Document Title:	**Error Rates for Latent Fingerprinting as a Function of Visual Complexity and Cognitive Difficulty**
Author(s):	**Jennifer Mnookin, Philip J. Kellman, Itiel Dror, Gennady Erlikhman, Patrick Garrigan, Tandra Ghose, Everett Metler, Dave Charlton**
Document No.:	**249890**
Date Received:	**May 2016**
Award Number:	**2009-DN-BX-K225**

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

ERROR RATES FOR LATENT FINGERPRINTING AS A FUNCTION OF VISUAL COMPLEXITY AND COGNITIVE DIFFICULTY

NIJ Award 2009-DN-BX-K225

Jennifer Mnookin (P.I.), Philip J. Kellman (Co-P.I.), Itiel Dror, (Co-P.I.)
Gennady Erlikhman, Patrick Garrigan, Tandra Ghose, Everett Metler, Dave Charlton

Abstract

Comparison of forensic fingerprint images for purposes of identification is a complex task that, despite advances in image processing, still requires highly trained human examiners to achieve adequate levels of performance. Latent fingerprints collected from crime scenes are often noisy, distorted, and represent only a portion of the total fingerprint area, making matching tasks difficult. While it is clear that expertise in fingerprint comparison, like other forms of perceptual expertise, such as face recognition or aircraft identification, depends on perceptual learning processes that lead to discovery of features and relations that matter in comparing prints, relatively little is known about the perceptual processes involved in making fingerprint comparisons, and even less is known about how the visual characteristics of fingerprint pairs relate to comparison difficulty. This project aims to determine more about the relationship between the measurable, visual dimensions of fingerprint pairs and the level of comparison difficulty for human examiners, both experts, and to a lesser degree, novices. For this research, we assembled a new database of latent fingerprints, matching tenprints, and close, non-matching tenprints. Using this database, we measured expert examiner performance and judgments of difficulty and confidence in a variety of settings. For the experts, we developed a number of quantitative measures of image characteristics and used multiple regression techniques to discover predictors of error as well as perceived difficulty and confidence. A number of useful predictors emerged, including variables related to image quality metrics, such as intensity and contrast information, as well as measures of information quantity, such as the total fingerprint area. Also included were configural features that fingerprint experts have noted, such as the presence and clarity of global features and fingerprint ridges. Within the constraints of the overall low error rates of experts, a regression model incorporating the derived predictors demonstrated reasonable success in predicting difficulty for print pairs, as shown both in goodness of fit measures to the original data set and in a cross validation test. The results indicate the plausibility of using objective fingerprint image metrics to predict expert performance and subjective assessment of difficulty in fingerprint comparisons. We also examined the extension of these results to settings that better approximate real-world fingerprint examiner scenarios, and found our regression model continued to provide significant explanatory value for a substantial portion of the prints. While further research is necessary, this research provides strong support for the plausible but previously untested assumption that for expert fingerprint analysis, difficulty (and by extension, error rate) is in significant part a function of measurable, visual dimensions of print comparison pairs. In addition to this primary focus, we also conducted several extensions to this research, involving expert metacognition and novice comparison. These experiments showed that experts have substantial, albeit imperfect, subjective knowledge about the difficulty of print pairs. Our experiments also showed that novices perform very poorly and showed no consistent pattern of feature use. This research thus also contributes to our understanding of the source and extent of human expertise in latent fingerprint analysis.

Table of Contents

Executive Summary

There has been a longstanding belief in the scientific validity of fingerprint evidence, based both on the apparent permanence and uniqueness of individual fingerprints and on the experience-based claims of trained fingerprint examiners. In the past, fingerprint evidence, in the hands of an experienced examiner appropriately applying the methods of the field, was often claimed to be "infallible" or to have a "zero error rate" (Cole, 2005; Mnookin 2008b). Yet systematic scientific study of the accuracy of fingerprint evidence is a rather late development, still very much in progress. The traditional claim of infallibility for fingerprint identification has been brought to the spotlight and questioned in light of high-profile cases in which errors have been discovered. While it is likely that well-trained, experienced examiners are highly accurate when making positive identifications, it is also clear that errors still occur. Recently, with the National Academy of Sciences (2009) inquiry into forensic science, new research has begun to emerge. The available data now suggest a low level of erroneous match determination by experts under experimental conditions and a higher rate for erroneous exclusion determinations (e.g., Ulery, Hicklin, Buscaglia, & Roberts, 2011; Tangen, Thompson, & McCarthy, 2011).

At present, however, fingerprint identification is a strikingly subjective process (e.g., NIST 2012, chapter 3). There are not validated, objective metrics for any meaningful step in the comparison process, from the determination of whether there is sufficient information to warrant a comparison to the final judgment of match or non-match (e.g.,Mnookin, 2010). Not only are these judgments and conclusions subjective, but at present, there is no method by which the relative simplicity or difficulty of a print pair can be determined. While examiners may, based on their experience, have a view about the relative ease or difficulty of a comparison, whether these subjective judgments are warranted has not been known. The main goal of the project was therefore to identify and quantify fingerprint image features that are predictive of identification difficulty and accuracy. A quantitative way of assessing fingerprint image quality and comparison difficulty would be an extremely useful development. First and foremost, objective metrics for measuring difficulty create the possibility of associating error rates with the level of difficulty. It is a matter of common sense to recognize that if some print comparisons are unusually hard, examiners will therefore be more prone to make possible mistakes in their analysis; conversely, easy comparisons would be expected to produce fewer errors than hard ones. But common sense or not, prior to this research, no research focused on examining this issue, or attempting to look at the relationship between error rate and difficulty.This project provides foundational steps toward the possibility of associating error rates with difficulty. An objective metric of difficulty has other benefits as well. Such a metric can be used to alert examiners when additional care is warranted (i.e., for a particularly difficult comparison), to caution examiners who are inclined to label a print pair as inconclusive that further examination might be prudent, and to create a set of fingerprint images with objective difficulty ratings that can be used for training examiners.

In addition to creating a fingerprint image quality metric, the project had several other objectives: (1) to create a realistic fingerprint image database with known ground truth (i.e., true matching prints and close non-matches); (2) to add to the rather modest scientific literature investigating fingerprint expertise; that is, to assess objectively expert performance both in terms of accuracy in fingerprint identification and in terms of the image characteristics that related to performance; (3) to examine the relationship between objective expert performance and subjective assessments of difficulty and confidence, to evaluate experts' metacognitive abilities vis-à-vis the comparisons they make; (4) to compare the use of visual information by experts in their fingerprint analysis to the use made by novices.

The report is roughly divided into two sections. In the first part, we identify and describe the computations for several image features that we hypothesized might correlate with identification accuracy. The features were a mixture of image properties such as contrast and brightness levels, traditional (i.e., Level I, II, and III) fingerprint features such as visibility of deltas and clarity of ridges, and relational features that applied to the *pair* of prints in a comparison. The final point is important since a comparison does not depend only on the quality of the latent, but also on the shared information between it and the known print.

The fingerprint images that were used were latent prints collected from volunteers. Each volunteer left several impressions on a variety of glass objects. Practicing fingerprint examiners verified that the impressions we collected were realistic exemplars of the kinds of images that could be found in forensic settings. Corresponding known prints were also collected from each individual, providing a set of known matching prints. The latent prints were submitted to an AFIS system and close, non-matching prints were selected for each of the collected latents. The final print database contained over 500 pairs of matching and 500 pairs of non-matching prints.

In the second section, we describe three behavioral studies in which we measured expert examiner and novice performance on a subset of the database. In Experiment 1, expert examiners made timed comparison judgments for print pairs via an online interface that we designed for this purpose. Examiners were shown a pair of prints and made an identification or exclusion judgment and provided difficulty and confidence ratings for each such comparison. We collected data from 56 examiners at a forensic conference. Examiner accuracy was high: there were 200 errors made out of 2292 total comparisons (overall accuracy of 91%) with more false negative or incorrect exclusions (14%) than false positives or incorrect identifications (3.2%). Difficulty and confidence ratings were highly correlated with accuracy, indicating that experts were often able to identify which comparisons were difficult or likely to be error-prone. We fit a regression model to the accuracy data to predict examiner performance from image features. We identified several features that were predictive of accuracy, including the ratio of the image areas of the latent and known print, the combined reliability of ridge information in both images, the variability of contrast across small regions of both images, and the visibility of deltas in the latent print. The model was also fairly successful in predicting the accuracy on a held-out set of fingerprints that were not used in the regression ($R^2_{adj} = 0.64$). A separate, classification analysis was able to identify print pairs for which at least one examiner made a mistake with a 75% (15/20) accuracy.

Experiment 2 sought to compare expert performance with that of novices. Naïve participants with no fingerprint examination training made fingerprint comparisons in an interface we designed. Not surprisingly, they performed far below experts, indeed approximately at chance. For a second group of subjects, we showed a brief video that highlighted various fingerprint features that could be used in making comparisons such as *minutiae*. Participants who viewed the video only showed a mild improvement in performance. However, there were noticeable changes in their biases to make an identification; in particular, trained participants seemed more hesitant to label a fingerprint pair as coming from the same source, perhaps because the training video emphasized the cost of making an incorrect identification and made the difficulty of fingerprint examination more apparent. We fit the novice data to the same regression model from Experiment 1. We found that untrained novices relied on different features than experts and that some features that experts used had the opposite effect on novice performance. Trained novices had the weight given to certain features shifted closer to those of experts, indicating that with training novices may learn which image information is relevant for identification, while learning to ignore irrelevant features.

Experiment 3 served as an extension and validation of Experiment 1. We used the data from Experiment 1 to make performance predictions for a new subset of fingerprints from the database.

We also designed a new web interface that incorporated many of the image processing tools that are typically available to experts. For example, website users could reverse the contrast of the images, zoom in and out, mark *minutiae*, rotate the image, and increase or decrease brightness and contrast. In addition, we gave examiners unlimited time for each comparison. We recruited a new set of 34 expert examiners and allowed them to perform the experiment on their own time. Performance was slightly higher than in Experiment 1, with 10% incorrect exclusions and 0.25% incorrect identifications. Fitting a regression model identified many of the same predictors as in Experiment 1. Using the model from Experiment 1 to predict performance for this set of prints was qualitatively successful for comparisons that were not labeled inconclusive by any examiners, but performed poorly for those that were. Further work needs to be done to incorporate image processing tool use into the regression model and to account for performance on comparisons that are labeled inconclusive.

We have taken several foundational steps in this project. First, and most importantly, we have successfully demonstrated that there are image features that can be quantified and used to predict examiner performance. While we do not claim our list to be complete or exhaustive, this research is a vital first step in proving that such features can be found and provides a useful guide for future research. In addition, these experiments provide persuasive evidence that there is a meaningful and important correlation between comparison difficulty and error rate. Errors were not distributed randomly across our exemplars, but rather, significantly clustered, revealing that difficulty is, to a significant extent, a function of visual aspects of the specific comparison. This recognition also provides evidence that it is not especially helpful to seek a field-wide "error rate" for latent fingerprint identification. Instead, as our scientific knowledge of fingerprint identification continues to progress, it will be more useful to seek error rates for different categories of comparisons, based on objective difficulty level.

Second, we have provided useful evidence for and quantification of examiner expertise, both in controlled (Experiment 1) and more realistic (Experiment 3) settings. This has also allowed us to examine inter-rater reliability and metacognitive judgments (i.e., whether examiners are aware which prints are actually difficult) and to contrast performance with novices (Experiment 2). Furthermore, similarity in examiner performance between Experiments 1 and 3 suggests that one can study examiner expertise without needing to perfectly replicate work conditions and still get a useful estimate of performance. This concern has previously limited research, and the comparative experiments in this study provide valuable information for future researchers to consider when engaging in study design. Third, we found preliminary evidence that even a small amount of training can adjust novice performance. This suggests that with more extensive bursts of training we can examine, in a controlled manner, the process by which an examiner becomes an expert. This will allow for various interventions in the training process, allowing for more efficient and automated training techniques. Fourth, we have constructed an independent fingerprint database, with ground-truth and difficulty ratings that can be used for future studies or for examiner training. In addition, we have created several web-deliverable tools that can also be used for evaluation or training that replicate many of the image processing features available to examiners.

Beyond the scientific findings, there are important aspects of the research that can impact policy. A more sophisticated understanding of the relationship between error rate and difficulty is, or should be, important for the courts in weighing fingerprint evidence. Courts are instructed, when assessing expert evidence, to focus on the "task at hand", and this research helps to show that fingerprint examination may vary in difficulty in ways that may be relevant to its evaluation as evidence (Daubert vs. Merrell Dow Pharmaceuticals, 1993; Kumho Tire Co. vs. Carmichael, 1999). More nuanced assessments of fingerprint task difficulty might, for example, affect how a judge understands admissibility of that specific conclusion, or what degree of certainty the expert will be allowed to express, or it might appropriately impact the weight given to a specific match conclusion by the fact-finder (Faigman, Blumenthal, Cheng, Mnookin, Murphy & Sanders, 2012).

The implications of these findings go beyond the court; they provide vital insights that can considerably enhance the procedures used in forensic laboratories. For example, similar to medical triage, the need for different procedures and checks can be made to fit the difficulty of the comparison. The understanding of what makes some comparisons more difficult than others also has implications for the selection and training of fingerprint examiners. During selection, benchmarks and skill sets can be set as criteria to ensure candidates have the acquired the necessary cognitive abilities needed to perform their job adequately. In addition, in evaluating the significance of errors for trainees, better information about difficulty level will be of great assistance. Trainees who make mistakes on simpler stimuli can be distinguished from those whose errors occur only on more difficult materials; for evaluating performance, all errors are not – and should not be treated as – equal.

While further research is clearly necessary to build on these results, this research provides significant steps forward for helping to establish that error rates are related to difficulty; for beginning to provide evidence for what visual dimensions of fingerprint comparison pairs are associated with difficulty; and for helping to tease out both examiner's metacognitive abilities and the substantial degree of examiner expertise in this domain.

I. Introduction

There has been a longstanding belief in the scientific validity of fingerprint evidence, based both on the apparent permanence and uniqueness of individual fingerprints and on the experience-based claims of trained fingerprint examiners. In the past, fingerprint evidence, in the hands of an experienced examiner appropriately applying the methods of the field, was often claimed to be "infallible" or to have a "zero error rate" (Cole, 2005; Mnookin 2008b). Yet systematic scientific study of the accuracy of fingerprint evidence is a rather late development, still very much in progress. The traditional claim of infallibility for fingerprint identification has been brought to the spotlight and questioned in light of high-profile cases in which errors have been discovered. While it is likely that well-trained, experienced examiners are highly accurate when making positive identifications, it is also clear that errors still occur. Recently, with the National Academy of Sciences (2009) inquiry into forensic science, new research has begun to emerge. The available data now suggest a low level of erroneous match determination by experts under experimental conditions and a higher rate for erroneous exclusion determinations (e.g., Ulery, Hicklin, Buscaglia, & Roberts, 2011; Tangen, Thompson, & McCarthy, 2011).

Contrary to popular belief and its depiction on many television shows, fingerprint identification – matching a fingerprint from a crime scene to one on file – is not a fully automated process. While algorithms can compare known prints (fingerprints collected in controlled conditions such as in a police station where the fingerprint images are clear) with high accuracy, identifying latent prints (those found at a crime scene) falls to individual fingerprint examiners who are extensively trained (Vokey, Tangen, & Cole, 2009). However, the nature and extent of examiner expertise has only recently come under scientific scrutiny (e.g., Busey & Parada, 2010; Busey & Vanderkolk, 2005; Dror & Charlton, 2006; Dror, Charlton, & Péron, 2006; Dror, Champod, Langenburg, Charlton, Hunt, & Rosenthal, 2011; Tangen, Thompson, & McCarthy, 2011; Ulery, Hicklin, Buscaglia, & Roberts, 2011). While several proficiency tests have been used to evaluate expertise, many may have used an overly limited number of prints; this may have led to inaccurate estimates of examiner performance because of idiosyncratic fingerprint properties that made a particular identification easy or difficult (Cole, 2006, 2008; Vokey, Tangen, & Cole, 2009).

Mistakes in fingerprint matching are costly and can put lives and livelihoods at risk. Errors in fingerprint matching are of two types that have different implications. A *false negative*, where a matching pair is labeled as non-matching, could, in a criminal proceeding, allow a guilty suspect to be set free. A *false positive*, where a non-matching pair is labeled as a match, could lead to, in a criminal proceeding, the conviction of an innocent person. Existing data suggest that fingerprint experts err more on the side of false negatives (about 8% of total judgments made on a match/non-match task for fingerprint pairs) than false positives (about 0.1%) (Langenberg, 2009; Tangen, Thompson, & McCarthy, 2011; Ulery, Hicklin, Buscaglia, & Roberts, 2011). Experts perhaps tend to incorporate the presumption of innocence, erring on the side that would free the guilty rather than convict the innocent, although false positive rates are not zero.

The practical importance of understanding when and why fingerprint comparison errors occur is likely to increase as technology advances. Current Automated Fingerprint Identification Systems (AFISs) retrieve from a database a number of prints associated with known individuals that could be potential matches for a particular latent. Under typical procedures, the intervention of a human expert is required for deciding which if any candidates generated by an AFIS comprise a match to a latent. Candidate matches selected by a properly functioning AFIS should often appear similar to the latent entered into the system, a fact that likely increases the potential for human error. Imagine, by way of contrast, a situation in which an examiner is asked to compare a latent print to a known print of a particular suspect in a criminal case. Assuming the two prints are not from the same individual, it would be a remarkable coincidence if the prints were highly similar. Use of an AFIS has high value in extracting candidates from a database, but it puts the examiner in the position of routinely needing to distinguish actual matches from close (highly similar) non-matches. The likelihood of human error increases with the degree of similarity of the potential candidates extracted by AFIS, thereby making the comparison process more difficult (Ashworth & Dror, 2000; Vokey, Tangen, & Cole, 2009). With the increase in size of AFIS databases, the possibility of finding a look-alike non-match increases, thereby increasing the potential for false positive errors (Cole, 2005; Dror & Mnookin, 2010; Dror, Péron, Hind, & Charlton, 2005).

From a visual information processing perspective, it is interesting and important to determine what visual characteristics of fingerprints influence the ease and accuracy of comparisons. Ultimately, it may be possible to evaluate a fingerprint comparison in terms of the quantity and quality of visual information available (Pulsifer et al., 2013) in order to predict likely error rates and to assess when there is insufficient information to warrant any conclusion.

Perceptual Aspects of Fingerprint Expertise

If asked to give reasons for a conclusion in a given comparison, fingerprint examiners will report significant explicit knowledge relating to certain image features, such as global configurations, ridge patterns and minutiae, as these are often explicitly tagged in comparison procedures. They are also pointed out during training of examiners. It would be a mistake, however, to infer that the processes of pattern comparison and the determinants of difficulty are in general available for conscious report or explicit description. As in many other complex tasks in which learning has led to generative pattern recognition (the ability to find relevant structure in new instances) and accurate classification, much of the relevant processing is likely to be at least partly implicit (Chase & Simon, 1975; Schneider & Shiffrin 1997; for a review, see Kellman & Garrigan, 2009).

Like many other tasks in which humans, with practice and experience, attain high levels of expertise, feature extraction and pattern classification in fingerprint examination involves *perceptual learning* -- experience-induced changes in the way perceivers pick up information (Gibson, 1969; Kellman, 2002). With extended practice, observers undergo task-specific changes in the information selected – coming to discover new features and relationships that facilitate classification in that domain. Evidence supporting this claim comes from increased perceptual learning when these features are exaggerated during training (Dror, Stevenage, & Ashworth, 2008). While several studies have explored the influence of bias and emotional context on fingerprint matching and classification (e.g., Dror, Péron, Hind, & Charlton, 2005; Dror & Charlton, 2006; Dror, Charlton, & Péron, 2006; Dror & Cole, 2010; Dror & Rosenthal, 2008; Hall & Player, 2008), there has been relatively little work investigating perceptual aspects of expertise among examiners or perceptual learning processes that lead to expertise.

There are also profound changes in *fluency*: What initially requires effort, sustained attention, and high cognitive load comes to be done faster, with substantial parallel processing and reduced cognitive load (Kellman & Garrigan, 2009). In turn, becoming more automatic at extracting basic information frees up resources for observers to discover even more subtle or complex structural information (see, e.g., Bryan & Harter, 1899). This iterative cycle of discovery and automaticity followed by higher-level discovery is believed to play a significant role in attaining the impressive levels of performance humans can attain in areas such as chess, chemistry, mathematics, and air traffic control, to name just a few domains (Kellman & Garrigan, 2009; Kellman & Massey, 2013).

These considerations motivate the research presented here. The primary goals were to: (1) create a fingerprint database with ground-truth (true matches) information and sufficiently difficult comparison to use as a testing base for future experiments that evaluate expert performance, (2) measure expert examiner performance on a variety of prints including difficulty comparisons, (3) measure novice performance to create a basis of comparison for expert skill, and (4) create a predictive framework by which one could assign an appropriate level of confidence in expert decisions derived from an objective assessment of characteristics of the pair of images involved in a particular fingerprint comparison. These goals are interconnected. Examiner performance levels (error rates) are likely to depend on the complexity and difficulty of the comparison: as comparisons get more difficult, errors are more likely to occur. Hence, the characterization and prediction of error rates should relate to the perceived difficulty of the comparison. Notwithstanding this relationship, no previous research on fingerprint identification has attempted to generate objective models for the assessment of perceived fingerprint comparison difficulty. Note that we use the term *comparison* difficulty advisedly. One of the insights guiding our research was that the right question is not merely whether a particular print is 'easy' or 'difficult,' clear or unclear, rich in information or less so. Rather, the right question is the difficulty of a given comparison. While latents may vary in quality more than tenprints, and thus may be the primary driver of difficulty, the specific comparison will also be relevant to determining difficulty. (To see the point most clearly, consider: a low quality print might nonetheless be part of an easy comparison when the tenprint is of a different pattern type. Similarly, a high quality latent might be part of a difficulty comparison when it bears an unusually high degree of similarity to the tenprint to which it is being compared.)

