerous factors. These marks should be treated with additional caution prior to rendering a conclusion of identification. [↑](#footnote-ref-2)
3. It will be argued in this mémoire that marks of sufficiently high quality do not carry the same necessity for documentation and QA policies, and thus may enjoy an abbreviated procedure that will save time for practitioners. [↑](#footnote-ref-3)
4. There are marks that contain sufficient quality and quantity of information to perform a comparison, and from which some information may be gleaned that could be of use to a case, but will not rise to a threshold of “identification”. For these marks, while some information may be found in agreement between the mark and a subject print, there is not a high enough quantity of data to reduce the possibility that someone else may have left the mark to the point at which it may be disregarded. For these marks, a conclusion of “identification” would be inappropriate, even if the examiner felt that the mark and the print were likely from the same source. These marks would fall under the category of “Cannot Exclude” for agencies such as the Las Vegas Metropolitan Police Department (LVMPD), which will be discussed in greater detail later. The same logic can be applied at the other end of the spectrum, to marks that may show some indistinct differences from the print, but not enough to rise to the threshold of an exclusion. For this reason, these marks are being designated by this research as “definitive conclusion not warranted”, rather than the narrower category of “identification not warranted”. [↑](#footnote-ref-4)
5. It is recognized that several of these categories may overlap. Part of the work of validating the metric will include defining where those areas lie. A graphic representing the possible relationship between some of the categories is presented in Appendix C; however, the true relationships will not be clear until data have been collected. [↑](#footnote-ref-5)
6. Here the author takes issue with the standard dichotomy of value options traditionally presented in the discipline. It has been represented by SWGFAST (2013) and in the Blackbox Study (Ulery *et al.* 2011) that there are two approaches to the value decision: Value for Identification (VID) and Value for Exclusion Only (VEO). These two approaches are presented as hierarchical in nature: a mark that is VID is of high quality, while a mark that is VEO is still useful, but typically of *lower quality* than the VID mark. The author believes there is a fundamental flaw to approaching the issue in this way.

   As any examiner practicing in a laboratory that allows an exclusion decision knows, reaching an exclusion threshold can be a very difficult thing. While an examiner who has found a match *knows* that they have reached the identification decision because they see all the corresponding data before them, knowing that an exclusion has been reached can be far more difficult, particularly with a mark that lacks clear focal points or orientation and anatomical source cues. Finding sufficient evidence of non-similarity can in fact be a much *higher* threshold than finding sufficient evidence of similarity, depending on the situation. For this reason, the author feels that treating a VEO mark as somehow inferior in quality to a VID mark is mis-stating the situation – very often, a mark must be of higher quality to definitively exclude than to definitively identify.

   Because of this, the author will be steering clear of the VID-VEO paradigm and instead will focus on marks being either appropriate for a definitive conclusion (Identification or Exclusion), or not appropriate for a definitive conclusion (i.e. either existing along the spectrum of Inconclusive, or lacking any comparison value at all). [↑](#footnote-ref-6)
7. One well-known current model is that proposed by Neumann et al, which was introduced for three-minutiae configurations (Neumann *et al.* 2006), expanded to n-minutiae configurations (Neumann *et al.* 2007), and further presented after a lengthy validation study (Neumann *et al.* 2012). [↑](#footnote-ref-7)
8. It should be noted, however, that in many of the early studies, these frequencies were assumed to be roughly equal, or were arbitrarily made up; only a very few feature types were considered; assumptions of independence were made; and often only limited pattern types were considered (Stoney and Thornton 1986). [↑](#footnote-ref-8)
9. The author was cross-examined at some length on this very topic during a 2010 Motion to Exclude hearing (2010). [↑](#footnote-ref-9)
10. For example, a spur is more discriminating than a bifurcation and a ridge ending standing alone. [↑](#footnote-ref-10)
