

Are Forensic Scientists Too Risk-Averse?

Willa M. Mannering¹, Macgregor D. Vogelsang¹, Thomas A. Busey¹, Fred L. Mannering²

¹Department of Psychological and Brain Sciences, Indiana University

²Department of Civil and Environmental Engineering, University of South Florida

This version is the text just prior to formatting by the journal. For the formatted version, consult the following reference:

Mannering, W. M., Vogelsang, M. D., Busey, T. A., & Mannering, F. L. (2021). Are forensic scientists too risk averse? *J Forensic Sci.* doi:10.1111/1556-4029.14700

Abstract

Fingerprint examiners maintain decision thresholds that represent the amount of evidence required for an Identification or Exclusion conclusion. As measured by error rate studies (1), these decision thresholds currently exhibit a preference for preventing erroneous identification errors at the expense of preventing erroneous exclusion errors. The goal of this study is to measure the decision thresholds for both fingerprint examiners and members of the general public, to determine whether examiners are more risk-averse than potential jury members. To externally measure these decision thresholds, subjects manipulated decision criteria in a web-based visualization that reflects the tradeoffs between erroneous identification decisions and erroneous exclusion decisions. Data from fingerprint examiners and the general public were compared to determine whether both groups have similar values as expressed by the placement of the decision criteria. The results of this study show that fingerprint examiners are more risk averse than members of the general public, although they align with error rate studies of fingerprint examiners. Demographic data demonstrate those factors that may contribute to differences in decision criterion placement, both between the two groups and between individuals within a group. The experimental methods provide a rich framework for measuring, interpreting, and responding to the values of society as applied to forensic decision making.

Keywords: *decision making, erroneous identification, signal detection theory, fingerprint examination, econometrics, multivariate data analysis*

HIGHLIGHTS

- No sufficiency standards for fingerprint comparisons exist.
- We created a tool to allow participants to assign relative importance values to errors.
- Examiners were more risk-averse than non-examiners.
- Multivariate statistical analyses identify demographic variables associated with criterion placement.

The tension between acquitting the guilty and convicting the innocent can be traced as far back as the 12th century scholar Moses Maimonides, who is quoted as: “*It is better and more satisfactory to acquit a thousand guilty persons than to put a single innocent one to death.*” (2) In the West, this became known as Blackstone’s Ratio: “*It is better that ten guilty persons escape than that one innocent suffer.*” (3) Both of these philosophical statements express the idea that any decision in the judicial system has the potential for two kinds of error, and both attempt to make reasoned arguments for where the decision thresholds should lie with respect to each kind of error. This tradeoff between errors does not lie solely at the level of acquittal or conviction at trial, but instead extends to *every decision made during a criminal investigation*. For example, what is the standard for a warrant? If the standards are relaxed, it may reveal more criminal activity but more innocent may be inconvenienced or even unjustly charged.

This warrant example has a statute determined by policymakers or judges, who act as proxy for the greater society. However, within the fingerprint comparison discipline (or indeed most pattern comparison disciplines), *there are no institutionally-mandated standards for what constitutes sufficient evidence to render a conclusion*. Instead, the community of fingerprint examiners has collectively determined what constitutes sufficiency, and this has been expressed by the governing body SWGFAST into a definition for Identification (termed ‘Individualization’ at the time that the document was created):

Individualization is the decision by an examiner that there are sufficient features in agreement to conclude that two areas of friction ridge impressions originated from the same source. Individualization of an impression to one source is the decision that the likelihood the impression was made by another (different) source is so remote that it is considered as a practical impossibility. (4)

Note that this definition of the identification threshold contains no quantification, and merely attempts to capture the relative strength of evidence that lead to the examiner adopting the Identification conclusion. Updated language has been proposed by the Organization of Scientific Area Committees (OSAC), but the overall sentiment remains the same: The definition fails to make reference to quantitative standards, and in fact explicitly rejects numerical thresholds such as ‘7 corresponding minutiae’. Instead, the examiner is allowed to specify their own personal decision threshold for what they consider sufficiency for purposes of an identification conclusion when conducting casework. In addition, the evidence is accumulated in the mind of the examiner, and compared against an internal decision threshold that is then expressed using categorical conclusions such as Identification or Exclusion.