Several studies have attempted to quantify expert performance. Tangen, Thompson, and McCarthy (2011) generated a testing set of 36 simulated latent prints from the Forensic Informatics Biometric Repository. Twelve were paired with a corresponding known print match, 12 were paired with a randomly selected, non-matching print from the same database, and 12 were paired with similar prints found by submitting the latent prints to the Australian National AFIS. This resulted in a testing set in which the ground truth was known, i.e., for each latent print there was a corresponding, correctly matching known print. Thirty-seven experts and 37 novices made similarity ratings on a scale of 1 (different) to 12 (same). Judgments of "inconclusive" were not allowed. Only accuracy information was computed from the rating scale. Performance in the dissimilar and similar non-matching conditions was highest for experts, at 100% and 99.32% respectively. Performance was lower when the latent and known prints matched: 92.12%, indicating that experts were more likely to "free the guilty" than "convict the innocent", although both kinds of errors were made. Novice performance was markedly lower than experts'. Their accuracy was best in the match and dissimilar conditions, with accuracies of 74.55% and 77.03% respectively, and worst in the similar non-match condition, with an accuracy of 44.82%. Similar performance in experts between similar and dissimilar non-matches may reflect the results of training that is absent in novices.

Despite the interesting findings, and the large quantity of test images and examiners, several important questions remain unaddressed by the Tangen et al (2011) study. While the fingerprints were collected in a realistic manner by having individuals grasp objects, it is unknown whether the set of prints is sufficiently representative. As with proficiency tests, perhaps this set of prints was particularly easy or difficult if they did not, for example, capture a sufficient variety of smudges and distortions that might occur. The prints were generated by having individuals grasp objects or push open a door; these kinds of manipulations may have yielded a disproportionate amount of relatively clean fingerprints with little distortion. An expert (one of the authors) determined that all of the prints used in the study had sufficient information to make a judgment (i.e., would not be judged as "inconclusive"), but (through no fault of the authors, given the lack of objective metrics available) there was not any other way to determine difficulty. Importantly, one would like to be able to somehow assess the difficulty of fingerprint comparisons, to be able to determine when a comparison should be easy and when it should be difficult and could lead to an increased error rate. For example, measuring novice performance only on matches and dissimilar non-matches would lead one to incorrectly estimate their average accuracy at comparing prints to be approximately 75%. The similar non-match condition in which accuracy is at chance is critical in demonstrating the difference between experts and novices. Without being able to quantify the degree of dissimilarity (difficulty of comparison) in the similar non-match condition, one can only say that expert performance is near perfect for this particular set of comparisons.

Other studies, using different fingerprint databases, have found novice accuracy to be closer to 85% (Vokey, Tangen, & Cole, 2009, Experiment 2). Discrepancy in accuracy estimates could be due to variability in the kinds of prints used for the study or the kinds of image manipulation tools (e.g., rotation of one of the images) available to participants. Without a quantitative measure of the properties of a fingerprint image that make a comparison difficult or easy, comparing accuracy rates across heterogeneous databases would provide little information about true ability, since the prints used could be substantially varied in difficulty.[1]

Such considerations naturally lead to the question of what features of the images make a particular comparison difficult or easy. For example, if many experts made errors in the match condition on the same fingerprints, it would be useful to know what features of those fingerprint images led to the errors. Identifying objective image features that correlate with accuracy would allow for predictions of comparison difficulty and could be used to tag print pairs that require additional scrutiny because they are more likely to be erroneously classified.

Ulery, Hicklin, Buscaglia, and Roberts (2011) have made an important first step toward addressing these issues. They created a large dataset of 744 print pairs including subjectively rated “low quality” latents that were rated as representative of those encountered in regular casework. In addition, the overall difficulty of comparisons was rated to be similar to casework by a majority of participants. A slightly greater proportion of images used in the study was rated as poor quality according to the NIST Fingerprint Image Quality Metric (NFIQ) compared to examples from AFIS. Non-matching pairs were selected by submitting latent pairs to an Integrated AFIS. 169 examiners participated, each evaluating approximately 100 randomly selected print pairs. Because the testing sets were generated randomly, there was large variability in the number of examiners that evaluated each pair. Examiners were given the option to label a comparison as “inconclusive”. Among 4,985 non-match trials, there were 6 false positives (accuracy: 99.89%), each on a different comparison, made by 5 unique examiners. There were 611 misses (matches evaluated as non-matching) out of 8,189 comparisons (accuracy: 92.54%). These results were very similar to identification accuracy amongst experts in Tangen, Thompson, and McCarthy, (2009). Performance correlated with years of experience and certification suggesting that some variability is due to individual differences among experts (Ulery et al., 2011). Participants were also asked whether there was enough information in each latent image to make an identification, to make an exclusion (less information may be needed for exclusions since only one non-matching feature is needed between a latent and known print), whether an identification is possible conditional on the content of the known print, or whether there was not enough information in the latent to make a comparison (in which case the print was not shown in a comparison). For matching pairs, only 48% of latents were unanimously agreed to contain enough information to make an identification; agreement was 33% for non-matching pairs. One source of variability in performance is therefore individual differences among expert examiners. Some of this variability may be due to different amounts of expertise, since duration and type of

[1] Of course, the visual qualities that make comparisons easier or more difficult for novices may or may not bear much resemblance to the characteristics that make prints difficult for experts. One of the experiments discussed below (Experiment 2) has relevance to this point.

training correlate with performance. Other differences may be due to lack of a standard for determining what counts as sufficient information. Without some way of measuring information content and quality, it is impossible to know what makes a comparison difficult, which comparisons actually are difficult (without relying on subjective ratings), and whether an examiner is correct in determining that there is sufficient or insufficient information to make a comparison. Similarly, Langenberg (2009) had six examiners complete 120 comparisons in two phases. He investigated overall accuracy, verification accuracy, consistency within and across examiners, as well as type of conclusion (identification, exclusion, inconclusive, or no value). The resulting performance data are interesting and informative. However, this study did not quantify what *features* of the images may have resulted in errors or disagreement among examiners. While it is important to know what the average accuracy of an average examiner may be on an average fingerprint, that was not a primary goal of our project. Rather, our effort was to identify, using objective measures, whether a particular comparison is easy or difficult and whether it is likely or not to result in an error.

A recent NIJ report has made a valuable early attempt at measuring fingerprint quality and information content (Neumann, Champod, Yoo, Genessay, & Langenburg, 2013). Almost 150 examiners evaluated 15 fingerprint pairs for information "sufficiency". Examiners who participated in the study were asked to classify, using a web-based tool, regions of the images that had low, medium, or high quality. They were also asked to mark, by hand, as many minutiae as they could find and to classify them. In addition, they were asked to make several subjective assessments of quality regarding fingerprint properties such as amount of distortion or degradation. The authors examined relationships between marked features (minutiae), perceived quality metrics, and the conclusions reached by examiners. Interestingly, there was a great deal of variation across examiners in terms of assessment of finger ridge quality, degradation and distortion, and the number of minutiae. The researchers were unable to find a quantitative measure common to all examiners that predicted whether there was sufficient information to reach a conclusion. Other features, including demographic factors, seemed to have little effect. This report underscores the problem that the features examiners selected were ultimately subjective, and therefore dependent on the idiosyncrasies of the specific examiners making the comparisons. That is, different examiners would produce different features for the exact same fingerprint image. This research, while interested in questions related to ours, highlights the importance of our project, in which we strove to identify objective (i.e., observer-independent) image features that were predictive of accuracy. The features we identified can be computed automatically for any fingerprint pair and involve neither a laborious and subjective period of minutiae marking and classification, nor the concerns that arise from any subjective process about inter-examiner consistency and reliability.

Fingerprint Features in the Standard Taxonomy

The first step in latent print examination is often manual preprocessing. For example, the region of the image that contains the fingerprint could be selected from the background and oriented upright. If a fingerprint is to be submitted to a database for automated comparison, key features need to be identified and labeled. Automated searches are then carried out by software that finds fingerprints on file with similar spatial relationships among labeled features in the submitted fingerprint. This is the only part of the examination and comparison process that is automated. The software returns a list of potential matches, some of which will likely be quickly excluded. Some will likely be closer non-matches or a match, and these require further scrutiny by an examiner.

Whether examiners are provided with potential matches via automated database searches or via investigative work, they often make their match decisions using the approach: Analysis, Comparison, Evaluation, and Verification (Ashbaugh 1999; Mnookin, 2008a). The examiner first inspects the two prints individually (analysis), then compares them relative to each other, looking for both similarities and differences (comparison). They then evaluate those similarities and differences to arrive at a decision about whether the prints match or not. In the final step, a second examiner independently validates the comparison. Mnookin points out that there is no formalized process for any of these steps. There is no method or metric for specification of which features should be used for comparison, nor any general measure for what counts as sufficient information to make a decision. Examiners rely on their experience and training rather than formal methods or quantified rubrics for making a decision. Despite the lack of a formal, standardized procedure, attempts have been made to formally describe and classify the kinds of features that might be found in a fingerprint.

Three types of features are commonly used to describe the information used for fingerprint comparison (for a complete discussion, see Maltoni, Maio, Jain, & Prabhakar, 2009). Level I features are global descriptors of ridge flow easily seen with the naked eye. The pattern in the central region (the "core") of the fingerprint can be classified as one of several common types: left- and rightward loops, whorls, tented-arches and arches. Deltas are triangular patterns that often occur on the sides of loops and whorls. A leftward loop and a delta are indicated by the yellow and green boxes respectively in Figure 1. Level I features are too common to be sufficient for identification, but they can be used for exclusion purposes as well as to guide inspection of the more detailed Level II and Level III features.

Level II features include *minutiae* such as ridge bifurcations and ridge endings. Level II features are found where fingerprint ridges and valleys split or end. *Minutiae* are highlighted in red circles in Figure 1. The uniqueness of fingerprints for identification purposes is largely due to the high variability in the existence and the relative positions of these features across fingers and individuals. Scarring, which occurs naturally with age and wear, can also add unique ridge patterns to a fingerprint. However, while scars can be used to compare the fingerprint found at a crime scene to that of a suspect in custody, they may not always exist in fingerprints on file that may predate the markings.

Level III features are the smallest fingerprint features used by some examiners for comparison. These include the positions of sweat pores and ridge thickness. Pores are indicated in light blue circles in Figure 1. The visibility of Level III features depends on the quality of the prints and examiners do not uniformly make use of them for comparison purposes.

Training may lead to an increase in the number of detailed local characteristics (*minutiae*) noticed by participants in a given print (Schiffer & Champod, 2007). With brief presentation times (under a second), when a viewer may not have enough time to compare many local features across two images in a matching task, experts utilized configural fingerprint information more efficiently than novices, focused on different information, and/or more effectively filtered out irrelevant information (Busey & Vanderkolk, 2005). What that information was, however, was not a primary focus of the research. Marcon (2009) had naïve observers rate "high quality" known prints and "low quality" latents for distinctiveness. Performance for categorizing pairs of prints as coming from the same source or a different source was higher for high-quality and high-distinctiveness images. These results suggest that performance suffers when fingerprint image quality is low, but do little to determine what makes a print low quality in the first place.

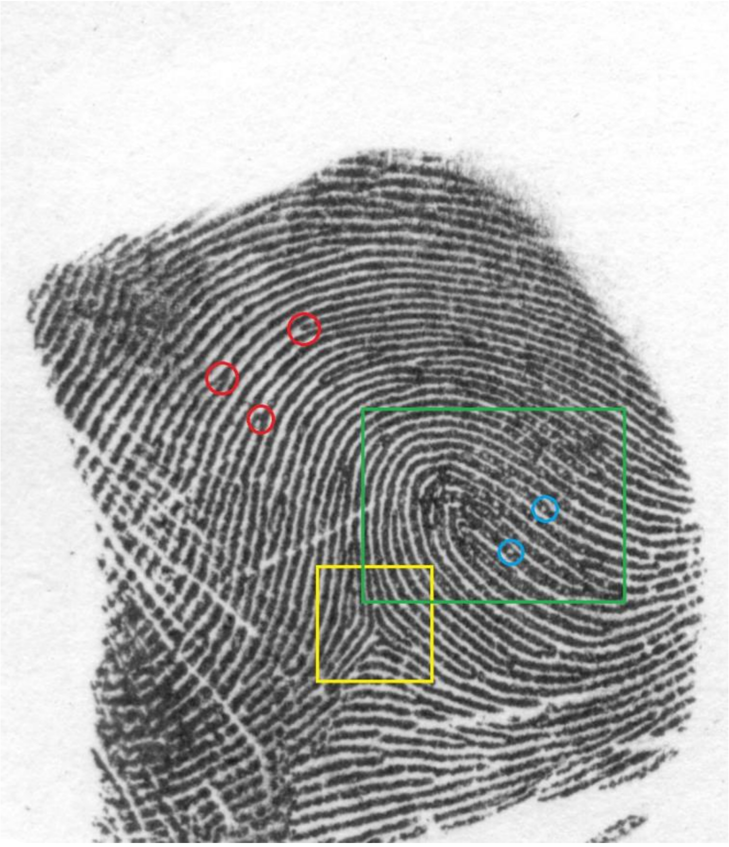


Figure 1. Depiction of various image features commonly identified by expert examiners. Red circles indicate *minutiae* (ridge bifurcations or endings); blue circles indicate pores (they appear as small white dots along a ridge); the yellow square indicates the delta; the green rectangle indicates the core, in this case a leftward loop.

Whereas the kinds of visual structures that may match or differ across fingerprints (core patterns such as whorls or loops and *minutiae*) have received some consideration, almost no analysis has been devoted to characteristics of image quality that may affect the fingerprint comparison task. These considerations apply mainly, but not exclusively, to latent prints. Intuitively, we would expect that a partial latent showing a small percentage of the full print, made on a surface unfavorable for extracting prints, and moved or smudged when the impression was made would present a more challenging matching problem than a clearer latent of larger area. For known prints, there is also variability in contrast, smudging, collection of excess media, and so forth that can affect the visual information available. There may also be relational variables involving the print pair: for example, comparing two prints of similar contrast may be easier than comparing a high-contrast known print and a low-contrast latent. Image processing measures extracted from latents and known prints, and the relations among them, may be useful for predicting the difficulty of a given comparison.

Predictor Variables

What properties of the images in fingerprint pairs are most important and informative in comparing fingerprints, and therefore most strongly predict matching performance? Although we relied on regression methods to provide answers to this question, it was important to develop, as inputs to the regression analyses, a wide variety of possible image characteristics that could be relevant. To generate such factors, we were guided by visual science, intuition, insights from fingerprint examiners, and prior work on image processing of fingerprints (e.g., Maltoni, et al., 2009), as well as the standard taxonomy of levels of pattern information in fingerprints. Some variables intuitively relate to the quantity of available information; for example, having greater print area available for comparison might make comparisons more accurate. However, this view might well be oversimplified; quality of information might matter as much or more than total print size. We created and adapted several image processing techniques sensitive to smudging, missing regions, poor contrast, etc. In short, these algorithms were used to create variables with values for each print pair that likely relate to the visual information relevant to examiner performance.

As mentioned above, we hypothesized that difficulty would be a function both of the characteristics of the individual prints (the latent and the potential match) and also of the characteristics of the *pair*. Because known prints are obtained under relatively standardized conditions, they are subject to significantly less variability than latent prints obtained from crime scenes. Accordingly, we expected that more of the variability in visual information quality affecting fingerprint comparisons would be determined by characteristics of latent prints. An especially poor quality latent might be more difficult to assess than a higher quality one, all else being equal. However, we also believed that comparison difficulty would be a function of interaction effects between the latent and the known, not simply a function of the information quality and quality of each alone. We therefore developed quantitative measures involving both individual prints and print *pairs*.[2]

A general description and motivation for the image features we selected or developed is provided below. Except where noted, we assessed each predictor variable for the latent print and the known print. For many variables, we also derived a variable that expressed an interaction or relationship of the values for the latent and known print combined (such as the ratio of latent print area to the known print area, or the Euclidean sum of contrast variability for the latent and known print combined). Details about the procedures used to derive the measures are described below.

Total Area. This variable was defined as number of pixels in the fingerprint after the fingerprint was segmented from the background. Although machine vision algorithms exist that could have been used for determining the region of usable print image, those algorithms we examined were not as good as human segmentation, and different human observers in pilot work produced strong agreement. Accordingly, we segmented fingerprints from their surrounds by having human observers designate their boundaries. In general, we expected that larger areas, especially of latent prints, would provide more information for making comparisons.

While there are a variety of automated computer algorithms to segment a fingerprint from its background (Shen & Eshera, 2003), we opted to manually segment the images because, although the automated methods we tested worked well for most known prints and high quality latent prints, they failed for many low-quality latent prints. Since many of our latent prints were intentionally low quality (e.g., low contrast, smeared, etc.), the automated approach was not adequate. Furthermore, fingerprint technicians often manually specify the region in which a fingerprint is to be found, and so manually specifying the print region was not a great departure from standard procedures (observations from Los Angeles Forensics Lab). To calculate Total Fingerprint Area, a graphic user interface (GUI) was developed in MATLAB that displayed each image, one at a time. Two of the authors segmented all images by clicking and selecting points on the boundary of the print. A

[2] One of the anonymous reviewers of the draft report made the interesting observation that to look at the characteristics of 'pairs' rather than individual prints could be seen to violate the principle of separating the analysis phase from the comparison phase, of ACE-V a separation which many fingerprint analysts adhere to, and which has been recommended by some as a method for controlling the risk of bias (Expert Working Group on Human Factors in Latent Print Analysis, 2012). It is true that assessing the comparison exemplars in conjunction does not adhere to the principle of a complete separation of these phases. However, the purpose of this separation is as a mechanism to control cognitive bias. If a metric makes use of automated, objective measures for each print, that obviates the need for separation. To whatever extent a metric incorporates subjective dimensions of measurement, the reviewer's point would indeed have purchase.

polygon was fit through the points and the number of pixels within its boundaries was used as a measure of print area. Since each print was scanned at the same resolution, number of pixels is proportional to physical area.

Area Ratio. To relate the relative area of a latent to a potentially matching known print, we divided the area of the latent fingerprint by the area of the known print. Typically the known print, obtained under controlled conditions, presents a more complete image. Thus, *Area Ratio* relates to the proportion of known print information potentially available in the latent print. However, for non-matching prints, the area of the latent may be larger than that of the known print because of differing finger sizes. Occasionally, even for a matching latent and known print, the latent could be larger than the known print due to smearing. The ratio was therefore not strictly in the range [0,1] and cannot be considered a true proportion.

Image Intensity. We measured the mean and standard deviation of pixel intensity taking into account all of the pixels in each fingerprint image (with intensities scaled in the range of [0,255]). The mean intensity and standard deviation of intensity provide two related but different measures, sensitive to different image characteristics. Very dark images (low mean intensity) might indicate the presence of large smudges that produce large, dark areas. Low standard deviation in intensity would make ridges (transitions from light to dark) difficult to detect.

Block Intensity. The image was divided into 50x50 pixel regions and the average pixel intensity was computed within each region. The mean of the block intensities is the same as the overall mean *Image Intensity.* The standard deviation of these regional averages (*standard deviation of block intensity*), however, can provide additional information about variability in image intensity across the image. Low variability is indicative of many similar areas across the image, but does not provide information about whether those regions have low or high contrast (i.e., an all black image and an image with 50% white and 50% black pixels, evenly distributed across the image, would have low *Block Intensity* variability). When pixel intensities are not uniformly distributed across the image, variability of block intensity is high (i.e., some regions of the image are darker than others). For latent images, this may indicate the presence of a smudge or worse contact (lighter impression) in some regions of the image.

Deviation from Expected Average Intensity (DEAI). Intensity, as coded above, may be a useful predictor variable, but both intuition and pilot work led us to believe that it might not capture some significant aspects of intensity variations. We therefore developed a separate intensity measure – deviation from expected average intensity. In an ideal fingerprint image, one might expect approximately half of the pixels to be white (valleys) and half to be black (ridges). The expected mean intensity would therefore be half of the range, or 127.5.[3] The absolute deviation of the observed average from the expected average was computed using the following formula:

$$DEAI = -|mean\ pixel\ intensity - 127.5|$$

Using absolute value here ensures that deviations from the midpoint of the intensity range in either direction are scored as equivalent; the negative sign ensures that the measure increases as the mean pixel intensity approaches 127.5 (large deviations produce a large negative value of the measure).

Contrast. Michelson contrast was computed for each the segmented fingerprint. Michelson contrast is defined as:

$$Contrast = \frac{Maximum\ Intensity - Minimum\ Intensity}{Maximum\ Intensity + Minimum\ Intensity}$$

[3] Ridges, on average, are thicker than valleys so the expected average would be slightly lower since there would be more black pixels than white.