11. This resolution was hotly debated at the time and was eventually passed despite the arguments of objectors such as Moenssens and Davis, who argued that the Resolution represented a step backward for the legitimacy of friction ridge examination as a science, noting particularly that no information on similarity at all could be offered in court without an absolute conclusion under this Resolution (Moenssens 1979) and that marks have exculpatory as well as incriminatory value and that information less than “proof-value” could still be probative (Davis 1979). Interestingly, these arguments are still being made today and are at the backbone of the pressure that resulted in the IAI rescinding this Resolution in 2010. [↑](#footnote-ref-11)
12. This was argued by Champod well before the development of the current statistical models. He characterized the interpretation of friction ridge evidence as “an increasing scale [that] runs from exclusion to identification” (Champod 1995). More recently, Neumann has also argued against the dichotomous model of conclusion reporting (Neumann 2012). [↑](#footnote-ref-12)
13. Lennard espouses a compromise of sorts, in which the expert’s opinion is presented along with statistical information to back it up; but he also presents the research of Martire, which suggests that jurors are poor Bayesians and may struggle to understand evidence presented in a statistical framework. Nevertheless, he conclusively states that the Inconclusive category is “overly broad and uninformative”. (Lennard 2013) [↑](#footnote-ref-13)
14. Limited research on juror understanding of statistical forensic testimony seems to indicate that they struggle to understand this information (McQuiston-Surrett and Saks 2009; Lennard 2013). So should it be presented at all? On the other hand, McQuiston and Saks’s data suggest that simply stating the expert’s opinion leads jurors to overestimate the weight to assign to forensic evidence. [↑](#footnote-ref-14)
15. This may be different from the participant’s confidence in a given feature, as an examiner may assign a high weight to a feature in which they have low confidence due to its high discriminating power. Likewise, an examiner may assign low weight to a feature in which they have high confidence due to its lack of discriminating power (e.g. a ridge ending). [↑](#footnote-ref-15)
16. For ease of reading/writing, this category will be referred to as “Not Appropriate for a Conclusive Decision” throughout this document. [↑](#footnote-ref-16)
17. As noted above in the State of the Art, several of these categories may partially overlap. Therefore, it may be possible and acceptable for a particular utility score to reside in more than one category at once. For example, a mark may be simultaneously Of Value; Complex; and AFIS Quality, but proceed with caution. [↑](#footnote-ref-17)
18. A mock-up of the short form is provided in Appendix D and can give further insight to the information being encapsulated by each category. These categories are, naturally, subject to change in light of data collected during Experiments One and Two. [↑](#footnote-ref-18)
19. Whether the examiner can determine where the mark came from (e.g. distal phalanx, palm, etc) [↑](#footnote-ref-19)
20. Whether the examiner can determine the distal orientation of the mark [↑](#footnote-ref-20)
21. For instance, the metric is designed to provide Analysis threshold decisions in multiple categories, which theoretically should provide useful information about many aspects of each mark. However, if, in practice, the metric clusters the scores for the marks around two main categories (e.g. value and no value), then the available categories are of no practical value and it could be found that all the extra flexibility available provides no practical benefit over the existing tools. [↑](#footnote-ref-21)
22. As discussed earlier, such cases are given the designation “Cannot Exclude” at the LVMPD to indicate that some detail was found in agreement that did not rise to the threshold of an identification. However, this author dislikes the Cannot Exclude label for two reasons: (1) it is misleading, as it tends to imply that the result of the comparison was nearly, but not quite, an Exclusion, rather than being at the opposite end of the spectrum, near an Identification; and (2) it is too weak a statement for the situations the author describes. The author’s objections appear to be echoed by Cole (2011). For these reasons, the author prefers the term “Insufficient Agreement”. [↑](#footnote-ref-22)
23. Since the close non-matches in this experiment will be generated by an AFIS system, any errors made involving marks in this experiment may be considered to be *de facto* AFIS errors; therefore, the results are directly applicable to errors that may realistically be expected during casework in an AFIS environment. [↑](#footnote-ref-23)