This approach gives the examiner wide latitude when conducting comparisons, in that they can rely on information other than corresponding minutiae. However, the decision thresholds adopted by examiners directly affect factors that are valued by the public: the safety of our communities as well as the number of innocent persons in prison. Changes to these decision thresholds affect both factors in complex ways. More importantly, whether the decision thresholds adopted by examiners corresponds to the values of society is, as yet, untested, primarily because it is difficult to measure the values given the complex tradeoffs that occur when different thresholds are adopted.

Dror (5) makes the important point that an understanding of errors rates is necessary for both the improvement of a discipline as well as determining the reliability of the forensic evidence: While there are potentially serious consequences to sending an innocent person to prison, there are also potentially serious consequences to letting a guilty person go free. Resolving the issue of which error is more important, and by how much, is central to the issue of

measuring the error rate for a discipline. Most discussions of error focus on the false positive errors (6), and examiners report worrying more about making erroneous identifications than missed identifications:

“Fear? Only fear of making a wrong decision...I think. I think that’s the fear. So you just wanna be sure that you have made the right decision so you will probably err on the side of safety because it’s better to let the ident go than to make the wrong ident.” (7).

This one-sided focus on erroneous identifications can be problematic: Dror (5) argues persuasively that if error rates are only based on false positive errors, examiners can artificially reduce errors by simply making more inconclusive decisions (therefore increasing the missed identification rate).

How do we determine the relative weight to assign to each of the two types of errors? The present work introduces a novel paradigm that measures the sufficiency standards within an interactive visualization that captures the complex tradeoffs that occur during forensic decision making. We use this paradigm to compare the values expressed by the public with those expressed by fingerprint examiners. The decision thresholds used by fingerprint examiners may introduce an inherent risk aversion, which may not align with the values that members of the general public hold with respect to important issues such as keeping innocent persons out of prison and keeping communities safe.

Below we provide brief context for the latent print examination, to demonstrate how an examiner’s decision thresholds can alter the outcomes of the comparison. A latent print examiner compares latent fingerprints from a crime scene to candidate exemplar fingerprints taken from a suspect to determine whether the two impressions may have originated from the same finger. To make this decision, they compare the amount of perceived detail in agreement between the two impressions and look for unexplainable differences. Most fingerprint comparisons are conducted

by human examiners, not computers, and the lack of guidance from policymakers on what constitutes sufficient evidence for a conclusion leads to a morally ambiguous task for fingerprint examiners. As an alternative to a mandated, quantifiable standard, the forensic community has developed a set of procedures known as ACE-V, which stands for Analysis, Comparison, Evaluation and Verification (8-10). ACE-V documents the steps examiners take to extract and compare features (typically minutiae). However, ACE-V does not specify the complete set of features that can be used. In addition, while some labs have informal guidelines for thresholds (such as 12 matching minutiae or ‘points’), examiners are generally free to make decisions based on all of the available information (a point-based system also suffers from the ambiguity of what constitutes a point and whether it corresponds with a region on the comparison impression).

In lieu of an externally defined decision threshold, examiners typically rely on experience with known mated and non-mated impressions to develop a personal, internal threshold for what constitutes sufficiency for purposes of identification. Verification from one or more additional examiners improves the decision-making process through conflict resolution, but this demonstrates reliability rather than accuracy because the ground truth is rarely known in casework. Error rate studies with ground-truth stimuli address accuracy issues under experimental contexts (1). However, in casework, examiners are hampered by the fact that much of the evidence is perceptual in nature, may lie below the level of consciousness, and therefore be difficult to verbalize (11, 12). Thus, they face the dual challenge of communicating the results of their examination to other stakeholders, as well as establishing a shared decision threshold so that the outcome of a particular examination is not dependent on which examiner considers the evidence (13).

All of these procedures create an ambiguous and potentially concerning state of affairs: examiners have collectively decided what constitutes sufficiency for purposes of identification and have promulgated and enforced this threshold through a loose collection of verification, proficiency tests, and legal challenges. There are no governing bodies that have specified a particular threshold, nor what information must be used, although organizations such as the National Institute of Standards and Technology (NIST), the National Institute of Justice (NIJ), and the Organization of Scientific Area Committees (OSAC) have played a supportive role to develop extended feature sets and create standards for analyzing evidence and communicating results. Judges conduct evidence admissibility hearings to determine whether the science is sound, which typically involves validation through error rate studies. However, the thresholds that are revealed by error rate studies have not undergone the political and legal process by which other standards such as obtaining a warrant were established. Therefore, it is unknown whether these standards reflect the values of our legal system and society at large.