This contrast measure produces a value between 0 (least contrast) and 1 (most) by dividing the difference of maximum and minimum intensity values by their sum. Michelson contrast is typically calculated from luminance values. In our images, we calculate Michelson contrast from pixel intensity values, which is appropriate given that fingerprint images may be displayed on a variety of monitors with different Gamma coefficients.

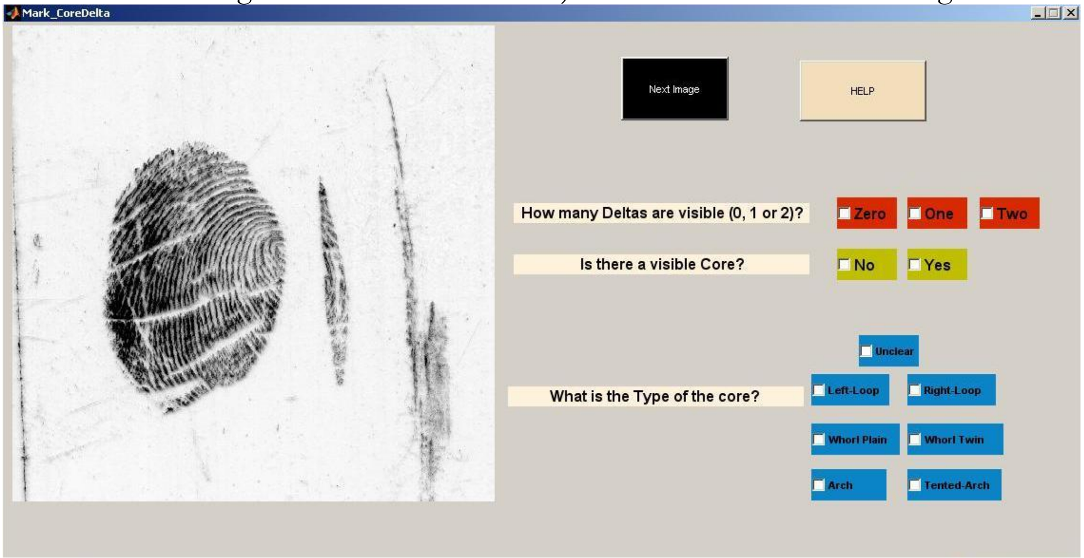
Block Contrast. The preceding measure obtained the Michelson contrast for an entire image. We also computed contrast for smaller image regions – block contrast – by segmenting the entire image into 50x50 pixel regions. *Block Contrast* is defined as the mean across the blocks. To illustrate the difference between overall contrast and block contrast, the Michelson contrast of an entire image containing all gray pixels, except for one white and one black pixel, would be 1. *Block Contrast*, however, would be very low, since most regions of the image would have 0 contrast. If black and white pixels were distributed more evenly across the image such that they appeared in each block, then *Block Contrast* would be high. High values of the measure may indicate the presence of clear ridges and valleys in many areas of the fingerprint. A separate but related predictor was the *standard deviation of block contrast* across blocks. Small standard deviation values could indicate high information content throughout the image (*Block Contrast* close to 1 everywhere) or that the image was uniformly smudged (*Block Contrast* close to 0 everywhere).

Ridge Reliability. Orientation-sensitive filters were used to detect edges in the fingerprint image. The relative responses of these filters were then used to identify "high reliability" regions where ridge orientation was uniquely specified. The proportion of high reliability regions was computed, resulting in an overall reliability score for each print. Ridge Reliability ranged between 0 and 1, with larger values indicating a greater proportion of print area with well-defined ridge orientation. An additional, relational predictor was computed by taking the Euclidean sum of the *Ridge Reliability* for the latent and known print (*Ridge Reliability Sum*). Large values of this measure indicate a high proportion of regions with well-defined ridge orientation in both the latent and known prints.

Fingerprint images were histogram equalized in blocks of 75x75 pixels to 256 gray levels. Local ridge orientation reliability was then computed for each pixel in each latent and tenprint using the MATLAB function ridgeorient.m (Kovesi, 2000). ridgeorient.m first computes the pixel intensity gradient within a 10x10 pixel region centered on each pixel. For that region, the direction of maximum change in intensity was identified. The area moment of inertia was then computed about this direction. This is the minimum moment of inertia, while the perpendicular direction is the maximum. The ratio of minimum to maximum inertia was computed and subtracted from one. If the two moments are close to each other, then the gradient in the maximum and perpendicular directions is similar, meaning that there is little variation in intensity in any direction that region of the image and it is unlikely to contain an edge. This would yield a ratio close to one, and, when subtracted from one, a reliability value close to zero. In contrast, a clear edge would produce a large difference between the minimum and maximum moments of inertia and therefore a small minimum to maximum ratio. When subtracted from one, it would yield a reliability score close to one. This code is available for download (see Kovesi, 2000). The local reliability values at each pixel were then averaged across 50x50 pixel regions. Regions in which the average reliability exceeded a threshold of 0.45 were classified as reliable. The proportion of reliable regions in the segmented fingerprint image was the *Ridge Reliability* measure. This measure is bounded between 0 and 1 and corresponds to the proportion of the area of each print that contains reliable ridge orientation information.

Visibility of Cores and Deltas. Earlier we described global configurations – *Cores* and *Deltas* – that provide Level I information to fingerprint examiners. The fact that ridge flow in fingerprints tends to follow a circular pattern dictates that there will be some global core (a whorl, loop, or arch) at or near the center of each print. Likewise the transitions from global cores, especially loops and whorls, to the circular ridge flow tends to give rise to deltas, triangular configurations (see Figure 1). As there will be only one core and at most a small number of deltas in any print, these serve as important reference points in making comparisons (Maltoni, et al., 2009). Unlike all of the other variables we used, which could take on a continuous range of values, *Cores* and *Deltas* are binary (either present or not).

A MATLAB-based GUI was developed and used by one of the authors to count the number of deltas present and whether or not the core was visible. Each print was also classified as left loop, right loop, whorl plain, whorl twin, arch or tented arch (or "unclear" if insufficient information was available for making a definitive classification). This interface is shown in Figure 2.



How many Deltas are visible (0, 1 or 2)? □ Zero □ One □ Two

Is there a visible Core? □ No □ Yes

What is the Type of the core? □ Unclear □ Left-Loop □ Right-Loop □ Whorl Plain □ Whorl Twin □ Arch □ Tented-Arch

Figure 2. The interface used for counting deltas, marking the presence or absence of a core, and labeling the core type.

Relations Among Basic Predictors

To remove effects on regression coefficients of differing scales of various predictors, we standardized all continuous metrics by subtracting the mean and dividing by the standard deviation. Standardization made some measures that were strictly non-negative (like *Standard Deviation of Intensity*) take on negative values. As is often recommended in using regression methods (e.g., Neter, Kutner, Wasserman, & Nachtsheim, 1996), we also examined the features for collinearity and found that several predictors were highly correlated. For example, the mean and standard deviation *Intensity* measures were correlated (Pearson's r = -0.77 for latents and -0.44 for known prints). High correlation among predictors is an undesirable feature for regression models (Neter et al. 1996) because it makes it harder to assess the individual effect of those predictors. If two predictors had a correlation of greater than 0.5, we removed one of them. After removal, the variance inflation factor, a measure of collinearity, for all continuous metrics was less than 5, indicating that collinearity was sufficiently reduced (Chatterjee & Price, 1991; Booth, Niccolucci, & Schuster, 1994; Neter, et al. 1996).

In addition, we included two-way interactions between all predictors that applied to both a latent and known print. For example, in addition to the *Standard Deviation of Block Contrast* for the latent and known print, we included the interaction between the two terms. In addition to *Area Ratio* and *Ridge Reliability Sum*, these are *relational* predictors that encode something about the relative quality of information in a latent and known print.

Present Studies

The main goal of the project was to identify and quantify fingerprint image features that are predictive of identification difficulty and accuracy. Low quality prints recovered from crime scenes are often distorted, smudged, or contain only partial impressions. Experts may disagree whether the prints contain enough information or are of sufficiently good quality to make a determination (identification or exclusion / match or non-match judgment) or whether there is not enough information (i.e., the print is of insufficient quality) to make a judgment. It would be useful to have a quantitative way of assessing fingerprint image quality and comparison difficulty. Such a metric can be used to alert examiners when additional care is warranted (i.e., for a particularly difficult comparison), to caution examiners who are inclined to label a print pair as inconclusive that further examination might be prudent, and to create a set of fingerprint images with objective difficulty ratings that can be used for training examiners.

If fingerprint comparisons are generally accurate but occasionally not so, characterizing the sources of difficulty and the quality of information in fingerprint pairs becomes crucial. Ultimately, it may be possible to evaluate a fingerprint comparison in terms of the quality of visual information available in order to predict likely error rates in fingerprint comparisons. Such a metric would have great value in both adding confidence to judgments when print comparisons are uncomplicated in terms of having high quality visual information, and it would allow appropriate caution in cases that are, from an objective standpoint of the quality of visual information, more problematic.

In a typical fingerprint evaluation, an expert examiner is given a latent and a known fingerprint pair, which they evaluate for identity (i.e., whether the two images came from the same finger or not). For the present study, we created a database of latent prints, matching known prints, and non-matching prints retrieved from an AFIS database. Comparisons involving prints retrieved by AFIS resemble those performed in realistic settings where candidate matches are also generated by AFIS. Since the system attempts to find similar prints, the comparisons in our study may reveal error rates that are higher than that would occur if the non-matches were randomly selected, but would be more representative of real-world comparisons.

In all studies reported here, participants performed a two-alternative forced choice task in which they evaluated whether two fingerprint images came from the same source (matched) or from different sources (did not match). The latent print was not a cropped version of the known print; rather, the two prints were retrieved in different instances and the task was to determine whether the same finger created both. Images were presented side-by-side on the computer screen.

In addition to creating a fingerprint image quality metric, the project had several other objectives: (1) To create a realistic fingerprint image database with known ground truth about each pair (i.e., true matching prints and relatively close non-matches). (2) To add to the rather modest, albeit growing scientific literature investigating fingerprint expertise; that is, to assess objectively expert performance in terms of both accuracy in fingerprint identification and the image characteristics that related to performance. (3) To compare expert performance with novice performance, and in that manner quantify the degree of expertise. An additional benefit of studying novices was our ability to study how performance changes when a group of novices was made familiar with some characteristic visual features of fingerprints through brief training and to investigate how and to what extent this training changes cognitive strategies by altering the relative importance assigned to various visual characteristics in fingerprint pairs. Although not an original goal of the project proposal, this was a natural extension and resulted in an interesting finding.

We created a database of matching and non-matching fingerprint pairs that were used in all studies described in this report. The details of the database are described in the following section. Fingerprints were chosen to be a realistic example of the kinds of prints that are normally found in evaluation settings. Care was taken to attempt to generate pairs that varied in difficulty. First, this was important in order to ensure that sufficiently difficult comparisons were included to try to simulate difficult comparisons in the real world and potentially generate errors. Second, a range of difficulty allows for the database to be used in other settings, for example, as a training set from which examiners can select, easy, medium, and difficult comparisons. We used this database for several experiments reported below.

In Experiment 1, fingerprint examiners recruited from a forensics conference made match/non-match comparisons for a subset of print pairs from the database. There were several important differences between the experiment and typical comparison settings. Normally, examiners have the choice to label a print pair as "inconclusive", which means that the examiner deems that there is not sufficient information available to unambiguously say whether two prints come from the same finger or not. This creates the possibility of a different kind of error from saying that two non-matching prints are from the same person (false alarm) or saying that two prints from the same person are from different people (miss): incorrectly deeming that there is not enough information to make a conclusive evaluation when there is in fact sufficient information. In all experiments, we asked participants to provide difficulty and confidence ratings for each comparison. While this procedure is different from the operation of fingerprint analysis in normal forensic settings, it has two important advantages. Firstly, errors in this forced-choice framework likely have a more direct relationship to fingerprint quality. Second, we were able to examine the relationship between fingerprint information quality and confidence. This experiment was an important first proof-of-concept to demonstrate that under at least restricted settings, errors could be made. If it had turned out that experts made no mistakes given the constraints of the task, then there would have been little hope of artificially creating other situations in which errors could occur. To foreshadow some of the findings, several features of the fingerprint images were found to correlate with performance.

In Experiment 2, we used an overlapping subset of the fingerprints to test performance of novices. This served as a baseline comparison for examiner expertise. We expected that novices with absolutely no training would perform very poorly at this task since expertise requires extensive practice, in a same way that an amateur would have difficulty in classifying birds or determining whether an x-ray image contained evidence of cancer. However, novice performance seems to vary greatly depending on the type of study and materials, and can be as high as 75% for matching prints (e.g. Vokey, Tangen, & Cole, 2009; Tangen et al., 2011). It was therefore important to get performance measures for this particular set of images. One group of novices provided this performance baseline. A second group was shown a brief video that highlighted the kinds of image features used by experts in fingerprint matching and explaining how one might go about comparing fingerprints. It would be unreasonable to expect that watching a short video would drastically improve performance (otherwise, experts would not need such extensive training); however, the training video might cause novices to begin to use and be affected by the same information that experts use. We hypothesized that we might find that the kinds of features that were important for accurate performance for experts might receive a greater weighting or become more important for novices who watched a short video. For example, if it was pointed out that *minutiae* or ridge flow could an important factor in determining fingerprint identity, then perhaps measures like *Ridge Reliability* would become more important (meaning that performance would be higher for prints with greater values of this predictor) for the task. By comparing what predictors correlate with accuracy between experts and novices, we can examine differences between the two groups and identify which features may be most important to focus on.

Experiment 3 was an extension and validation of Experiment 1 with a different set of experts and an expanded set of tools in a substantially more realistic setting. Participants had access to a wide range of image processing tools to manipulate the images in the study, via an interface we built. They had unlimited time to make their comparisons. They also had the option of indicating that a particular comparison should have been deemed inconclusive. However, they still had to provide difficulty and confidence ratings, as well provide a best guess as to whether the prints were a match or non-match. The prints used in this study were selected based on predictions of difficulty from Experiment 1. Some of the prints in Experiment 3 were used in Experiment 1 and some were new. We expected to find generally comparable performance in this new group of experts to those tested in Experiment 1, but we did not know the extent to which the other manipulations (additional time and tools) would impact performance. If there was no difference in performance, then, moving forward, that would suggest that findings from experiments using simplified testing materials might be able to be extrapolated to more realistic settings; if there were very substantial performance differences, that would show that some of the simplifications of the sorts we took dramatically altered performance. We found a high correlation in accuracy for print pairs that were used in both Experiment 1 and Experiment 3, although there were some differences. The model was successful in predicting accuracy for many print pairs in Experiment 3, despite the differences between the two studies. There were several inconsistencies in predictions, however, particularly for print pairs that were labeled as inconclusive in Experiment 3. We explore some of the implications of these results and suggest further analyses and studies.

II. Methods

Database Creation

Fingerprints were collected from 103 individuals. Each individual first used a single finger to produce a clear, known print using ink as is done in police stations. Then, using the same finger, they touched a number of surfaces in a variety of ways (with varying pressure, smudges, etc.), to create a range of latent fingerprint marks that reflect those found in a crime scene. Professional fingerprint examiners who participated in the study reported that these prints were similar to those that they encounter in their everyday casework. The latent fingerprints were lifted using powder and were scanned at 500 dpi using the FISH system. Image dimensions ranged from 826 pixels in height to 1845 pixels and from 745 pixels in width to 1825 pixels. The latent prints that were created varied in clarity, contrast, and size. For each individual who contributed to the database, we collected a total of six prints – one known print and five matching latent prints. Across individuals we varied the fingers used. Each scanned fingerprint was oriented vertically and approximately centered.

To create the non-matching pair of prints, we did not want to randomly choose a known and a latent, as such pairs may be too obviously different. This would make the "non-match" decisions nearly uniformly easy, and would also, by default, indicate which were the "matching" pairs. Therefore, we obtained similar, but non-matching known prints by submitting the latent prints to an AFIS search. An expert selected from the AFIS list what he deemed as the most similar print. That enabled us to produce non-matching pairs that were relatively similar. The final database consisted of 1,133 fingerprint images – five latent prints from 103 fingers (515), 103 known prints that matched (103), and another 515 known prints for the non-match for each of the latents. Since we used an AFIS with a database from another country, it was practically impossible that a match would be in the database. Furthermore, the expert who selected the most similar print from the AFIS candidate list verified that each was a similar print, but not an actual match.

Experiment 1 – Experts at Conference

Subjects

Fifty-six fingerprint examiners (18 male, 35 female, three not reported) participated in the study. Forty participants self-reported as latent print examiners, three as known print examiners, ten as both, and three did not report. Years of experience were reported between the range of 1 and 25 years (Latent: Mean = 9.54, SD = 6.97 ; Ten-Print: Mean = 10.45, SD = 8.07). Twenty-seven participants reported being IAI (International Association for Identification) certified. 32 reported that their labs were accredited.

Participants were either directly recruited at the 2011 IAI Educational Conference or via a flyer sent out in advance of the conference. As incentive, participants were told they would be entered into a raffle to win an iPad 2. All participants signed informed consent forms prior to participating. As indicated above, some limited demographic information was collected, but it was stored separately from individual participant IDs such that the two could not be linked.

Apparatus

All stimuli were displayed on laptop computers with 17-inch monitors at a resolution of 1024 x 768 pixels. Stimuli were presented using a program accessed online; data were stored on the website's server. A screenshot of the program is shown in Figure 3.

Stimuli

Of the 1,133 fingerprint images, 200 latent and known print pairs were selected and used for the study; half were a match and half were a close non-match. Individual print metrics were computed for each image or image pair (see below) and prints were selected to (approximately) uniformly sample each feature space. Known prints were sampled without replacement, but multiple latent prints from the same finger were occasionally selected since each latent could be paired with a different known print image (the match or a close non-match). Print pairs were then grouped into batches of 20, each containing ten matches and ten non-matches. Latent prints from the same finger did not appear within the same batch.

Design

A group of experts made match / non-match judgments and provided confidence and difficulty ratings on a subset of 200 print pairs selected from a database of over a thousand fingerprint images. Two fingerprint images that were either from the same finger (match) or from two different fingers (non-match) were presented side-by-side. Images were presented on computer screens and were always oriented upright. Examiners had a maximum of three minutes to evaluate each pair of images. Performance was recorded for each print-pair tested.

Procedure

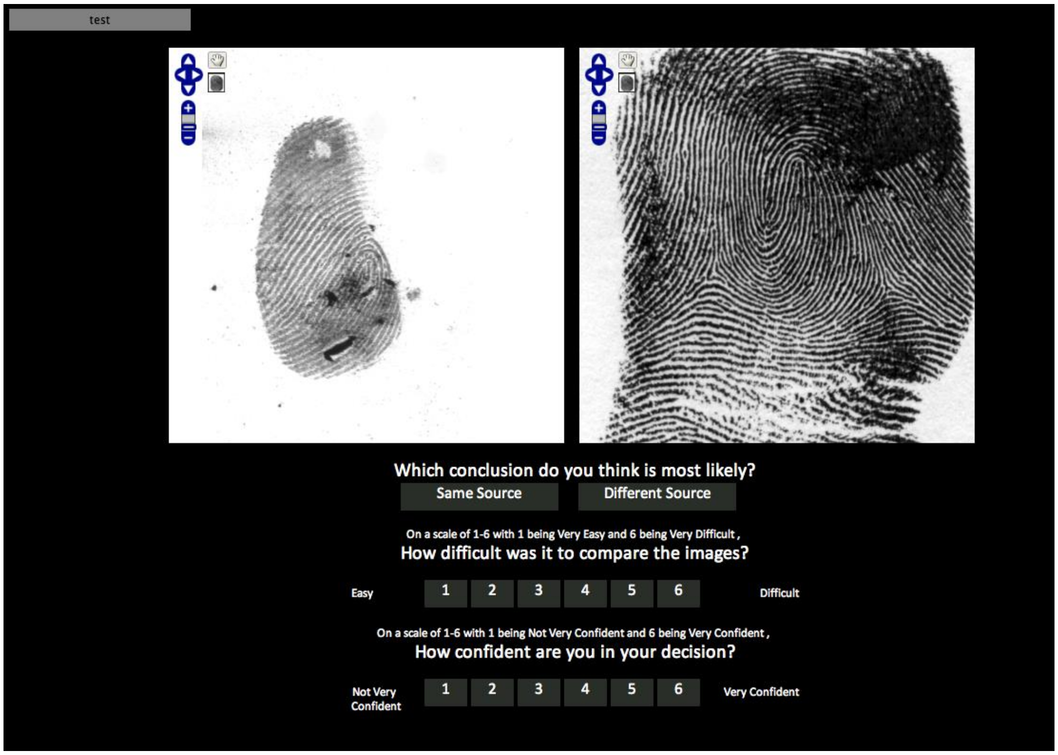
Participants were tested in a large room, seated at desks with individual laptop computers. Before data collection began, each participant was asked to sign a consent form, and then given written instructions detailing how the stimuli would be presented and the judgments they would be required to make. Participants were told that they would be asked to compare latent-known print pairs and determine whether they were matches or non-matches (without the option to choose "inconclusive" as a response). Participants were also told that they would be asked for confidence and difficulty ratings for each of their judgments. The instructions emphasized that this procedure was not intended to replicate real-world conditions and that participants should simply try to maximize accuracy. Participants were also instructed to refrain from using any fingerprint examiner tools not provided by the experimenter, such as a compass.

When the experimental program was initiated, participants were asked to report their age, gender, years of experience, specialization, IAI certification, lab accreditation, and lab affiliation. Reporting this information was optional.

Next, the experiment began. On each trial, two fingerprints were presented side-by-side. The latent print was always on the left. A button in the top-left corner of each image window allowed participants to zoom in on each image individually. Fingerprint size was constrained within the bounds of each window, so that each print was always viewed through an aperture of 460 pixels by 530 pixels. The initial presentation of the images had them scaled to fit entirely in this window. A single level of zoom allowed participants to magnify the image. Participants could also translate each image independently within its window (both when the image was zoomed or unzoomed) either by dragging it with mouse or by using arrow buttons in the top-left corner of each image window. No other image manipulation features were available.

Participants made a match/non-match judgment by clicking a button at the bottom of the screen. Specifically, participants were asked: "Do these prints come from the same source or a different source?" Participants then made difficulty and confidence ratings by clicking on a Likert scale. The participants were asked: "How difficult is the comparison?" and "How confident are you in your decision?" On the Likert scales, "1" corresponded to least difficult / least confident and "6" corresponded to most difficult / most confident. Once all responses were recorded, an additional button appeared allowing the participant to advance to the next trial. Figure 3 shows a sample screenshot of the experiment.

Participants had three minutes to complete each trial. A message was given after two and a half minutes warning them that the trial will end in 30 seconds. If the full three minutes elapsed without a decision, the trial was ended, and the participant moved on to the next trial. After presentation of a set of 20 print pairs, participants were given a short break and asked if they wanted to complete another set of 20 comparisons.



Which conclusion do you think is most likely?

On a scale of 1-6 with 1 being Very Easy and 6 being Very Difficult ,
How difficult was it to compare the images?

On a scale of 1-6 with 1 being Not Very Confident and 6 being Very Confident ,
How confident are you in your decision?

Figure 3. Screenshot of a sample trial from Experiment 1. Examiners could use the keys in the windows to change position or zoom level. Responses were made by clicking on the buttons shown in gray. Once all responses were provided, a button appeared allowing the user to advance to the next trial.

Each set of 20 print pairs contained ten match and ten non-match comparisons. The order in which print pairs were presented within a set was randomized across subjects. The sets were presented in a pseudo-random order so that approximately ten participants completed each set. Although the number of trials completed by individual participants varied based on their availability and willingness to do more comparisons, most participants completed two sets of prints (40 print pairs).

Experiment 2. Novices

Subjects

Participants were undergraduate students at University of California, Los Angeles, who participated in the experiment for partial course credit. 36 novices were randomly assigned to either the "no training" or the "training" groups, with 18 subjects per group.

Apparatus

All stimuli were presented using Matlab and routines from the Psychophysics Toolbox (Brainard, 1997). Displays were presented on one of three 16" x 12" ViewSonic Graphic Series G225f computer monitors in the UCLA Human Perception Laboratory, each with a resolution of 1024 x 768 pixels and a refresh rate of 60 Hz. The observer sat approximately 40 cm from the screen. Participants responded using a keyboard.

Stimuli

Fingerprint pairs (latents and known prints) were selected from the fingerprint database. As in Experiment 1, latents could be paired with a corresponding known print (match) or with a close non-match retrieved from AFIS (see Database Creation).

Subjects in the training group watched a 5 minute video ("How to Compare Fingerprints – The Basics") before beginning the experiment. The video described the fingerprint comparison process, identified cores and deltas and how they could be used in fingerprint matching, as well as other fingerprint features, such as *minutiae* and ridge counts. A sample comparison was performed in which *minutiae* were used to match two fingerprints.

Design

One hundred print pairs were selected randomly from the database. Each print pair came from a different individual. Half of the print pairs were matches and half were non-matches. Based on pilot data, novices went through comparisons fairly rapidly and could complete all 100 in approximately 40 minutes.

Prints were displayed side-by-side with the latent print always on the left-hand side of the screen and known prints on the right. Images were large, approximately 6 inches x 6 inches in size, although the size of the fingerprint within each image varied. Fingerprints were roughly 4 inches x 5 inches. The presentation order of comparisons was randomized across participants.

Procedure

Subjects sat at desks with computers in a well-lit room. On each trial, two fingerprints were presented side-by-side. The latent print was always on the left. Subjects responded whether the two prints were the same or different by pressing the Y or N keys on the keyboard. Each participant also made difficulty and confidence ratings. Participants were asked: "How difficult is the comparison?" (with 1 as easiest and 6 as most difficult) and "How confident are you in your decision?" (with 1 as least confident and 6 as most confident). Subjects responded by pressing a number key on the keyboard. Once responses to all three questions were entered, the participants could proceed to the next trial by a key-press. Participants were instructed that they should try to maximize accuracy. No other fingerprint examiner tools (e.g., a compass) were made available.

For the training group, subjects first watched an approximately 5-minute long YouTube video describing the fingerprint comparison process. Novices who received no training immediately started the experiment without watching any video.

The study began with a practice session of 6 comparisons on which subjects received feedback (correct or incorrect). After making a match response and submitting confidence and difficulty ratings, the two print images from the trial were shown again on the screen with the feedback printed above them to allow subjects to re-examine the images.

Experiment 3 – Experts with Advanced Tools

Subjects

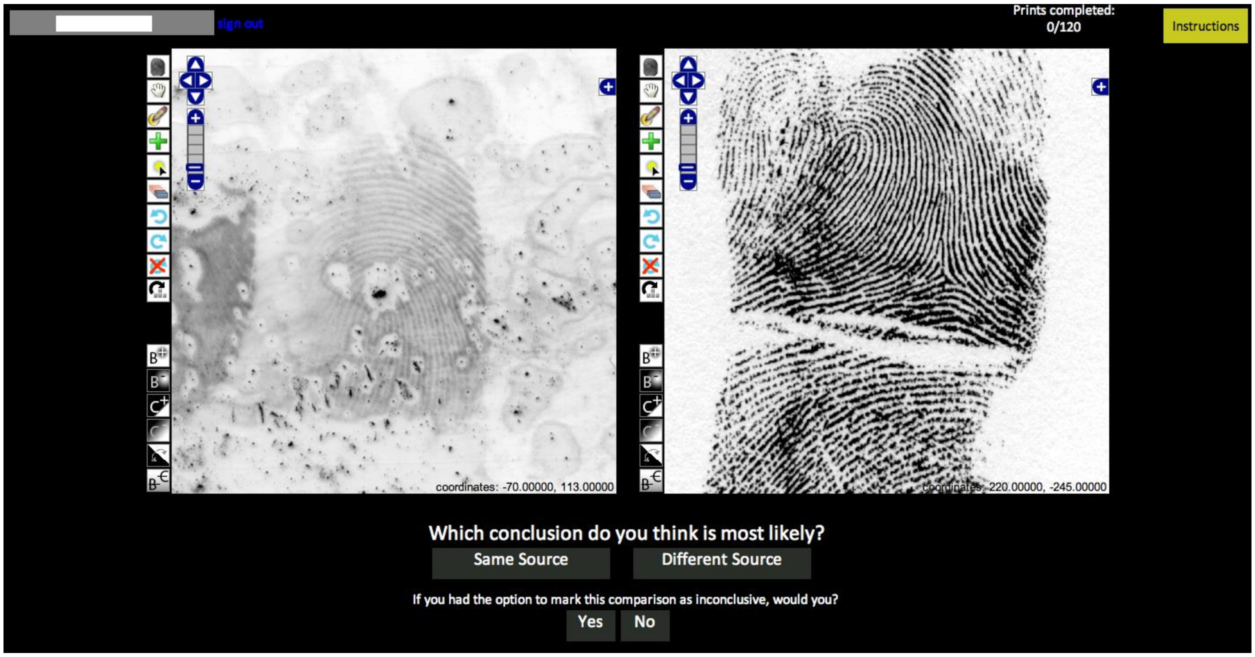
Thirty-four examiners (16 male, 18 female; age range: 29-62), were recruited via personal contact. Twenty identified as latent examiners, one as a tenprint examiner, and 13 as both. Years of experience for latent examiners ranged from 1 to 36. Eight reported being IAI accredited. Thirty reported as coming from accredited labs or offices.

Apparatus

The experiment was conducted over a specially designed website that was a modification of the one used in Experiment 1. The basic structure was the same, including a login screen, consent form screen that included an electronic signature and a link to a downloadable pdf document that contained the consent information, a demographic form sheet that was optional, and the actual experiment page that displayed two fingerprint images. The welcome screen also included a password and login section. Passwords were e-mailed to users individually during recruitment and they were allowed to generate their own login names. Users were able to re-login as often as they liked and their progress was saved across sessions. The instruction screen was greatly expanded to include participation guidelines, system requirements, and image manipulation button control. All of these are described below.

Because subjects were allowed to complete the experiments remotely, no information about monitor size is available. In the instructions, users were asked to ensure that their monitors had a minimum resolution of 1200x720 pixels. Users were asked to click on a calibration button to adjust resolution and monitor brightness and contrast settings. Four shapes were shown and users were asked to adjust monitor resolution until all appeared to have equal side lengths with no distortion (pixel height and width should be equal). A brightness bar with 32 levels from black to white was shown below. Instructions stated that all 32 colors on the bar should be visible, with equal steps from bar to bar. In particular, users were instructed to adjust monitor contrast and brightness if the darkest bar was not seen or if there was a very large change in brightness between the final two bars. Users were expected to make these adjustments on their own. No feedback was provided and no measurements were taken by the website. Users were also instructed to use an up-to-date browser from among the following list: Firefox, Chrome, Safari, or Opera.

Unlike Experiment 1, the website had added functionality meant to reproduce some of the image processing features typically available to examiners in actual practice. Each fingerprint image (both the latent and known print) had a toolbar on the left-hand side with a 16 buttons. In addition, a navigation cross appeared within the boundary of the image that enabled users to pan across the image (up, down, left, or right) by clicking on the arrows of the cross. A zoom bar was located directly below it that allowed someone to step through four levels of zoom. All images began maximally zoomed out. A screenshot of this design is visible in Figure 4.



Which conclusion do you think is most likely?

If you had the option to mark this comparison as inconclusive, would you?

Figure 4. Screenshot of a sample trial from Experiment 3. The rest of the screen that included questions about difficulty and confidence ratings is not shown in this figure. Those questions would appear lower down on the page. The general layout was similar to that used in Experiment 1. In addition to the basic tools available in Experiment 1, an additional toolbar appeared to the left of each image to allow independent image manipulation. Hovering over an icon in the toolbar caused some hypertext to appear that described the tool. Clicking on the instructions button in the upper-right-hand corner provided access to detailed descriptions of each tool. A progress count indicating how many comparisons were completed appeared to the left of the instructions button.

A description of the image manipulations available in the toolbar appears below:

1) Zoom out completely (revert to original zoom level).
2) Free pan. By clicking on this button, a user would be able to manipulate the region of the fingerprint image that appears in the viewing window by clicking and dragging the image allowing different areas to become visible in the window. When maximally zoomed out, the entire image fit within the window.
3) Place new markers (on/off). Clicking on this button enabled marker placement. By default, the markers were green circles, however, the marker symbol and color could be changed. Markers remained in the correct positions on the image through zooming, panning, and rotation.
4) Change marker symbol. There were 5 marker symbols that could be used: circles, stars, crosses, triangles, and lightning bolts.
5) Change marker color. There were 4 marker colors: light green, dark green, red, and blue.
6) Erase marker. After clicking on this button, when the cursor hovered over a marker on the image, that marker was highlighted. Clicking on the marker removed it from the image.
7) Rotate image left (counter-clockwise) by 15 degrees.
8) Rotate image right (clockwise) by 15 degrees.
9) Reset rotation to original orientation.
10) Free rotate. Enable rotation one degree at a time by pressing the up and down keys.
11) Increase brightness of the image.
12) Decrease brightness of the image.
13) Increase contrast of the image.
14) Decrease contrast of the image.

15) Invert brightness of the image (black to white).
16) Undo all brightness and contrast manipulations (revert to values of original image).

An instructions button appeared in the top-right corner of the display. That could be used to re-access the screen shown at the start of the study. Image manipulations were not saved across login sessions. For example, if a user added markers to an image, but did not make a comparison judgment by clicking on the appropriate button and exited the experiment by closing the browser or logging out, the markers would not be visible when they logged back in.

Stimuli

Fingerprint pairs were selected from the database with the following constraints. Half of the examples came from Experiment 1. A subset was chosen that spanned a range of accuracies and included an equal number of matches and non-matches. This allowed us to validate the model on a new set of subjects. The regression model was then used to predict performance for the remaining print pairs that were not used in Experiment 1. Print pairs were selected to have a range of predicted accuracies from this set. In total, 120 pairs were chosen, 60 from Experiment 1 and 60 new pairs. Each set of 60 was composed of half matches and half non-matches.

Design

The design was similar to Experiment 1. Participants performed a two-alternative, forced choice (2AFC) task to determine whether two fingerprint images were a match (came from the same source) or not. There were 120 fingerprint pairs in the experimental set. The order in which they were presented was randomized across participants. Print images were presented side-by-side with the latent always on the left and the known print always on the right (see Figure 4). Each image had a manipulation toolbar to the left that allowed for a variety of image manipulations to be performed (see Apparatus section for details).

Procedure

Subjects were emailed a link to the experimental website as well as a password to access the site. On the welcome page, subjects read a brief description of the study and were asked to generate a username to use for accessing the site across sessions. Once a username and their password were entered, users were provided with a link to a pdf of the consent form and were instructed to download and read the form and to provide an electronic signature on the website. The following screen prompted users to provide some optional demographic information similar to Experiment 1. The next screen showed instructions for the experiment, the purpose of the study, and a description of the image manipulation tools as outlined above. The icon for each tool was shown, along with a description. Once the instructions were read, users continued to the actual experiment. Users made one comparison at a time and had to respond to all questions before continuing, similar to Experiment 1. Users were asked if the two images came from a same source or a different source and were asked to provide a difficulty and confidence rating for the comparison, similar to Experiment 1. In addition, and in distinction from Experiment 1, after making an identification or exclusion response, users were asked whether they would have marked the pair as inconclusive. If users logged out in the middle of a trial, the same trial would resume when they logged back in. However, all image manipulations that they had performed up to that point were reset. On a trial, users could adjust each of the two images independently with the manipulation tools. In addition, in this experiment, there were no time limits on each trial; examiners could take as much time as they wished. In all other respects – apart from the additional image manipulation tools, the absence of time limits, and the inquiry into whether the examiner would have selected 'inconclusive' if that had been an option – the procedure for this experiment was identical to that of Experiment 1.

Analysis Methods

Data Preprocessing

For the first experiment, if the examiner made a match/non-match judgment, but time expired before they could make difficulty or confidence ratings, the data were retained. If only difficulty and confidence ratings were provided, but a comparison judgment was not made before time expired, the trial was excluded from the analyses. Twenty such trials were excluded from the total of 2,312 comparisons (fewer than 1%) in Experiment 1. No other special preprocessing steps were undertaken for any other studies.

Descriptive Statistics and Correlations

For each experiment, average performance was computed for all comparisons, separately for match and non-match trials, separately for each individual, and for each print pair. Average difficulty, confidence, and response time ratings were also computed for each print pair and subject. Correlations were computed between accuracy, difficulty and confidence ratings, and response time. In Experiment 3, data were split by what tools were used and whether a print pair was rated as inconclusive or not.

Regression Analysis

We fit a crossed, logistic regression model in which print pair performance (1 = accurate; 0 = inaccurate) was crossed with expert and print identity. This is a type of mixed-effects model and is appropriate for analyzing these data for several reasons (Breslow & Clayton, 1993; Baayen, Davidson, & Bates, 2008). First, not every subject evaluated every print pair. A mixed-effects approach enables the examination of both the predictor variables and the "random effects" due to inter-subject differences (i.e., differences between expert performance and differences between evaluations of the same print pair by multiple experts). Second, a mixed-effects approach allows one to model individual item differences by fitting data from individual trials instead of aggregating across all presentations of an item (Dixon, 2008; Jaeger 2008). Differing levels of expertise and experience, as well as differences in comparison strategy and decision thresholds, could give rise to variability in participant performance independent of the fingerprint features. Variability across items could occur if some comparisons were easier than others irrespective of differences in measured image features. Including these sources of variability in the model allows us to test whether print comparisons and experts differed from one another, instead of assuming they were all equivalent and simply averaging across participants and items. Data were fit using the "arm" (Gelman & Su, 2013) and the "lme4" (Bates, Maechler, & Bolker, 2012) R packages for R version 2.15.2.

For each of i print-pair comparisons (items) and j experts (subjects), we define $y_{i,j}$ as

$$y_{i,j} = \begin{cases} 1 & \text{if print} - \text{pair } i \text{ is accurately classified (correct identification or rejection) by expert } j \\ 0 & \text{if print} - \text{pair } i \text{ is inaccurately classified (false alarm or miss) by expert } j \end{cases}$$

Accuracy for any particular print pair and expert, $\Pr(y_{i,j} = 1)$, was modeled with a logistic regression:

$$\Pr(y_{i,j} = 1) = \text{logit}^{-1}(X_{i,j}\beta + \text{printID}_i + \text{expertID}_{j}), \text{ for } i = 1, \ldots, 200, \;\; j = 1, \ldots, 56 \qquad (1)$$

where $X_{i,j}$ is a vector describing the features measured on a print pair, β is a vector of coefficients (the fixed effects; one coefficient for each feature), expertID_j is the expert-specific random effect, which allows the intercepts to vary across experts, and printID_i is the item-specific random effect. expertID and printID were normally distributed.

The regression equation can be rewritten and expanded as:

$$\text{logit}(\Pr(y_{i,j} = 1)) = \beta_0 + \text{x}_{1\,i,j}\beta_1 + \text{x}_{2\,i,j}\beta_2 + \ldots + \text{x}_{n\,i,j}\beta_n + \text{printID}_i + \text{expertID}_j \qquad (2)$$

where n is the number of predictors. In this form, it can be seen that printID and expertID can be grouped with β_0 as intercept terms. Because printID and expertID are vectors, the equation reflects that each combination of print and expert has its own intercept term. It is this combined term (β_0 + printID$_i$ + expertID$_j$) that varies across experts and items. Multi-level modeling allows one to capture possible differences between individual subjects or test items without fitting a separate regression equation for each item (by applying a distribution over the terms that vary, in this case printID and expertID; see Gelman & Hill, 2007).

The parameter expertID is defined as:

$$\text{expertID}_j \approx \frac{\frac{n_j}{\sigma_\mu^2}}{\frac{n_j}{\sigma_\mu^2} + \frac{1}{\sigma_{\text{expertID}}^2}}\left(\bar{y}_j - \beta\bar{x}_j\right) + \frac{\frac{1}{\sigma_{\text{expertID}}^2}}{\frac{n_j}{\sigma_\mu^2} + \frac{1}{\sigma_{\text{expertID}}^2}}\mu_{expertID}$$

Where n_j is the number of print-pairs evaluated by expert j, σ_μ^2 is the within-expert accuracy variance, $\sigma_{\text{expertID}}^2$ is the variance among the average accuracies of different experts, $\bar{y}_j$ is the average accuracy for expert j, and $\mu_{printID}$ is the overall average accuracy across experts. From this equation it can be seen that expertID is a weighted average between the individual estimates of the intercept for each expert $\left(\bar{y}_j - \beta\bar{x}_j\right)$ and the average intercept across experts, $\mu_{expertID}$. When $\sigma_{\text{expertID}}^2$ is small, the right-most term dominates and the model approaches a regular regression model with a single intercept for all experts. When $\sigma_{\text{expertID}}^2$ is large, greater weight is placed on individual intercepts, and it is as if there is a separate regression model for each expert's data. expertID$_j$ is therefore a pooled estimate of the intercept term for each expert, taking into consideration across-expert differences in performance.

Each expertID is assigned the probability distribution

$$\text{expertID}_j \sim \text{N}(\mu_{\text{expertID}}, \sigma^2_{\text{expertID}}), \text{ for } j = 1, \ldots, 56$$

with the parameters of the distribution estimated from the data. One can see from this distribution that it has the effect of pulling the overall intercept closer to the average accuracy (μ_{expertID}) if there is little variability among experts (when $\sigma^2_{\text{expertID}}$ is small), and pushing toward individual regression equations for each expert when variance is large. The ratio of individual (within-examiner) and group (across examiners) variances is the intraclass correlation. It is defined as:

$$\frac{\sigma_{\text{expertID}}^2}{\sigma_{\text{expertID}}^2 + \sigma_\mu^2}$$

When the intraclass correlation is close to 0 ($\sigma^2_{\text{expertID}}$ is small and σ^2_y is large), it indicates that differences between examiners contribute little to accuracy. Intraclass correlations close to 1 (large $\sigma^2_{\text{expertID}}$ relative to σ^2_μ) indicates that group differences contribute a lot to accuracy and that there is little variability within groups. Defining expertID$_j$ in this way allows the model to incorporate potential individual differences among experts. printID$_i$ is defined in a similar way.