How can we determine whether the current thresholds adopted by examiners for fingerprint comparisons are appropriate? The appeal to philosophy such as Blackstone's Ratio provides little solace, because it tends to lead to 'virtue one-upmanship': you might want 10 criminals to go free for every innocent in prison, but someone else might argue for 100 or 1000. Optimizing this ratio and therefore a sufficiency threshold is complex, because an optimal solution depends on several factors, including: (1) the cost and benefits to society of various outcomes, including both erroneous identifications and correct identifications; (2) the ability of examiners to separate mated from non-mated pairs (which depends on both the quality of the images and the training of the examiners), (3) the prior probability of mated and non-mated pairs (essentially how good the detectives are at finding the correct suspect) and (4) how the evidence

is used to ultimately achieve a result in a criminal trial. As we will see, there is no perfect placement of the thresholds that will eliminate all errors. However, there may be an *optimal* placement of the thresholds that reflects the values of society and therefore indirectly estimates the cost of various outcomes.

We have three goals. First, measure the decision thresholds for identification and exclusion decisions in both fingerprint examiners and members of the general public, to determine whether they align. Second, determine whether the measured decision criteria for fingerprint examiners aligns with the thresholds inferred from error rate studies (1). Finally, determine which demographic attributes are associated with different decision threshold placement. Although the focus of our particular measurement tool is on the placement of a decision criteria, our methods allow us to infer the value that each participant places on the errors that contribute erroneous information that might put an innocent person in prison or allow a guilty person to avoid the consequences of their actions. Below we provide additional details and context for latent print comparisons, which will illustrate several assumptions and models that provides a structure to address these questions.

2. Background and Motivation

During normal casework, fingerprint examiners compare latent fingerprints processed from a crime scene and exemplar prints taken in a controlled environment with a known donor individual. Latent prints are often distorted, incomplete, and corrupted by visual noise. Exemplar prints are typically higher quality than the latent prints. The ground truth (typically unknown and unknowable by the examiner) can either be mated fingerprints or non-mated fingerprints. Mated fingerprints are pairs of latent and exemplar fingerprints that originate from the same finger.

Non-mated fingerprints are pairs of latent and exemplar fingerprints that originate from different fingers. Because the status of a pair of fingerprints is rarely known outside of an experimental context, the examiner must make a conclusion that represents their expert opinion. To do this, an examiner conducts an analysis and comparison of the two prints to make one of three decisions about a pair of fingerprints: identification, exclusion, or inconclusive. An identification decision means that the examiner believes there is enough perceived detail in agreement between two fingerprints to say the fingerprints came from the same finger, and most examiners subscribe to the SWGFAST definition described above. An exclusion decision means that the examiner believes there is either not enough detail in agreement or that there are sufficient details in disagreement between the two fingerprints to say they did not come from the same finger. An inconclusive decision means the examiner believes there is not sufficient detail in agreement or disagreement to make an identification or exclusion decision. While there are variations on this decision structure (for example, some agencies might use a ‘could not exclude’ conclusion in place of an identification), these are the categories that are typical across agencies.

3. Error Rates in Fingerprint Comparisons

Ulery et al. (1) measured the accuracy and reliability of latent fingerprint examiners’ decisions with a study of 169 latent print examiners who each compared approximately 100 pairs of latent and exemplar fingerprints from a pool of 744 pairs. There are two kinds of errors: Five examiners made erroneous identification errors for an overall erroneous identification rate of 0.1%, while 144 out of the 169 examiners made at least one erroneous exclusion error for an overall erroneous exclusion rate of 7.5%. Further, 31.1% of the total mated fingerprints were

classified as inconclusive and 11.1% of the non-mated fingerprints were classified as inconclusive.

This study brought attention to what might be viewed as risk-aversion in latent fingerprint conclusions: examiners are much more likely to make an erroneous exclusion error than an erroneous identification error. This asymmetric error pattern has also been replicated: Tangen, Thompson, and McCarthy (14) found a 7.8% erroneous exclusion rate and a 0.68% erroneous identification rate. These results reinforce the attention given to erroneous identifications as opposed to erroneous exclusions when considering the risk of different errors.