Individual differences amongst experts may arise due to differences in experience, training, and other factors. These could manifest as different baselines of performance, or intercept terms in the model. All else being equal, one expert might do better with the exact same print pair than another expert. This variability is captured by the expertID term in the model. It is also possible to model item-specific (in this case, print-pair-specific) effects; these are represented by printID. printID captures differences in print comparison difficulty inherent to individual print pairs and not related to the features used to predict print pair accuracy. In constructing a model, it is assumed that the error terms are uncorrelated; however, it is possible that print pair errors are correlated across participants. Inclusion of the item-specific term captures this potential non-independence (Baayen et al., 2008).

The regression model gave a predicted accuracy for each fingerprint pair. This was compared to the average observed accuracy. Model performance was measured as root mean squared error given by the following equation:

$$RMSE = \sqrt{\sum\nolimits_{print\ pairs} (observed\ accuracy - predicted\ accuracy)^2}$$

RMSE values close to 0 indicate close agreement between observed and predicted accuracy across many print pairs; values closer to 1 indicate a poor fit. Further, we plotted observed versus predicted accuracy, fit a straight line to the data points, and computed R^2, a measure of linear fit.

Model Validation

In addition to creating models of accuracy, we fit similar models to difficulty and confidence ratings and response time data. Overlap in selected predictors with appropriate signs provides additional evidence for their importance. If, for example, *Area Ratio* was a significant positive predictor of accuracy, but was irrelevant for predicting difficulty and confidence ratings, we may have reason to be suspicious of its import.

We withheld 20% of the collected data from model fitting to use as a testing set. Models were fit on the remaining 80% of the data (the training set) and were then used to generate predictors for the withheld 20%. Performance was measured for both the training and testing sets. Testing sets are important to use to ensure that models are not over-fit to the specific sample.

Signal Detection Theory Measurements

In addition to basic accuracy information, one can distinguish between sensitivity and bias in subject responses. This is the basis of signal detection theory (Green & Swets, 1966). Sensitivity describes a sensor's ability to detect a signal. Once a signal is detected, the observer needs to make a decision about how to classify the signal, e.g., whether two prints were from the same source or not. Because sensors are subject to both external and internal noise, the exact same stimulus may elicit different responses across presentations. Sensitivity, *d'* (pronounced "dee-prime"), was computed with the following formula:

$$d' = Z(hit\ rate) - Z(false\ alarm\ rate)$$

Where *Z* is the inverse of the cumulative Gaussian distribution, *hit rate* is the proportion of "match" responses to match trials out of the total number of match trials, and *false alarm rate* is the proportion of "match" responses to non-match trials out of the total number of non-match trials. Values close to 0 indicate poor discriminability (inability to tell apart matching print pairs from non-matching pairs); higher values indicate better discrimination performance. For details, see, e.g., Green and Swets (1966).

Because of the high accuracy among experts found in other studies (e.g. Tangen et al., 2011), we expected their sensitivity to be very high, even for Experiment 1, which had time and tool constraints. We did not have a firm expectation for novices since reports of novice performance were quite varied in terms of their performance. For example, Tangen et al. (2011) found accuracy to be around 75% for matches and around 50% for similar non-matches.

Bias was computed by calculating the log β. The measure can be thought of as the bias to respond "yes" or "no" in a forced-choice signal detection task. For the current study, the two alternatives can be thought of as "match" or "non-match" responses. It is also the log likelihood-ratio for a statistical decision test (see e.g., Wickens, 2002). The bias is computed by the following formula:

$$\log \beta = \log \frac{\varphi(\lambda - d')}{\varphi(\lambda)}$$

where φ(x) is the Gaussian density function, *d'* is the sensitivity, and λ is the decision criterion boundary given by –Z(false alarm rate).

A score of 0 indicates no bias. That is, no preference for saying "match" vs. "non-match" irrespective of one's discriminative ability (i.e., expertise). Deviations away from 0 indicate a preference toward saying "match" or "non-match". Positive bias scores indicate a more conservative decision criterion, a propensity to say "non-match" more often. Negative bias scores indicate a more liberal decision criterion, a propensity to say "match" more often.

We expected that experts would show a slight conservative bias, favoring "non-match" responses because of the high cost of making a false identification. Novices who received no training might not have the same associations with the fingerprint matching task and might show no bias. Tangen et al.'s study, however, indicates that there may be a bias toward saying "match". This would explain the significantly greater accuracy for matches compared to non-matches. Novices who watched the brief training video were made aware of the importance and difficulty of matching fingerprints and so might show a bias similar to experts or a reduction of the bias toward saying "match" shown by novices who did not watch the video.

III. Results

Experiment 1

Descriptive Statistics

Responses were aggregated across participants and prints. Overall accuracy (percent of correctly classified latent-known print pairs, averaged across subjects) was 91% (range: 8.3 -100%, SD 17%). Average accuracy was 86% for "match" trials (14% false negatives) and was 96.8% for "non-match" trials (3.2% false positives). Of the 2,292 comparisons, there were 200 errors, resulting in an overall error rate of 9.6%. Accuracy for particular print pairs ranged from 86% to 95%. There was some variability in performance among experts (range: 79-100%, SD 5%).

Non-matching trials include prints that do not originate from the same source; participants responded to a total of 1144 of these trials. Participants correctly labeled 96.8% of the non-matching trials as "no match" (correct rejections), and incorrectly labeled 3.2% of the non-matching trials as "match" (false alarms). In absolute terms, participants correctly labeled 1107 of the 1144 non-matching trials as "no match" (correct rejections), and incorrectly labeled 37 out of the 1144 non-matching trials as "match" (false alarms). Nineteen examiners made at least one false alarm, and twenty-seven of the non-matching fingerprint stimulus pairs caused at least one false alarm.

At the level of the stimulus, nineteen fingerprint stimulus pairs accrued one false alarm each; six accrued two false alarms each; and two accrued three false alarms each. At the level of the participant, twelve participants made one false alarm; three participants made two false alarms; three participants made three false alarms; and one participant made ten false alarms.

Across all participants, 118 of the 200 print pairs produced 100% accuracy. Mean difficulty and confidence ratings for these pairs were 2.62 and 5.23 respectively, compared to ratings of 4.06 and 4.15 for prints that were misclassified by at least one participant. Of the 118 pairs that produced no errors, 72 were non-matches and 46 were matches. The lowest accuracy, 8.3% (1/12), was for a "match" print-pair. Average accuracy for each print pair is shown in Figure 5 sorted by increasing accuracy.

There was a significant difference between average ratings of difficulty for hits (*M* = 2.95, *SD* =1.58), misses (*M* = 4.57, *SD* =1.25), correct rejections (*M* = 3.17, *SD* = 1.60), and false alarms (M = 5.16, SD = 1.04), *F*(3, 2278) = 69.51, *p* < .001. Post-hoc comparisons revealed that all pairwise differences are significant at the 0.05 level (Bonferroni adjusted *p* < 0.001) except for the comparison of misses and false alarms (Bonferroni adjusted *p* = 0.22).

In assessing average confidence when participants were correct versus when they were incorrect (i.e., collapsing hits and correct rejections and misses and false alarms), there was a significant difference between average ratings of confidence for correct (*M* = 4.96, *SD* = 1.41) versus incorrect responses (*M* = 3.30, *SD* = 1.57), t(2280) = -15.80, *p* < .001.

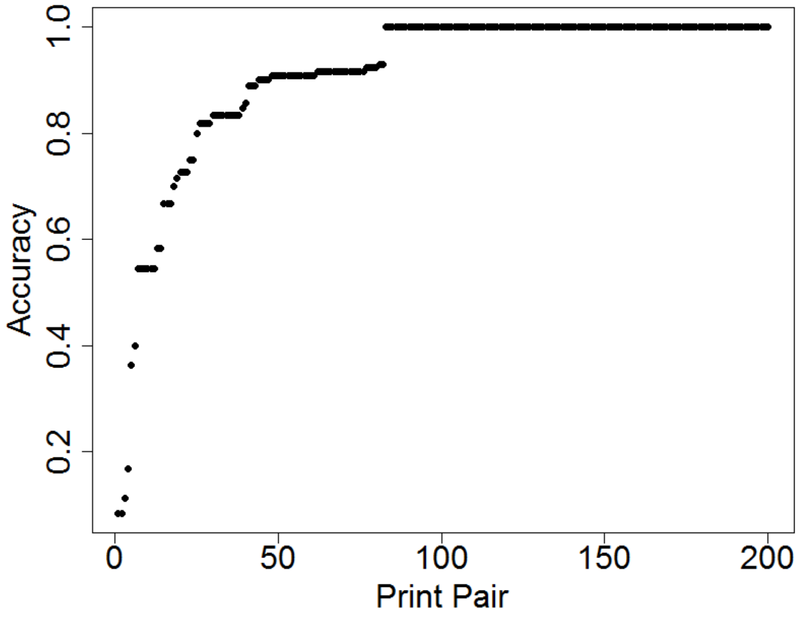


Figure 5. Sorted average accuracy for each print pair. Print pairs are numbered along the x-axis from 1-200 in order of increasing accuracy.

There was a significant difference between average ratings of confidence for hits (*M* = 5.21, *SD* = 1.25), misses (*M* = 3.42, *SD* = 1.57), correct rejections (*M* = 4.74, *SD* = 1.50), and false alarms (M = 2.76, *SD* = 1.44), *F*(3, 2278) = 106.64, *p* < .001. Post-hoc comparisons reveal that all pairwise differences are significant at the 0.05 level (Bonferroni adjusted *p* < 0.001) except for the comparison of misses and false alarms (Bonferroni adjusted *p* = 0.06).

There was a significant difference between average confidence ratings for trials in which participants responded "match" (*M* = 5.12, *SD* = 1.34) versus "no match" (*M* = 4.57, *SD* = 1.57), *t*(2280) = -8.76, *p* < .001. There was a significant difference between average confidence ratings for matching trials (M = 4.95, *SD* = 1.44) and non-matching trials (*M* = 4.68, *SD* = 1.53), t(2280) = -4.39, *p* < 0.001.

There was a significant difference between average difficulty ratings for trials in which participants responded "match" (*M* = 3.03, *SD* = 1.62) versus "no match" (*M* = 3.35, *SD* = 1.63), *t*(2280) = 4.6867, *p* < .001. There was no significant difference between average difficulty ratings for matching trials (*M* = 3.18, *SD* = 1.64) and non-matching trials (*M* = 3.23, *SD* = 1.63), *t*(2280) = 0.8306, *ns.*

For the seventy-four trials on which a particular examiner got a trial correct (hits + correct rejections) and for which at least two other examiners got incorrect (misses + false alarms), the average confidence rating was 3.51 (*SD* = 1.75) and the average difficulty rating was 4.59 (*SD* = 1.32). For the 837 trials on which a particular examiner got a trial correct (hits + correct rejections) and for which all other examiners got correct (hits + correct rejections), the average confidence rating was 5.00 (*SD* = 1.34) and the average difficulty rating was 2.86 (*SD* = 1.53). The difference between the confidence and difficulty ratings for the two sets were significant (confidence: *t*(909) = 9.83, *p* < 0.001 ; difficulty: *t*(909) = -9.39, *p* < 0.001).

There are many hits (438) and correct rejections (435) that the experts rated as not difficult (difficulty rating of 1 or 2) (total = 873 or 41.9% of the total number of correct responses). There were fewer hits (190) and correct rejections (257) that the expert rated as difficult (difficulty rating of 5 or 6) (total = 447 or 21.5% of the total number of correct responses). There were very few false alarms (1) and misses (9) that experts rated as not difficult (total = 10; 5.0% of the total number of incorrect responses). There were more false alarms (29) and misses (88) that the expert rated as difficult (total = 117; 58.8% of the total number of incorrect responses). This set of findings – showing that overall examiners have reasonably strong abilities to assess the difficulty of comparisons – offers interesting insights into examiners' metacognitive abilities, which we are currently in the process of analyzing further for an additional paper on the topic.

Correlations Among Dependent Measures

We measured the correlations of accuracy with the other three dependent measures. There was a strong negative correlation between average difficulty and confidence ratings (*r*(198) = -0.91, *p* < 0.001) and weaker correlations between average accuracy and confidence (*r*(198) = 0.52, *p* < 0.001), and between average accuracy and difficulty (*r*(198) = -0.50, *p* < 0.001). There was also a strong positive correlation between response time (RT) and difficulty (*r*(198) = 0.71, *p* < 0.001) and a negative correlation between response time and confidence (*r*(198) = -0.59, *p* < 0.001). Accuracy was highest and RT lowest for prints that were rated least difficult. Accuracy decreased and RT increased as print difficulty ratings increased. Excluding the 118 prints with 100% accuracy, the correlations between accuracy and confidence and difficulty were qualitatively weaker, but the difference did not reach significance. The full set of correlations is shown in Table 1.

Table 1. Correlations between dependent measures
All Fingerprint Pairs

	Accuracy	Confidence	Difficulty
Confidence	0.52***		
Difficulty	-0.50***	-0.91***	

Response Time	-0.48***	-0.59***	0.71***

Fingerprint Pairs with Accuracy < 100%

	Accuracy	Confidence	Difficulty
Confidence	0.36**		
Difficulty	-0.32**	-0.89***	
Response Time	-0.22*	-0.34**	0.45***

Note. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.5$

Regression Model

A cross, logistic regression model was initially fit to the entire dataset as described in the Analysis Methods section. A model was fit with all of the predictors (after removal of some to minimize collinearity). A likelihood ratio test showed that the model with the predictors fit the data better than a null model with only the random effects terms ($\chi^2(17) = 53.27, p < 0.001$).

Comparing a model that included the random expert effect (expertID) to one that did not, we found that the Akaike Information Criterion (AIC) was slightly smaller for the model that included the effect, but the Bayes Information Criterion (BIC) was smaller for a model that did not. Both of these measures are information-theoretic metrics of goodness-of-fit that take into consideration the overfitting the data with excess parameters. Qualitatively, a more parsimonious model that fit the data almost as well would have a smaller AIC and BIC (Akaike, 1973; Burnham & Anderson 2002). The fact that the criteria move in opposite directions when the model includes expertID suggests that any differences between the models should be treated with caution. A likelihood ratio test comparing the two models was significant ($\chi^2(1)=4.79, p<0.05$) (Zuur, Ieno, Walker, Saveliev, & Smith, 2009). expertID terms varied from between -0.52 ± 0.69 to 0.44 ± 0.77. All values of expertID were within two standard errors of zero. In terms of Equation 2, this means that β_0 + expertID was not reliably different from β_0. Based on these analyses, we felt justified in averaging across experts and ignoring between-expert differences in all subsequent modeling steps by removing the expertID term. This same analysis could not exclude the print-pair specific term, printID, which was retained in the model.

We simplified the model further by removing predictors (fixed effects) based on minimization of the AIC (Zuur et al., 2009). A likelihood ratio test revealed no statistically significant difference between a model that included all of the predictors and the reduced model ($\chi^2(11) = 9.55, p > 0.05$), indicating that the removal of predictors increased parsimony without significantly impacting predictive ability. Similar methods were applied to novice data from Experiment 2 and expert data from Experiment 3.

The model obtained for accuracy was:

$$\begin{aligned} Accuracy = \; & logit^{-1}(3.385 + 0.798 * Delta\ (L) + 0.534 * Mean\ Block\ Contrast\ (K) \\ & - 0.471 * Area\ Ratio - 0.451 * \ SD\ Block\ Contrast\ (L \text{ x } K) + 0.419 \\ & * Sum\ of\ Ridge\ Measures + 0.334 * \ DEAI\ (L \text{ x } K) + printID) \end{aligned}$$

Where L and K indicate whether the predictor applies to a latent or known print image respectively, and LxK indicates predictors that apply to print pairs. printID is the item-specific, random effect. The parameters of the fitted model are shown in Table 2. All predictors were significant (Wald's z, ps < 0.05), except for Delta (L) and DEAI (LxK) which were marginally significant (p = 0.054 and p

= 0.053 respectively)[4]. To get a more intuitive notion of model performance, we used the predicted proportions from the logistic regression as estimates of average performance across experts. The resulting fit was very good (R^2_{adj} = 0.91). We also computed the root mean squared error (RMSE) by taking the sum of the squared differences between predicted and observed values. Values closer to 0 indicated better performance. The error for the fitted model ($RMSE_{model}$ = 0.06) was lower than for a null model that only included the printID random effect ($RMSE_{null}$ = 0.18).

Table 2. Predictors for accuracy model.

Fixed Effects	Coefficient Estimates	Standard Error	z
Intercept	3.385	0.197	17.167***
Delta (L)	0.798	0.415	1.923
Mean Block Contrast (K)	0.534	0.164	3.268**
Area Ratio	-0.471	0.156	-3.010**
SD Block Contrast (LxK)	-0.451	0.128	-3.530***
Ridge Sum	0.419	0.154	2.715**
DEAI (LxK)	0.334	0.173	1.938

Random Effects	Variance
printID	2.154

Note: *** p < 0.001, ** p < 0.01, * p<0.05. p-values are reported here, but should be interpreted with caution. They were not used for model selection (see Footnote 2). Estimates are arranged by coefficient magnitude in descending order (see text). L – latent, K – known print, LxK – interaction.

Validation of the Regression Model for Accuracy

The dataset was then split into training and testing sets. First training on the full dataset was used as a check to make sure that the model could fit at all. If it had failed to fit on the full dataset, there was no point in training on a subset of the data. The training set contained 180 (90%) of the print pairs (2063 individual observations), and the testing set contained the remaining 20 print-pairs (10%, 229 observations). The testing set print pairs were a representative sample of the overall dataset, containing 12 pairs with perfect accuracy and 8 pairs with less-than-perfect accuracy. This was important in order to ensure that the training set did not have too few pairs with low accuracies (there were only 24 pairs in all with average accuracies below 80%). We replicated the model selection procedure for data only from the training set. The same predictors were selected with comparable coefficients, except for Delta (L) which was replaced with Core (L). For both the full and training datasets, the coefficients for these two predictors, Delta (L) and Core (L), were not significantly different from zero and were within two standard deviations of zero. Nevertheless, they could not be excluded based on the selection procedure described above. The fit of the model to the training set was comparable to that of one on the full set (R^2_{adj} = 0.89, $RMSE_{train}$ = 0.07). The results of fitting on the full set were therefore not likely due to overfitting.

We used this regression model fitted to the training set to predict accuracy for the withheld training set of 20 print pairs. Less variance could be accounted for the testing set than for the training set, suggesting some amount of overfitting (R^2_{adj} = 0.64). The error, however, was comparable between the training and testing sets ($RMSE_{test}$= 0.07). The model's predictions are shown in Figure 6.

[4] *p*-values for the Wald statistic in unbalanced, mixed-effects data are difficult to define due to difficulty in determining the appropriate degrees of freedom and therefore should be interpreted with caution (Agresti, 1996, 2002; Baayen et al., 2008).

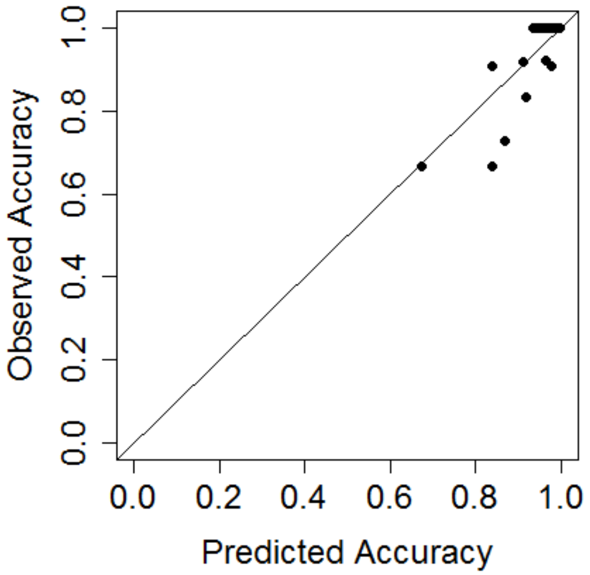


Figure 6. Model predictions of average accuracy for 20 test print pairs plotted against observed average accuracy.

As a secondary assessment of model performance, we used the model to predict whether at least one expert made an error on a print pair. We divided the set of print pairs into two classes: those that had 100% accuracy (perfect pairs) and those that had less than 100% accuracy (non-perfect pairs). A naïve classification strategy not based on the model and that assumes no errors are ever made would have a classification accuracy of 107/180 or 59%. Using the model fitted to the training set, we parametrically varied a classification threshold such that print pairs with a predicted accuracy greater than or equal to that threshold were classified as perfect pairs and those below that predicted accuracy were classified as non-perfect pairs. A threshold setting of 94% resulted in the best classification performance of 164/180 or 91% correctly labeled pairs.

The classification procedure described above was repeated for predictions generated for each left out (testing) pair using the threshold optimized on the training set. 75% (15/20) of the pairs were correctly classified as either having perfect (9/15) or non-perfect accuracies (6/15). The classifier was slightly better at correctly identifying print pairs that had at least one error than those that were perfect: 3 perfect prints were misclassified as having an error and 2 non-perfect pairs were misclassified as perfect.

Difficulty Ratings

Difficulty ratings showed a reliable negative correlation with accuracy (see Descriptive Statistics, above), indicating that experts had reasonable metacognitive awareness (i.e., print pairs that were thought to be difficult tended to have low accuracy across experts). Accuracy for trials with a difficulty rating greater than 3 (on a scale of 1 to 6) was 84% compared to 91% for all comparisons. We compared the fitted model from the previous section to one that also included difficulty rating as a predictor. The resulting model had significantly better goodness of fit than the model from the preceding section that did not include it as a predictor ($\chi^2(1)=81.1, p<0.001$, $\text{RMSE}_{\text{model+difficulty}} = 0.05$, $R^2_{\text{adj}} = 0.95$).

We added difficulty rating as a predictor for the regression model applied to the training set described above. Predictive performance on the testing set was worse (decreased R^2) than when the difficulty rating was not included. However, classifier performance on the testing set was slightly improved, with 85% (17/20) of the pairs classified correctly. One perfect print was misclassified as non-perfect, and two non-perfect prints were misclassified as perfect. The discrepancy between the relatively worse regression fit and the improvement in classifier performance is due to two non-perfect print pairs that had a predicted accuracy that was much lower than their true accuracy. These were classified correctly as non-perfect, but contributed significantly to the error.

The inclusion of difficulty ratings in applications of this model must be made with caution. All other measures capture objective features of the fingerprint image, while difficulty ratings are subjective and therefore may vary across individuals and rely on the good faith of the raters. Therefore, while difficulty rating may be informative to include, in subsequent models we opt to exclusively deal with objective factors. We return to this point in the discussion.

Regression Analysis of Other Dependent Measures

Difficulty ratings, confidence ratings, and response times were reliably correlated with accuracy and so ought to also depend on print pair information content. If similar features are predictors for many measures, then they are likely capturing something important about the fingerprint images. Here, we fit models of the other dependent measures to the training dataset as a further validation step: the importance of particular image features as valid predictors of accuracy is bolstered if those same features are shared in models of other dependent measures.

Unlike accuracy, response time varied greatly across experts, with some taking much longer times on comparisons that others evaluated fairly quickly. There are several possible reasons for this variability. Less experienced examiners may take longer to come to the same conclusion than a seasoned examiner (a perceptual fluency that comes with expertise; see Kellman & Garrigan, 2009). Some subjects may have completed the comparison quickly, but then taken time to deliberate confidence and difficulty ratings since response time was recorded only once all answers were given, and not when the subject selected “match” or “non-match”. Also, the self-confidence of the examiners in their abilities may have affected response time. Only a small component of the variability in response time was likely to be due to differences in attention or interest since such differences would presumably have led to greater variability in accuracy, which was not observed.

We fit a linear, mixed-effects model to normalized response time data for the training set following the same model selection steps as for the accuracy model in the preceding sections. Due to the variability in response time across experts, the random effect of expertID was retained in the model. The results of the regression are shown in Table 3. Three features, Core (L), Mean Block Contrast (K), and SD Block Contrast (L) were found to be predictive of response time using the same model selection procedure that was used for the analysis of predictors of comparison accuracy. The latter two predictors were also selected in models of accuracy (SD Block Contrast as part of an interaction term). Visibility of cores instead of deltas was selected as a predictor of response time. Interestingly, it also appears as predictor when the model is fit to a testing set. Visibility of a core might make it simple to compare latents and known prints: if the cores do not match then no further comparison is required, so a comparison can be made quickly. Absence of a core could also make it difficult to orient the latent and known prints, since, as noted earlier, these features could act as landmarks for orienting two prints during comparison.

Linear mixed-effects models were also fit separately for difficulty and confidence ratings. Like response time, there was a great deal of inter-subject variability for both measures. Variability in confidence and difficulty ratings may be due to differences in degree of expertise and self-confidence in the task. Variability in ratings may also be due to differences in interpretation of the rating task and therefore in response strategy. One expert, for example, responded with maximum confidence to all comparisons, saying to the experimenter that in real-world situations an expert would be 100% confident or rate a comparison as inconclusive.

Table 4 contains the coefficient estimates for the model of difficulty rating. As in the model of accuracy, Ridge Sum, Area Ratio, and Core (L) were selected as predictors. Similar to response time, difficulty was also negatively correlated with accuracy, so the regression coefficients have opposite signs to those in the accuracy model. In addition, visibility of Cores in the known print and the interaction of the Core terms were also selected. Delta (L) appears in this model as well as in the model of accuracy.

Table 4. Predictors for difficulty rating model.

Fixed Effects	Coefficient Estimates	Standard Error	T
Intercept	2.748	0.301	9.121***
Core (L x K)	-2.104	0.722	-2.913**
Core (L)	1.719	0.705	2.437**
Core (K)	0.935	0.324	2.883**
Delta (L)	-0.778	0.191	-4.082***
Ridge Sum	-0.207	0.079	-2.631**
Area Ratio	0.202	0.078	2.571**

Random Effects	Variance
printID	1.076
expertID	0.301

Note: *** p < 0.001, ** p < 0.01, * p<0.05. Estimates are arranged by coefficient magnitude in descending order (see text). L – latent, K – known print, LxK – interaction.

A similar model was fit for confidence ratings. The results are shown in Table 5. Identical predictors with comparable magnitudes were selected as for the difficulty rating model. The coefficients have opposite signs since high difficulty ratings correspond to low confidence ratings. Because difficulty and confidence are so strongly correlated (-0.91), it is not surprising that the exact same predictors are selected for in both models.

Table 5. Predictors for confidence rating model.

Fixed Effects	Coefficient Estimates	Standard Error	t
Intercept	5.248	0.247	21.255***
Core (L x K)	2.034	0.564	3.604***
Core (L)	-1.644	0.551	-2.983**
Core (K)	-0.920	0.253	-3.631***
Delta (L)	0.581	0.149	3.899***
Area Ratio	-0.162	0.062	-2.647**

Ridge Sum	0.155	0.062	2.517**

Random Effects	Variance
printID	0.616
expertID	0.488

Note: *** p < 0.001, ** p < 0.01, * p<0.05. Estimates are arranged by coefficient magnitude in descending order (see text). L – latent, K – known print, LxK – interaction.

Experiment 2

Descriptive Statistics

Overall accuracy for untrained novices was 53%. There was a significant difference for the average accuracy for "match" trials (62%) and "non-match" trials (45%, $t(49) = 3.51$ $p < 0.001$). Accuracy, averaged across prints for individual participants, ranged from 42% to 62% (M = 53%; SD = 5.8%). Of the 1800 comparisons, there were 838 errors, resulting in an overall error rate of 47%.

Overall accuracy for trained novices – and by 'trained' we mean only exposure to the short video presentation about fingerprint evidence and how it functions – was 54%. There was no significant difference ($p > 0.05$) for the average accuracy for "match" trials (54%) and "non-match" trials (54%). Accuracy, averaged across prints for individual participants, ranged from 43% to 75% (M = 54%; SD = 7.5%). Of the 1800 comparisons, there were 826 errors, resulting in an overall error rate of 46%.

The five highest accuracies for trained novices were 58%, 59%, 60%, 64% and 75%. The five highest performing untrained novices had accuracies of 57%, 60%, 61%, 61%, and 62%. Accuracy and rating scores are depicted in Figure 7. Expert scores from Experiment 1 are included for comparison.

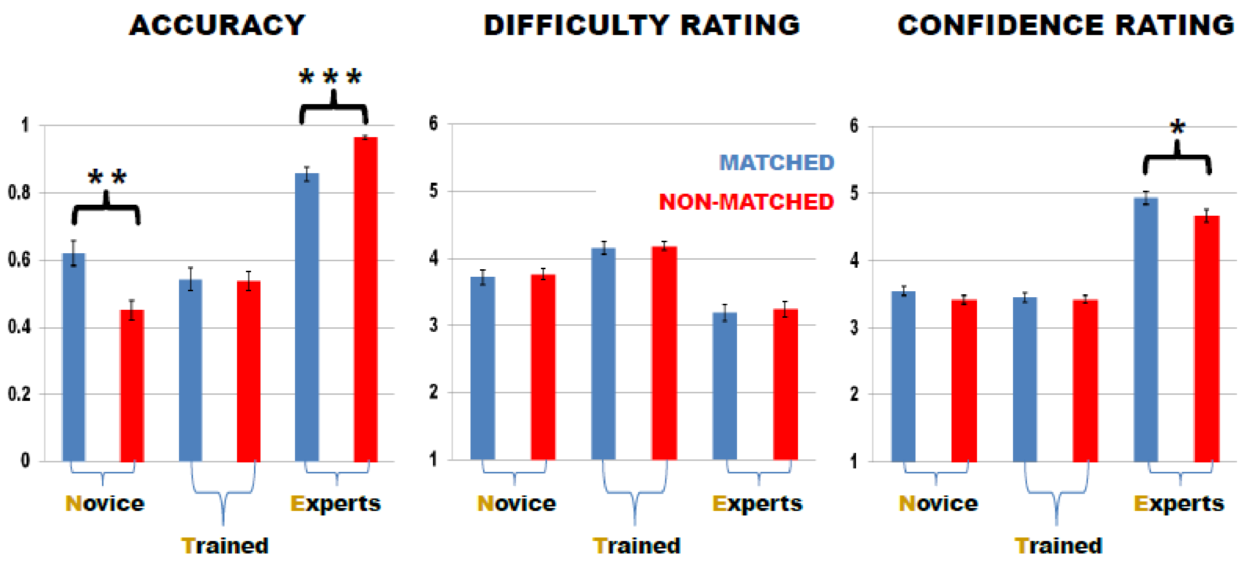


Figure 7. Mean accuracy, difficulty and confidence ratings for untrained novices, trained novices, and experts. Data are split by matching and non-matching comparisons. Error bars are standard errors. *s indicate significance levels of independent t-tests between matching and non-matching comparisons. * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

For both groups of novices, there was a negative correlation between difficulty and confidence ratings ($r(198) = -0.87$ and -0.85, $p < 0.001$, with and without training respectively), and a weaker correlation between accuracy and confidence ($r(198) = 0.31$ and 0.33, $p < 0.001$, with and without training) and between accuracy and difficulty ($r(198) = -0.41$ and -0.31, $p < 0.001$, with and without training respectively). It is interesting to note that the latter correlation is higher for novices with training than without training, perhaps indicating that they were beginning to appreciate the visual features that make matching print pairs harder.

The variance for experts' accuracy was less than that for either group of novices. This reflects the near-ceiling performance of the experts. The average difficulty rating was lower for experts (3.2/6) than either group of novices (untrained: 3.7/6, $t(30) = 3.50$, $p < 0.01$, trained: 4.2/6, $t(27) = 5.81$, $p < 0.001$). There was also a significant difference in difficulty ratings between the two groups of novices ($t(33) = 2.25$, $p < 0.05$). Confidence ratings of experts (4.8/6) were also higher than those of novices (untrained: 3.4/6, $t(37) = 7.26$, $p < 0.001$, trained: 3.6/6, $t(39) = 7.18$, $p < 0.001$). There was no significant difference in confidence ratings between the groups of novices. For the experts there was a significant difference between confidence ratings for match trials (4.9/6) and for non-match (4.7/6; $t(99) = 1.98$, $p < 0.05$)). No such asymmetry was found for the novices.

Signal Detection Measures

To assess participants' sensitivity in discriminating matches and non-matches, we submitted accuracy scores from the assessed print pairs to a signal detection analysis (Green & Swets, 1966). The average sensitivity for the expert group ($d' = 2.64$) was much higher than for the novices ($d' = 0.19$). There was no significant difference between the average sensitivity of untrained ($d' = 0.17$) and trained ($d' = 0.21$) novices. Despite low average sensitivities, the maximum sensitivity was 0.63 for untrained novices and 1.36 for trained novices. However, only 2/18 trained novices had sensitivities higher than the maximum untrained novice sensitivity.

Response bias ($\log \beta$) was computed for novices and trained novices. Mean bias (averaged across subjects) was 0.01 and 0.04 respectively. There was no significant difference between the two groups ($t(34) = -1.19$, $p = 0.24$). Average bias for the six highest performing untrained novices was slightly liberal (-0.06), while the average bias for the two highest performing trained novices whose sensitivity was greater than the maximum sensitivity of untrained novices was slightly conservative (0.12), but the difference between the two was not statistically significant ($t(6) = -2.13$, $p = 0.073$) perhaps because there were so few trained novices with high sensitivities.

Regression Analysis

The same crossed, logistic regression model was fit to the novice data as was used for experts in Experiment 1. Similar procedures were followed to remove variable and simplify the model. The results are shown in Table 6 with the coefficients from the fit to the expert data included for ease of comparison. Ridge Sum, Mean Block Contrast (K), SD Block Contrast (LxK), DAEI (LxK), and visibility of Cores (K) were selected as significant predictors for novices. Three predictors, Delta (K), Ridge Sum, and DEAI(K), were selected in the trained novice model.

Table 6.

	Expert	**Untrained Novice**	**Trained Novice**
Fixed Effects	Coefficient Estimate		

	(Standard Error)		
Intercept	3.385 (0.197) ***	0.521 (0.326)	0.628 (0.231) **
Area (K)			
Area Ratio	-0.471 (0.156) **		
Delta (L)	0.798 (0.415)		
Delta (K)			-0.523 (0.246) *
Ridge Reliability (K)			
Ridge Sum	0.419 (0.154) **	0.297 (0.112) **	0.312 (0.096) **
Mean Block Contrast (K)	0.534 (9.164) **	-0.540 (0.112) ***	
SD Block Contrast (L)			
SD Block Contrast (LxK)	-0.451 (0.128) ***	0.194 (0.085) *	
DAEI (K)			-0.253 (0.099) *
DEAI (LxK)	0.334 (0.173)	-0.213 (0.101) *	
Core (L)		0.463 (0.251)	
Core (K)		-0.752 (0.373) *	
Core (LxK)			
Random Effects			
	Variance		
Print Pair	2.154	0.809	0.658
Subject			0.077

* p<0.05   ** p < 0.01   *** p < 0.001

Table 6. Coefficient estimates for the three groups of subjects: experts, untrained novices, and trained novices, and for a high-performing subset of the trained novices. L – latent K – known print L*K - interaction. Because a model selection procedure was used to select the most parsimonious model, some parameters do not appear in all models. Fixed effects appear at the top of the table and random effects appear at the bottom. For random effects, the estimated variance is specified. *p* values correspond to significance tests on Wald statistics for each predictor, which are not shown in this table. For mixed-effects models, it is difficult to determine the appropriate degrees of freedom, so *p* values should be interpreted with caution. Instead, it may be more informative to examine whether predictor coefficient estimates are within two standard errors of the 0. Predictors are sorted first in descending order of coefficient magnitude for experts and then by L, K, and L*K.

The root mean squared error (RMSE) was used as a measure of model performance on a withheld dataset of 20% of the prints similar to Experiment 1. RMSE was computed by making individual accuracy predictions for each print pair and then comparing this predicted average accuracy to the observed average accuracy. Point estimates of the predictor coefficients and random effect terms were used. RMSE for the expert, novice, and trained novice testing sets were 0.07, 0.25, and 0.21, respectively. The larger RMSEs for both groups of novices indicate poorer model fits. Regression predictors can still be interpreted as important contributors in predicting accuracy, but the model should be interpreted with caution. The poor fit is not surprising given near-chance performance for both groups of novices. However, it is interesting that despite these worse prediction results, a different, almost completely non-overlapping set of predictors is selected for in the trained novice model, and that the prediction performance is slightly improved relative to the untrained novice model.

Experiment 3

Descriptive Statistics

Thirty-four examiners made a total of 1646 comparisons. Each print pair was evaluated by a minimum of seven distinct examiners. Average accuracy was 94.84%. Performance on matches was 90.00% while performance on non-matches was 99.75%. The lowest accuracy for any print pair was 10%. There were three print pairs out of 120 with an average accuracy less than 50%, three print pairs with an average accuracy between 50% and 75%, and 16 print pairs with an average accuracy between 75% and 100%. 98/120 print pairs had perfect accuracy.

Average examiner accuracy was 95.03%. The lowest accuracy was 81.82%, the highest was 100%. Four examiners computed fewer than 10 comparisons, but none made any mistakes. Eighteen examiners completed between 10 and 50 comparisons with an average accuracy of 93.55%. Twelve examiners completed more than 50 comparisons with an average accuracy of 95.6%.

Of the 1646 total comparisons, 126 were labeled as inconclusive, of which 73 were matches and 53 were non-matches. Average accuracy for prints labeled inconclusive was 76%; average accuracy for prints not labeled inconclusive was 96.38%. Average difficulty rating for inconclusive prints was 4.62; average difficulty rating for non-inconclusive prints was 2.27. For pairs that were labeled inconclusive by any examiner, an average of 23% of examiners labeled those prints inconclusive. At most, 7/9 examiners rated a particular pair inconclusive. Of the 42/120 pairs that had at least one examiner label inconclusive, five had 50% or more of examiners agree that they were inconclusive with an average accuracy of 59.78%. The remaining 37 comparisons had fewer than 50% of the examiners that rated them as inconclusive and had an average accuracy of 90.19%.

Half of the print pairs used in this experiment were also used in Experiment 1. Performance was strongly correlated across the two experiments on that subset of comparisons (Spearman's rho = 0.45, p < 0.001).. The accuracies for the two experiments are shown in Figure 8. Qualitatively, accuracy for many pairs was similar across both experiments. However, for several pairs there were marked differences. For two pairs, for example, accuracy in Experiment 1 was close to 50%, but was near 100% in Experiment 3. Another print pair had an accuracy of near 10% in Experiment 1 and an accuracy of approximately 55% in Experiment 3. We have not yet examined the kinds of tools that were used with each of these comparisons (see Conclusion for planned future analysis of these data).

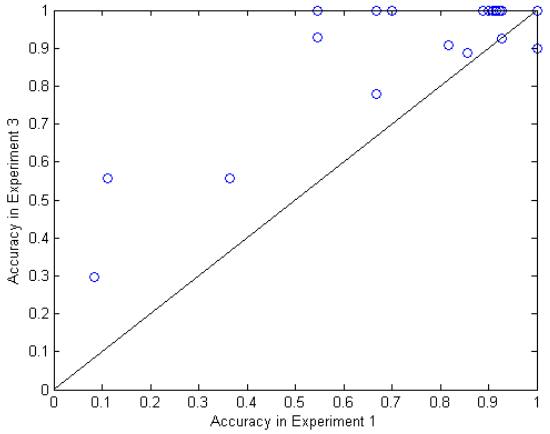


Figure 8. Average accuracy for the 60 print pairs in both Experiment 1 and Experiment 3.

Response Time

Minimum response time was 2.4 seconds. Maximum response time was 91.5 minutes. Average response time was 2.58 minutes. Average accuracy for comparisons that took less than a minute (21.43% of all comparisons) was 97.12%; average accuracy for comparisons that took more than a minute was 93.02%. Since users could leave the experiment window open (there was no time-out), the response time does not necessarily reflect the amount of time spent evaluating the fingerprints.

Tool Use

Across all comparisons, 82.38% made some use of the tools. Average accuracy for comparisons involving tool use was 95.86%, while average accuracy for comparisons without tool use was 94.62%. Average difficulty rating for comparisons on which tools were used was 2.61; average accuracy for comparisons without tool use was 1.72.

Minutiae were marked on 41.86% of the comparisons. On average, 3.74 *minutiae* were marked per comparison. Average number of marked *minutiae* was 4.42 for comparisons rated inconclusive and 3.68 for comparisons not rated inconclusive. Accuracy was 95.3% for comparisons with no *minutiae* marked and 94.19% for comparisons with at least one marked. Average difficulty rating for comparisons with no *minutiae* marked was 2.17 and 2.84 for those with at least one marked.

For the other tools, 79.65% of comparisons had the zoom tool used, 12.64% used rotation, 24.30% had a brightness or contrast adjustment. In all cases, average difficulty was rated as higher for comparisons that had tool use compared to those that did not (2.61 vs. 1.85, 2.90 vs. 2.39, 3.15 vs. 2.23 for each of the tools respectively).

Regression Analysis

The model fit in Experiment 1 was used to predict accuracy data from this experiment. The predictions were based on the un-edited images, i.e., it did not take into account if examiners used a tool to alter image properties like brightness or contrast. Since the model takes those features as inputs, the model predictions need to be interpreted with caution. Subsequent analyses will investigate how the model's predictive performance changes when image features are computed taking into consideration individual subject modifications.

Data were split two ways: First, by print pairs tested in Experiment 1 and those that were new to this experiment. Second, by whether the pairs were rated as inconclusive by at least one examiner. Model predictions are shown in Figures 9 and 10 respectively. While many of the pairs used in Experiment 1 had qualitatively good accuracy predictions, six had observed performances that were drastically different from predicted performance. Many more new pairs had inaccurate predictions. However, out of all of the pairs that were presented in Experiment 1 and had inaccurate predictions, only one had no examiners rate it as inconclusive (predicted accuracy: 79.3%, observed accuracy: 100%, number of examiners: 10).

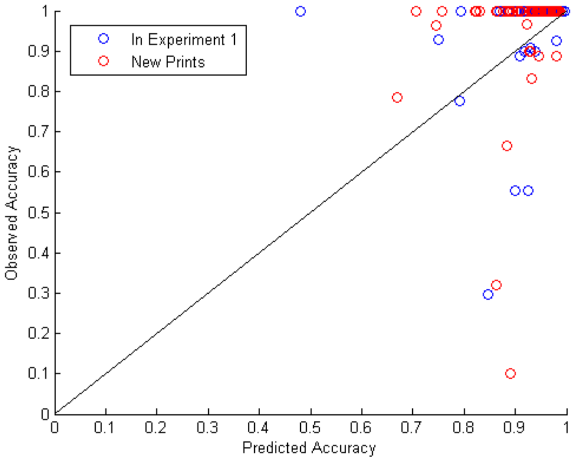


Figure 9. Predicted vs. observed accuracy for print pairs in Experiment 3. Predicted accuracy is obtained by fitting the model from Experiment 1. Pairs are split by whether they were included in Experiment 1 (blue circles) or not (red circles). There were 60 pairs in each group.

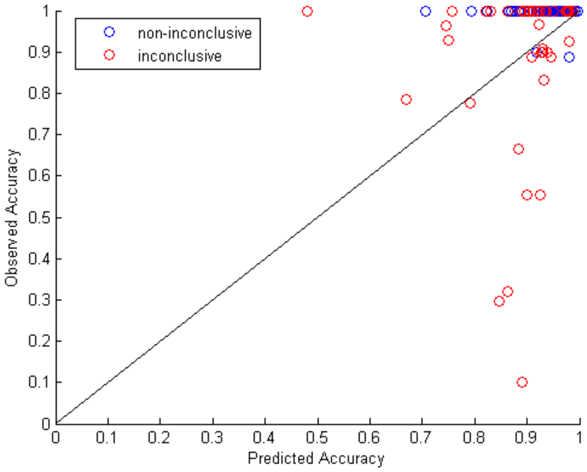


Figure 10. Predicted vs. observed accuracy for print pairs in Experiment 3 as in Figure 8. Pairs are split by whether at least one examiner rated that pair as inconclusive (red circles) or whether no examiner rated them as inconclusive (blue circles).

IV. Conclusions

In Experiment 1, we evaluated expert performance on a fingerprint matching task. Experts were highly accurate, committing few errors despite limited access to resources and restricted viewing time. Using a number of potential predictors derived from image processing algorithms, we were able to identify, using regression analyses, several image characteristics predictive of expert performance. Six features in particular were found to be important predictors of accuracy: Ridge Sum, Area Ratio, visibility of Deltas in the latent print, Mean Block Contrast of the known print, interaction between SD Block Contrast for latents and known prints, and the interaction between DEAI (deviation from expected average intensity) for the latents and known prints. Taken together, these features can explain 64% of the variance in performance accuracy on a novel set of print pairs that were withheld from those used to train the model. A classifier derived from the full data set identified the pairs on which at least one expert made a mistake with 91% accuracy, and a similar model derived from 90% of the data classified novel pairs with 75% accuracy.

Many of the same image characteristics were also predictors of subjective difficulty ratings, confidence ratings, and response times. We also found that difficulty ratings, a subjective measure, were moderately correlated with accuracy and could improve the performance of the classifier on novel print pairs.

There are several interesting observations that can be made about the set of features that we found to be predictive of accuracy (Table 2). First, four of the six features were relational, in the sense that they were calculated based on information from both prints in a comparison pair. This is a desirable feature of the model since a particular print could arise in two separate comparisons (e.g., a latent print compared to a matching and a non-matching print). In real world scenarios, a single latent print may be compared to a many known prints. In cases where one of the prints in the study was of very poor quality, such relational features might not matter. For example, if Mean Block Contrast (K) is low (i.e., for a very washed out or very dark print), then a comparison would be difficult irrespective of some relational features such as Area Ratio. Conversely, if two prints do not share Level 1 pattern type, they will not make for a difficult comparison regardless of the quality and quantity of information in each. In general, however, error rates and difficulty seem likely to be primarily characteristics of print comparisons, rather than individual prints, as difficulty for actual non-match comparisons will be most acute when the prints share significant similarities, and difficulty for actual matches will be most acute when latent quality or quantity is limited or misleading. Results in our regression models support this idea.

Second, the features within the model correspond to many types of information content. Mean Block Contrast (K), SD Block Contrast (L x K), and DEAI (L x K) capture properties of the image itself (i.e., dark or light, uniform or not). Area Ratio and Delta (L) reflect large-scale or configural (Level I) characteristics of prints, and Ridge Sum relates to visibility of fine detail in the image such as Level II features (see Introduction). These outcomes fit broadly with the idea that fingerprint examiners access different kinds of information in making comparisons and that basic image characteristics determine the detectability of relevant features and patterns.

Third, although not our primary result, the signs of the coefficients provide appealing interpretations and verify our expectations about the negative impact of low quality prints. That high contrast and clarity of ridges are predictors of accuracy should not be surprising. The DEAI measure increases as the average pixel intensity approaches 127.5, the mean expected pixel intensity for an image that contains 50% white and 50% black pixels. We assumed that this proportion would correspond to greater clarity, since a mostly light or dark image could be difficult to analyze. The positive coefficient found for this measure in the accuracy model indicates that as the proportion of white to black pixels approaches 0.5 in the latent and known print, accuracy increases. Visibility of deltas in the latent image also had a positive effect on accuracy perhaps because they provided orienting information, making it easier to match relative locations on the latent and known print. Accuracy decreased as SD Block Contrast (L x K) and Area Ratio increased. When SD Block Contrast is high in both the latent and known print, accuracy is low. In general, high variability in Block Contrast picks up variable image quality across image regions (e.g., due to gaps or smudging in portions of a print). In smudged regions, pixels would be uniformly dark, while in clear regions pixel intensity would be more variable, leading to higher contrast measures in those areas. If an image were more uniform in pixel intensities, it would have lower variability in contrast across regions and therefore lower SD Block Contrast measures. Area Ratio had a large, negative coefficient. This at first seems counterintuitive; higher area ratios tend to correspond to larger areas of latent prints. Larger areas, however, may make comparisons more difficult by making it more difficult to identify distinctive regions of the image. Since non-matching known prints were chosen by submitting the latents to an AFIS system, the non-matches likely shared many features. If experts were only shown a small latent region, it might have been easier to compare that region to the corresponding region on the known print and quickly exclude mismatched pairs as compared to a larger latent image with more accidentally matching regions.

In addition to being able to predict accuracy, it may be important to identify which comparisons are likely to yield an error and therefore may require more scrutiny. To address this issue, we created a classifier that sorted the print pairs into ones that had perfect accuracy (so-called “perfect pairs”) and ones on which at least one expert made a mistake (“non-perfect pairs”). The classifier was able to correctly sort the pairs with 91% accuracy on the training set and 75% accuracy on the testing set.

Difficulty ratings were used in two ways to add to the modeling results. We used difficulty rating itself as a predictor of accuracy. Difficulty ratings improved the fit of a model trained on all of the print pairs, but did not improve the predictive power of a model on a testing set of withheld prints. Classification performance, however, was improved. While ratings are not objective, there was nevertheless a moderate correlation between them and accuracy, suggesting that experts were aware of which comparisons were difficult, an issue we are also exploring in a paper in progress. Outside the experimental setting, it may be impractical to expect to be able to get a group of experts to provide ratings.

Difficulty ratings, confidence ratings, and response times were also evaluated as separate dependent measures. Because these measures correlated moderately with accuracy, we expected that similar features should be selected for when the same features were used to predict other dependent measures. Four of the six features that were significant predictors in the accuracy model were also significant predictors in the other models. A fifth feature, SD Block Contrast (K), which was included as part of an interaction term in the accuracy model was also included in the model of response time. Some features, such as visibility of cores, were significant predictors in the other models but not in the model of accuracy. Cores and deltas are global features. Their presence or absence can be used as a quick measure of assessing difficulty. However, global features on their own are not sufficient to make a comparison. Accuracy, therefore, depends to a greater extent on image quality, relational information, and ridge information.

These results suggest that physical characteristics of fingerprints, measured using automated image processing methods, may be valuable in predicting comparison difficulty and error rates for print pairs. Given that the present work is the first effort we know of to systematically predict errors from physical characteristics of print pairs, the predictive results are highly encouraging. Validation across larger data sets would be desirable for practical use of a predictive model such as the one derived here. Further developments along these lines, along with continuing progress in characterizing the physical quality of prints (e.g., Pulsifer et al., 2013), will likely prove to have practical value in quantifying the likely evidentiary value of expert assessments of fingerprint matches.

While these results on modeling print-pair difficulty are encouraging, there are also many differences between the paradigm used in the present study and the actual process of fingerprint comparison. Experts typically have unlimited evaluation time and access to image processing tools that were not available in the experiment described here. In addition, examiners typically are not in a “forced-choice” situation, and may decide that a real-world comparison is inconclusive. (Experiment 3, however, does attempt to correct in part for these limitations.)

Despite these limitations, there are several important dimensions to these results. The results show that even under constraints, experts were highly accurate. More than half of the print pairs had perfect accuracy, even in circumstances where the examiners’ time was limited, their access to tools constrained, and they were not permitted to select the option of “inconclusive”. Relatively few studies have examined expert performance in fingerprint matching tasks, and this study adds to that body of research. It is possible, however, that error rates in forensic laboratory settings are lower than those we observed. It is also possible that other aspects of real world settings – like the danger of cognitive bias, the pressure of casework, and knowledge of extraneous information about the case – could elevate error rates as compared to experimental conditions. Care must be taken not to generalize too quickly from experimental settings, but nonetheless, such experiments can reveal a great deal about examiner performance, albeit under constraints.

Experiments in ecologically valid settings are difficult to conduct. In such settings, there are many factors that may improve accuracy (such as more time to conduct the comparisons, verification checks, etc.), as well as factors that can reduce accuracy (such as biasing influences from extraneous contextual case information, see Kassin, Dror, & Kukucka, 2013; Dror & Rosenthal, 2008). Given the significant differences between our experimental conditions and ecologically-valid fingerprint identification, we wish to reiterate that the point of the experiment reported here is not to measure such error rates, and it would be a mistake to take these data as direct evidence of a specific error rate for the field (Koehler, 2008). Rather, we are interested in identifying the features that correlate with difficulty, in order both to understand what features of print pairs affect difficulty, and to begin to understand how error rate might *vary* with comparison difficulty.

Consistent with several previous studies, very large performance differences were observed in Experiment 2, between experts and novices. Experts committed relatively few errors (approximately 9%), while novices performed nearly at chance. Experts outperformed novices despite that fact that they were under time constraints and did not have access to typical tools (i.e., image manipulation software, compass, or magnifying lens). Novices who watched a brief training video prior to the task did not perform differently overall (54% accuracy); however, trained novices committed fewer false alarms than untrained novices and, in general, were more conservative in their responses. In this manner, they were, to a limited degree, in between untrained novices and experts in at least one respect: untrained novices had many more correct answers when the prints actually matched (hit) while experts had more correct answers when the prints were from different sources (correct rejections). Trained novices performed like neither of these other groups, in that they had similar performance for both kinds of comparisons. This may reflect a shift in bias regarding an implicit 'default' conclusion – when novices see two prints with a lot of information, they may be biased to say that they match, being at a loss of what parts of the image are relevant for comparison. Experts, on the other hand, have a better sense of what features are important for making comparisons and also may be biased against false alarms (which in real world settings would result in a false conviction), saying that two prints do not match when they are unsure and therefore leading to more correct rejections. This possibility was reflected in higher confidence ratings by experts for comparisons of matching prints than for non-matching prints. Trained novices may have picked up, even on the basis of a very short video training, some idea of what information to focus on in the print and so become less likely to say that two prints match when they are unsure. Furthermore, several trained novices greatly outperformed untrained novices, with one having an overall accuracy of 75%, while the highest untrained novice accuracy was 62%.

There were also marked differences in confidence and difficulty ratings between both groups of novices and experts. In general, experts were more likely to rate prints as easy and to have higher confidence in their ratings. The short training video did not have an effect on confidence ratings among novices, but trained novices did rate comparisons as more difficult overall than untrained novices. This confirms the notion that novices were guessing when it came to comparisons, which is why their accuracy was at chance. It was not surprising that the short five minutes training video did not drastically improve performance. What was surprising is that even such a short training session made the subjects more attuned to the difficulties of comparisons, perhaps by directing their attention to relevant features so that they became more aware of the difficulty of the task.

The same model-fitting procedure as in Experiment 1 was used to fit accuracy data for trained and untrained novices. Ridge Sum was the only predictor that appeared in all three models. Area Ratio was only selected for in the expert model. Mean Block Contrast (K), SD Block Contrast (LxK) and DEAI (LxK) appeared in models for both experts and untrained novices, but with opposite signs. This may mean that novices did not use the information appropriately. Features of images that normally help experts may have served to confuse novices. An overabundance of information may overwhelm a novice observer and lead them to incorrectly treat two complex visual stimuli as sufficiently similar. The fact that the predictors are not selected for in the model for trained novices may indicate that the training helped novices use the information in fingerprint images more appropriately. High information content did not bias them in the same way to label a comparison as a match. However, the model fits were much worse for both trained and untrained novices compared to experts, suggesting that the model did not provide a good fit to the data. This is not surprising given that accuracy was at chance for both groups.

Overall, we confirm that novices are very poor at fingerprint comparison, at least when tested on reasonably difficult exemplars. Similar to Tangen et al. (2011), we found that untrained novices had better performance for matches than non-matches. However, their match performance was not as high as that observed by Tangen et al., (62% vs. 75%). Averaged across match and non-match comparisons, novices in Experiment 2 performed at chance. Watching a short training video eliminated the difference in performance between matches and non-matches and slightly shifted bias for a subset of the subjects. This suggests that the advantage for matches for novices is due to a biased preference to label a comparison as a match. It is interesting to note that this pattern is reversed for experts. Experts have greater accuracy for non-matches than for matches. An opposing bias may exist for experts because they are more aware of the high cost of making an incorrect identification and would prefer to err on the side of caution; even under experimental conditions that instruct them to make their best guess, they may not view a false positive and a false negative as equivalent errors. Since the bias disappeared for trained novices, the training video may have emphasized the importance of correct identification, the difficulty of comparisons, and the high cost of errors. As a result, trained novices may have been more reticent to call a comparison a match by default. The first and most rapid effect of training may therefore be to alert the observer to structures in a fingerprint image that can be used to discriminate two images. As with other perceptual learning domains, more exposure is required to learn to exploit fingerprint information content to make comparisons. This demonstrates that fingerprint examiner expertise is a perceptual learning domain and is therefore likely amenable to the same kinds of training methods that have been used in mathematics and category learning (e.g., Mettler & Kellman, *in press*; Thai, Mettler, & Kellman, 2011).

Further studies need to be conducted to trace the effects of perceptual learning on accuracy and bias. A long-term study that tests examiners through various stages of their training might be able to identify gradual changes in accuracy. Changes in accuracy may correspond to a gradual reweighting of predictor variables. Examiners just beginning their training may give weights to image features in a manner similar to novices. As training progresses, a gradual shift of which variables matter most for accuracy may occur until weights match those of examiners in Experiment 1. It would be valuable and interesting further research to examine how quickly these shifts occur and how different kinds of training might affect them.

Experiment 3 sought to extend the findings of Experiment 1 by testing examiners within substantially more realistic settings for fingerprint comparison. Examiners were given unlimited time to make their comparisons and were provided with an array of image processing tools similar to those they would normally have access to in the course of their work. In addition, experts were given the opportunity, after giving a conclusion, to label a print pair as "inconclusive", an option they were not given in Experiment 1. However, even when a pair was labeled inconclusive, experts were still required to provide a "match" or "non-match" judgment, which was meant to reflect their best guess. In this sense, our protocol was quite different from laboratory practice, in which an 'inconclusive' determination means that the examiner does not offer any further speculation about whether the print pair does or does not share a common source. But this approach gave us important clues about the relationship between performance and an indication that a print pair lacked the quality to warrant evaluation, both for an individual examiner, and in aggregate.

Error rates in Experiment 3 were similar to those in Experiment 1 and those reported in other studies (e.g., Tangen et al., 2011). This is a valuable finding, because it suggests that error rates in the first experiment cannot therefore be solely attributed to lack of resources or time to perform comparisons. There was wide variability in the number and types of tools used by experts. Tool use was often, but not always, correlated with greater difficulty and worse performance. Intuitively, this may have occurred because more difficult comparisons necessitated additional image manipulations, but the use of manipulations was not associated with greatly improved accuracy.

There was very little agreement on which comparisons were inconclusive or not. One possible reason for this discrepancy is variation in expertise among examiners. Another reason could be variation in decision criteria – some examiners may be more willing to label a print as a match or non-match rather than inconclusive than others. If differences are due to decision criteria, then one may be able to determine objectively whether there is in fact enough information to make an identification. For example, if a model predicts very high accuracy for a particular comparison, then this may be used to encourage examiners to spend extra time evaluating a comparison before determining that there is insufficient information to make a match / non-match decision. That is, it may be possible to objectively determine whether there is or is not sufficient information in a particular print pair. This would allow one to judge whether a determination of inconclusive is correct or not. We are still actively exploring how to incorporate inconclusive judgments into the model and how they relate to measures of accuracy and performance.

The model fit in Experiment 1 was used to generate predictions for comparison accuracy in Experiment 3. While the model was successful in predicting the accuracy for many comparisons (see Figures 9 and 10), there were several comparisons for which the model made poor predictions. A close examination of those comparisons revealed that all but one of them were marked as inconclusive by at least one examiner in Experiment 3.

There are several alternative ways of analyzing the data that are still under investigation. First, as mentioned earlier, features could be recomputed based on the final settings instead of using initial values. For example, if contrast was manipulated, it may be more appropriate to use the final contrast setting since this reflects the status of the image at which the identification was made. Using the final settings would mean that the tested images may not directly correspond to those used in Experiment 1, since the performance predictions of the model defined in Experiment 1 were based on the original image settings. This would result in new model predictions for a majority of the tested prints. However, if only one subject made a particular contrast adjustment then there would only be that single evaluation from which accuracy is computed. This would make it difficult to know what true average accuracy would be for a large group of experts and one reason why we did not begin with this analysis. It is interesting to note that perhaps the sequence of image manipulations might collectively be informative for predicting accuracy. For example, seeing the same image at several contrast settings may improve performance compared to seeing an image at just one setting.

Second, instead of using the predictor weights from Experiment 1, a new model could be fit to these data. The weights may be different due to the addition of manipulation features and added evaluation time. For example, if the ability to mark *minutiae* or the number of *minutiae* marked was a very important feature in determining accuracy, the relative importance of the other predictor variables may have been degraded. Similarly, Area Ratio may matter less when more time is provided to compare two images; with little time, a smaller latent area might focus examiner attention in a way that larger areas would not. Given unlimited time, however, regardless of whether the latent area was small or not relative to the known print area, examiners could have compared sections of it at leisure. We would expect to find that many of the same predictors that were important predictors of accuracy in Experiment 1 continue to be so for this experiment. This would confirm that the originally identified image features are indeed relevant for fingerprint identification. How much those features matter, relative to one another, might depend on the exact manner in which the comparison task is set up.

Finally, the manipulations in Experiment 3 might be used to generate new features that reflect examiner behavior. Number and relative spacing of marked *minutiae*, degrees of image rotation, number of image contrast or brightness adjustment steps, or number of levels of zoom might interact with the original set of image features. For example, when Ridge Sum is low (clarity of ridges is poor), marking *minutiae* may correlate with improved accuracy, but may not matter when Ridge Sum is high. It is important to emphasize that such features are not properties of the image themselves, but decisions made by examiners. They cannot therefore be used alone to determine the difficulty of a print, but they may be informative about what kinds of behaviors and procedures are most beneficial to generating a correct identification.

In addition, we may be able to offer insight on the relationship between 'inconclusive' determinations and performance, as well as the relationship between examiners' subjective perceptions of difficulty and their objective performance. We are engaged in further analysis on both of these questions as well.

Policy Implications and Future Research

Experiment 1 was an important step in "unpacking" error rates and their relationship to difficulty, an endeavor that has great importance to forensic science and the legal system. The mere fact that some fingerprint comparisons are highly accurate whereas others are prone to error has a wide range of implications. First, it demonstrates that error rates are indeed a function of comparison difficulty (as well as other factors), and it is therefore very limited (and can even be misleading) to talk about an overall "error rate" for the field as a whole. In this study, more than half the prints were evaluated with perfect accuracy by examiners, while one print was misclassified by 91 percent of those examiners evaluating it. Numerous others were also misclassified by multiple examiners. This experiment provides strong evidence that prints do vary in difficulty and that these variations also affect the likelihood of error. Even though it was a logical assumption that print comparisons would have this quality, establishing this point empirically has significant value. Second, this study lays down a foundation for finding objective print characteristics that can quantify the difficulty of a comparison. The model we offer provides both evidence for what specific visual criteria seem to affect difficulty, as well as a model for combining these criteria to best predict accuracy. This model illustrates the benefits of creating objective measures of difficulty for print pairs, which could be substantially more efficient and consistent than more subjective approaches to assessing difficulty. It also lays the groundwork for further study that can examine the relationship between examiners' subjective assessments of difficulty (Neumann, et al, 2013) and a more objective approach to measuring difficulty of comparisons.

Experiment 2 confirmed the differences found by prior research between novices and experts. Novice performance was essentially at chance and we obtained similar measures to those found by Tangen et al. (2011). Interestingly, we found that even exposure to a short training video seems to alter the way that novices approach the assessment task (though it did not significantly alter their overall accuracy rate). We also found that the image features used by novices were different than those used by experts. This suggests that fingerprint expertise is a perceptual learning process that results in the improved detection of structure and relevant information in fingerprint images. Consequently, procedures that promote perceptual learning (such as sequencing techniques during training; Mettler & Kellman, 2014), may be leveraged to improve training efficiency for fingerprint examiners.

Experiment 3 demonstrated that error rates in more realistic environments were generally comparable to those in Experiment 1. This is a critical finding because it means that valuable experiments with fingerprint examiners can potentially be conducted rapidly, in controlled environments without needing to rigorously replicate the environmental settings in which identifications are normally made. This can save a great deal of time, effort, and money for future research. While realistic, rigorous examination methods are of course preferred in evaluating expertise, one may also be able to generate smaller, less realistic, but similarly accurate testing materials for use during training, for example in creating an online training curriculum. The relative consistency of results between Experiment 1 and Experiment 3 suggests that while greater ecological validity is always to be preferred, valuable information may be acquired through experiments with design constraints as well. Experiment 3 also revealed two additional important findings: (1) a lack of consistency among examiners about which prints were seen to be "inconclusive" and (2) poorer aggregate performance on prints rated "inconclusive" by anyone. This raises interesting questions for further research, as well as important policy questions about where the line between 'inconclusive' and a match conclusion should be drawn.

Consider: Of the 42/120 pairs that at least one examiner labeled inconclusive, five had 50% or more of examiners agree that they were inconclusive, with an average conclusion accuracy of 59.78%. It would seem relatively clear that if we could identify these comparisons in advance, via difficulty ratings, these would be comparisons that ought not to be assessed by examiners at all, given the susbstantial risk of error and the aggregate performance only modestly above chance. But the remaining 37 comparisons that some examiner(s) labeled inconclusive had an average accuracy of 90.19%. That is, to be sure, still a substantially higher error rate than that achieved for the prints no one deemed inconclusive, but it is also quite a high accuracy rate compared to many human endeaors. Would the better practice be for these prints, could they be identified in advance by their visual metrics, not to be assessed or no conclusion offered? Or is a roughly 10 percent chance of error low enough that we would rather have this information than otherwise? Or would it, perhaps, be best to design some special, distinct examination process for this category of prints, to gain the benefits of examiners' best judgments, while recognizing that their high degree of difficulty makes them unusually error-prone? We are still in the process of assessing the relationship between objective visual characteristics and examiner's 'inconclusive' determinations, but this example illustrates how and why objective metrics (either alone, or combined with subjective measurements by examiners) may help the design of appropriate laboratory protocols and more data-driven approaches to the field and its use of information.

Overall, a more sophisticated understanding of the relationship between error rate and difficulty should also be extremely important for the courts in weighing fingerprint evidence (and has been highlighted by the NAS (2009) inquiry into forensic science). Courts are instructed, when assessing expert evidence, to focus on the "task at hand", and this research helps to show that fingerprint examination may vary in difficulty in ways that may be relevant to its evaluation as evidence (Daubert vs. Merrell Dow Pharmaceuticals, 1993; Kumho Tire Co. vs. Carmichael, 1999). More nuanced assessments of fingerprint task difficulty might, for example, affect how a judge understands admissibility of that specific conclusion, or what degree of certainty the expert will be allowed to express, or it might impact the weight given to a specific match conclusion by the fact-finder (Faigman, Blumenthal, Cheng, Mnookin, Murphy & Sanders, 2012). It is possible, for example, that if we could accurately identify the most difficult comparisons, they could be made use of for investigative purposes but not used as evidence in the courtroom. In this way, it is possible that many prints which currently are deemed 'inconclusive' -- and may indeed be difficult enough that they are significantly more prone to error -- could be used to provide valuable, even if more error-prone, information that could assist investigations, rather than have their analysis entirely forgone.

While our model requires further testing, it is possible that it could be piloted in such a way. To be sure, our model does not yet offer the granularity to, say, associate error rates with a set number of distinct levels of difficulty, it could be adapted to examine comparisons and to predict whether they have an unusually high or low difficulty level. The implications of these findings thus have relevance both to the court and more broadly, in that they provide vital insights that can considerably enhance the procedures used in forensic laboratories. Similar to procedures for medical triage, the need for different procedures and checks can be made to fit the difficulty of a comparison.

The understanding of what makes some comparisons more difficult than others also has implications for the selection and training of fingerprint examiners. During selection, benchmarks and skill sets could be set as criteria to ensure candidates have acquired the necessary cognitive abilities needed to perform their job adequately. In addition, in evaluating the significance of errors for trainees, better information about difficulty level will be of great assistance. Trainees who make mistakes on simpler stimuli can be distinguished from those whose errors occur only on more difficult materials; for evaluating performance, all errors are not – and should not be treated as – equal.

While further research is clearly necessary to build on these results, this research therefore provides significant steps forward for helping to establish that error rates are related to difficulty; for beginning to provide validated evidence for what visual dimensions of fingerprint comparison pairs are associated with difficulty; and for helping to tease out both examiner's metacognitive abilities and the substantial degree of examiner expertise in this domain.

Acknowledgments

Portions of this report have been excerpted or adapted from the publications and papers in preparation listed in section VI Dissemination of Research Findings. We would like to thank Clara Sao, Jenny Chun, and Rachel Older for help in the data collection, and Joe Doherty, Matt Thompson, and Jaclyn Seelagy, for helpful comments and suggestions about data analysis and interpretation. We are also grateful to Gerry Laporte at the NIJ for helpful guidance, as well as UCLA for indirect support of the project. We would also like to thank the two anonymous reviewers who provided helpful evaluations and comments on this report. And finally, a great thank you to all the participants who volunteered to take part in the construction of the database and the experimental task.

V. References

Akaike, H (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.) *Second International Symposium on Information Theory*, pp. 267-281. Budapest, Hungary: Akadémiai Kiadó.

Agresti, A (1996). *An introduction to categorical data analysis*. New Jersey: John Wiley & Sons Inc.

Agresti, A (2002). *Categorical data analysis.* New Jersey: John Wiley & Sons Inc.

Ashbaugh DR (1999). *Quantitative-qualitative friction ridge analysis: An introduction to basic and advanced ridgeology*. Florida: CRC Press.

Ashworth, ARS, & Dror, IE (2000). Object identification as a function of discriminability and learning presentations: the effect of stimulus similarity and canonical frame alignment on aircraft identification. *Journal of Experimental Psychology: Applied, 6*(2), 148–157.

Baayen, RH, Davidson, DJ, & Bates, DM (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412.

Bates, D, Maechler, M, & Bolker, B (2012). Lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. http://CRAN.R-project.org/package=lme4

Booth, GD, Niccolucci, MJ, & Schuster, EG (1994). Identifying proxy sets in multiple linear regression: an aid to better coefficient interpretation. Research paper INT-470. United States Department of Agriculture, Forest Service, Ogden USA.

Breslow, NE, & Clayton, DG (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Society*, *88*(421), 9-25.

Bryan, WL, & Harter, N (1899). Studies on the telegraphic language: The acquisition of a hierarchy of habits. *Psychological Review*, *6*, 345-375.

Burnham, KP, & Anderson, DR (2002). *Model selection and multimodel inference: A practical information-theoretic approach.* Second edition. New York: Springer-Verlag.

Busey, TA, Schneider, B, & Wyatte, D (2008). Expertise and the width of the visual filter in fingerprint examiners. Poster presented at the 8th Annual Meeting of the Vision Sciences Society, Naples, FL.

Busey, TA, & Vanderkolk, JR (2005). Behavioral and electrophysiological evidence for configural processing in fingerprint experts. *Vision Research*, *45*, 431-448.

Busey, TA, Yu, C, Wyatte, D, Vanderkolk, J, Parada, F, & Akavipat, R (2011). Consistency and variability among latent print examiners as revealed by eye tracking methodologies. *Journal of Forensic Identification*, *61*(1), 60-91.

Chatterjee, S, & Price B (1991) *Regression analysis by example*. Second Edition. New York: John Wiley & Sons.

Charleton, D, Fraser-Mackenzie, PAF, & Dror, IE (2010). Emotional Experiences and Motivating Factors Associated with Fingerprint Analysis. *Journal of Forensic Science*, 55 (2), 385-393.

Cole, SA (2002). *Suspect Identities.* Cambridge, MA: Harvard University Press.

Cole, SA (2005). Is fingerprint identification valid? Rhetorics of reliability in fingerprint proponents' discourse. *Law & Policy*, *28*(1), 109-135.

Daubert v. Merrell Dow Pharmaceuticals 509 US 579 (1993).

Dixon, P (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, *59*, 447-456.

Dror, IE (2009). How can Francis Bacon help forensic science? The Four idols of human biases. *Jurimetrics*, 50(1), 93-110.

Dror, IE, & Charlton, D (2006). Why experts make errors. *Journal of Forensic Identification*, *56*(4), 600-616.

Dror, IE, Charlton, D, & Péron, AE (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International*, *156*(1), 74-78.

Dror, IE, & Cole, SA (2010). The vision in "blind" justice: Expert perception, judgment, and visual cognition in forensic pattern recognition. *Psychonomic Bulletin & Review*, *17*(2), 161-167.

Dror, IE, Péron, AE, Hind, S & Charlton, D (2005). When emotions get the better of us: the effect of contextual top-down processing on matching fingerprints. *Applied Cognitive Psychology*, *19*(6), 799–809.

Dror, IE, & Mnookin, JL (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, & Risk*, *9*(1), 47-67.

Dror, IE, & Rosenthal, R (2008). Meta-analytically quantifying the reliability and bias ability of forensic experts. *Journal of Forensic Sciences*, *53*(4), 900-903.

Dror, IE, Stevenage, SV, & Ashworth, A (2008). Helping the cognitive system learn: Exaggerating distinctiveness and uniqueness. *Applied Cognitive Psychology*, *22*(4), 573-584.

Expert Working Group on Human Factors in Latent Print Analysis (2012) Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach; NIST Interagency Report 7842.

Faigman, DL, Bllumenthal, JA, Cheng, EK, Mnookin, JL, Murphy, EE, & Sander, J (2012). Modern Scientific Evidence: The Law and Science of Expert Testimony: Fingerprints, Vol. 5 (Ch. 33)

Gelman, A, & Hill, J (2007). *Data analysis using regression and multilevel/hierarchical models.* New York: Cambridge University Press.

Gelman, A, & Su, Y-S (2013). arm: Data analysis using regression and multilevel/hierarchical models. R package version 1.6-04. http://CRAN.R-project.org/package=arm

Gibson, EJ (1969). *Principles of perceptual learning and development.* New York: Prentice Hall.

Green, DM, & Swets, JA (1966). *Signal detection theory and psychophysics.* New York, NY: Wiley.

Haber, L & Haber, RN (2008). Scientific Validation of Fingerprint Evidence under Daubert. *Law, Probability and Risk*, *7*, 87–109.

Hall, LJ, & Player, E (2008). Will the introduction of an emotional context affect fingerprint analysis and decision-making? *Forensic Science International*, *181*(1-3), 36-39.

"How to Compare Fingerprints – The Basics" http://www.youtube.com/watch?v=IrpTqKkgygA

Jaeger, TF (2008) Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434-446.

Kang, J, Bennet, JM, Carbado, D, Casey, P, Dasgupta, N, Faigman, D, Godsil, R, Greenwald, AG, Levinson, J, & Mnookin, J (2012). Implicit Bias in the Courtroom, *UCLA Law Review*, 59, 1124-1185.

Kassin, SM, Dror, IE, & Kukucka, J (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition*, *2*(1), 42-52.

Kellman, PJ (2002). Perceptual learning. In: Pashler H, Gallistel CR, editors. *Stevens' handbook of experimental psychology*, vol. 3 (3rd edition). New York: John Wiley & Sons. pp. 259-299.

Kellman, PJ, & Garrigan, P (2009). Perceptual learning and human expertise. *Physics of Life Reviews*, *6*(2), 53-84.

Kellman PJ, & Massey CM (2013) Perceptual Learning, Cognition, and Expertise. In: Ross BH, editor. *The psychology of learning and motivation* vol. 58. Amsterdam: Elsevier Inc. pp. 117-165.

Koehler, JJ (2008). Fingerprint error rates and proficiency tests: What they are and why they matter. *Hastings Law Journal*, *59*, 1077-1100.

Kovesi, PD (2000). MATLAB and Octave functions for computer vision and image processing. Available from http://www.csse.uwa.edu.au/~pk/research/matlabfns/

Kumho Tire Co. v. Carmichael 526 US 137 (1999).

Langenburg, G (2009). A performance study of the ACE-V process: A pilot study to measure the accuracy, precision, reproducibility, repeatability, and bias ability of conclusions resulting from the ACE-V process. *Journal of Forensic Identification*, *59*(2), pp. 219-257.

Langenburg, G, Champod, C, & Wertheim, P (2009). Testing for potential contextual bias effects during the verification stage of the ACE-V methodology when conducting fingerprint comparisons. *Journal Forensic Sciences*, 54, 571-582.

Maltoni, D, Maio, D, Jain, AK, & Prabhakar, S (2009) *Handbook of fingerprint recognition*. London: Springer.

Marcon, JL (2009) *The distinctiveness effect in fingerprint identification: How the role of distinctiveness, information loss, and informational bias influence fingerprint identification* (Doctoral dissertation). Available from *ETD Collection for University of Texas, El Paso (*Paper AAI3358893).

Mettler, E, & Kellman PJ (*In Press*). Adaptive response-time-based category sequencing in perceptual learning. *Vision Research.*

Mnookin, JL (2001). Fingerprint evidence in an age of DNA profiling. *Brooklyn Law Review*, 67, 13-70.

Mnookin, JL (2008a). Of black boxes, instruments, and experts: Testing the validity of forensic evidence. *Episteme*, *5*(3), 343-358.

Mnookin, JL (2008b). The validity of latent fingerprint identification: Confessions of a fingerprinting moderate. *Law, Probability & Risk*, *7*, 127-141.

Mnookin, JL The Courts, The National Academy of Science, and the Future of Forensic Science, 75 Brooklyn Law Review 75, 1209-1276 (2010) (The Ira. M. Belfer Lecture, 2009).

National Research Council, National Academy of Sciences (2009). *Strengthening Forensic Science in the United States: A Path Forward.* Washington, DC: National Academy Press.

NIST, Latent Print Examination and Human Factors: A Systems Approach (2012).

Neter, J, Kutner, MH, Wasserman, W, & Nachtsheim, C (1996). *Applied linear statistical models.* Fourth Edition. USA: McGraw-Hill/Irwin.

Neumann, C, Champod, C, Yoo, M, Gennessay T, & Langenburg, G (2013). Improving the understanding the reliability of the concept of "sufficiency" in friction ridge examination. https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=266312

Oppenheimer, DM (2008). The secret life of fluency. *Trends in Cognitive Science*, *12*(6), 237-241.

Pulsifer, DP, Muhlberger, SA, Williams, SF, Shaler, RC, Lakhtakia, A (2013). An objective fingerprint quality-grading system, *Forensic Science International*, *231*, 204-207.

Schiffer, B, & Champod, C (2007). The potential (negative) influence of observational biases at the analysis stage of fingermark individualisation. *Forensic Science International*, *167*, 116-120.

Schneider W, & Shiffrin RM (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychology Review*, *84*, 1-66.

Shen, W, & Eshera, MA (2003). Feature extraction in fingerprint images. In N. Ratha and R. Bolle (Eds.), *Automatic fingerprint recognition systems* (pp. 145-182). New York: Springer-Verlag.

Tangen, JM, Thompson, MB, & McCarthy, DJ (2011). Identifying fingerprint expertise. *Psychological Science*, *22*(8), 995-997.

Thai, K, Mettler, E, & Kellman, PJ (2011). Basic information processing effects from perceptual learning in complex, real-world domain. In L. Carlson, C. Holscher, & T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 555-560). Boston, MA: Cognitive Sciences Society.

Ulrey, BT, Hicklin, RA, Buscaglia, J, & Roberts, MA (2011). Accuracy and reliability of forensic latent fingerprint decision. *Proceedings of the National Academy of Sciences*, *108*(19), 7733-7738.

Ulery, BT, Hicklin, RA, Buscaglia, J, & Roberts, MA (2012). Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners. *PLoS ONE 7*(3), e32800.

Vokey, JR, Tangen, JM, & Cole, SA (2009). On the preliminary psychophysics of fingerprint identification. *Quarterly Journal of Experimental Psychology*, *62*, 1023-1040.

Wertheim, K, Langenburg, G, & Moenssens, A (2006). A report of latent print examiner accuracy during comparison training exercises. *Journal of Forensic Identification*, 56, 55-93.

Wickens, TD (2002). *Elementary Signal Detection Theory.* Oxford University Press. New York: New York.

Zuur, AF, Ieno, EN, Walker, NJ, Saveliev, AA, & Smith, GM (2009). *Mixed effects models and extensions in ecology with* R. New York: Springer Science+Business Media, LLC.

VI. Dissemination of Research Findings

Journal Articles

Kellman, PJ, Mnookin, J, Erlikhman, G, Ghose, T, Garrigan, P, Mettler, E, Charlton, D, & Dror, I. . Forensic Comparison and Matching of Fingerprints: Using Quantitative Image Measures for Estimating Error Rates through Understanding and Predicting Difficulty, PLOS ONE, 9(5): e94617. doi:10.1371/journal.pone.0094617

Erlikhman, G, Kellman, PJ, Dror, I,, Ghose, T, Garrigan, P, Charlton, D, & Mnookin, J Experts and Novices: Using Fingerprint Image Features to Understand Performance . *In Preparation*

Mnookin, J, Dror, I., Doherty, J., Seelgacy, J., Erlikhman, G, , Garrigan, P, & Kellman, PJ, Metacognition, Recognizing Difficulty, and Expert Fingerprint Examiners. *In Preparation*

Conference Presentations focusing on grant research include:

Ghose, T., Erlikhman, G., Garrigan, P., Mnookin, J., Dror, I., Charleton, D., & Kellman, P.J. (2013). Perception, Image Processing and Fingerprint-Matching Expertise. *European Conference for Vision and Perception, August 2013*

Erlikhman, G., Ghose, T., Garrigan, P., Mnookin, J., Dror, I., Mettler, E, Charleton, D., & Kellman, P.J. (2013). Fingerprint matching expertise and its determinants. *Vision Sciences Society, May 2013*

Additional Presentations referencing research (partial):

Fingerprint Evidence and Current Research: presentation at NACDL/Cardozo Law School National Forensic Science College, June 2014

Keynote presentation on "A Cognitive Perspective on Expert Evidence and the Administration of Justice" at Project Innocence Annual Meeting, Portland, 11 April 2014.

Invited presentation on "The Human Factor in Forensic Science" at the University of Amsterdam, 11 October 2013.

Keynote presentation on "Distributed Cognition Between Humans and Technology" at Biometrics Institute Technology Showcase, 27 June 2013.

The Sir Michael Davies Keynote presentation on "Experts: The Myth of Impartiality" at the Expert Witness Institute (EWI) Annual Meeting, London, 5 June 2013.

Second Workshop on "Cognitive Factors in Making Forensic Comparisons", at the London Metrapolitan Police, 22 May 2012.

Keynote presentation on "Distributed Cognition Between Human Experts and Technology" at the Annual User's Education Conference, Bellevue, 8 May 2012.

Invited presentation on "Psychology and the Law: Cognitive failing in administering justice, and how psychology can help the criminal justice system" at the University of Seattle, WA, 6 May 2013.

Keynote presentation on "Cognitive Forensic" at the Forensics Europe Expo Conference, London, 24 April 2013.

First workshop on "Cognitive Factors in Making Forensic Comparisons", at the London Metrapolitan Police, 17 April 2012.

Invited presentation on "Perception and Judgments of Human Experts: The Role of Contextual Information" at the Department of Psychology, University College London (UCL), 12

February 2013.

Invited presentation on "Cognitive Underpinning of Expertise: Why & How Forensic and Medical Experts Make Errors, and How to Minimise Them" at the Department of Psychology, Warwick University, 7 February 2013.

Second 1-day Workshop on "Cognitive Factors in Making Forensic Comparisons", at The Netherlands Forensic Institute (NFI), 10 December 2012.

A 2-day workshop on "The Human Element and Cognition in Biometric Identification" at the Biometric Institute, 28-29 November, 2012.

Keynote presentation on "Brain Friendly Biometric Systems: Effective Distributed Cognition Between Humans and Technology" at 8th Biometrics Institute Technology Showcase & Exhibition, 27 November 2012.

Invited presentation on "Contribution of Cognitive Psychology: Reliability and Biasability of Experts in the Court Room" at the University of New South Wales, Sydney, 23 November 2012.

Invited presentation on "Cognitive Forensics: Increasing expertise in forensic science" at the Centre for Forensic Science, University of Technology, Sydney, 21 November 2012.

A 2-day workshop on "Cognitive Factors in Making Forensic Comparisons" at the Australian National Institute of Forensic Science (NIFS) and Australian New Zealand Policing Advisory Agency (ANZPAA), 19-20 November 2012.

A 2-day workshop on "Cognitive Factors in Making Forensic Comparisons" at Victoria Police, 15-16 November 2012.

Invited presentation on "Cognitive Forensics: Identifying and Mitigating Bias in Criminal Cases" at the Criminal Bar Association Autumn Conference, 3 November 2012.

A 1-day workshop on "Improving Forensic Decision Making" at the Colorado Bureau of Investigation (CBI), 17 September 2012.

Plenary presentation on "Cognitive Forensics, Expertise, the Biasing Snowball Effect, and Context Management in Forensic Investigations' at the 6th Annual European Academy of Forensic Science Conference, 23 August 2012.

Invited presentation on "Expert Evidence: The Good, the Bad, and the Ugly" at the After Court Seminar program for High Court Judges, Deputy High Court Judges, Judges of the Court of Appeal, and Justices of the Supreme Court, Royal Courts of Justice, 13 June 2012.