Ulery et al.'s (1) study also revealed a large inconclusive rate, raising the possibility that examiners fall back on inconclusive decisions in order to prevent making (potentially serious) errors. Of course, if examiners make a large number of inconclusive decisions then fewer crimes will be solved. This complicates the set of tradeoffs that occur when examiners must decide where to place their decision criteria, which then determines how much evidence is required before making an exclusion or identification conclusion. As we measure the values of the different outcomes of latent print examinations, we might find that the general public is less concerned with innocent people being put in prison and are intolerant of the large percentage of inconclusive decisions found by Ulery et al. (1) that might contribute to increases in crime rates.

We cannot simply ask participants whether it is better to put innocent persons in prison or let guilty persons go free, because the tradeoff is quite complex and depends on multiple factors, including how easy it is to separate mated from non-mated pairs, as well as the proportion of mated pairs in the comparison set. We need an accurate representation of the quantitative nature of the tradeoffs that occur given these factors. To do this, we modeled the error rate data of Ulery et al. (1) using signal detection theory (15), as described next.

4. Modeling the Decision Tradeoff with Signal Detection Theory

In order to estimate the tradeoffs that occur as an examiner adopts different decision thresholds, we constructed a mathematical representation of the underlying distributions of mated and non-mated fingerprint pairs taken from the Ulery et al. (1) error-rate data as reproduced in Table 1.

[Insert Table 1]

These response proportions are the result of both the ability of examiners to separate mated from non-mated pairs, as well as the decision thresholds adopted by examiners. However, by constructing a model of the underlying distributions that produced these response proportions, we can estimate the consequences of other choices of decision thresholds. To model these underlying distributions of mated and non-mated pairs of impressions, we assume that a comparison results in some amount of perceived detail in agreement, and that larger values along this unidimensional evidence axis are more likely to produce an identification decision. If there is a large amount of perceived detail in agreement, an examiner is likely to produce an identification decision. However, this internal evidence value is not typically stated by the examiner, who instead only reports the results of the comparison as exclusion, inconclusive, or identification. As a result, the distribution of internal evidence values across many different comparisons must be inferred using the assumptions underlying signal detection theory (SDT).

To characterize the distribution of the amount of perceived detail in agreement for both mated and non-mated distributions, we fit a model to the Table 1 proportion data using the following assumptions. First, we assumed that the distribution of evidence scores is normally distributed and allowed the mated distribution to have a different variance than the non-mated

distribution. The non-mated distribution is fixed with a mean of zero and a standard deviation of 1.0, which sets the scale of the evidence axis. Four free parameters were then adjusted: the location of the mated distribution along the evidence axis, the standard deviation of the mated distribution, and the two decision criteria (one that separates exclusion from inconclusive responses and the other that separates inconclusive from identification responses).

These parameters were fit to the response proportions shown in Table 1 using maximum likelihood estimation, fit to the entire dataset rather than individual subjects because the discipline-wide data is most relevant when considering error rates for the field as a whole. We use optimization procedures to find parameter values such that the predicted proportions of different responses were as close as possible to the obtained proportions of responses. Figure 1 is a graphical representation of the results of the signal detection modeling. The best fitting parameters were: mated mean = 3.42, mated standard deviation = 1.54, exclusion criterion = 1.21, and identification criterion = 2.97. These estimates imply that the mated distribution is slightly more spread out than the non-mated distribution, and that examiners have adopted an extremely conservative threshold for the identification criterion, given that it is almost 3 standard deviations away from the center of the non-mated distribution. The predicted proportions from signal detection theory were able to reproduce the results illustrated in Table 1.

[Insert Figure 1]

This close correspondence is of course expected, because there are four free parameters and four degrees of freedom in the data. However, we argue that the utility of this model derives from its ability to accurately capture the nature of the decision tradeoff near where estimates were actually derived (i.e. near the identification decision criterion) and therefore this model

serves as a useful formalization of the decision tradeoff despite the inability to directly assess the goodness of fit.

The summary of the Ulery et al. (1) data by the signal detection theory model allows for the exploration of the consequences of implementing different decision thresholds. For example, if examiners were to adjust the identification criterion to the left (say adopting a value of 2.5 instead of 2.97) both the number of potential criminals in prison would increase and the number of potential innocent people in prison would increase. However, each group would not increase by the same amount, or even proportionately. Instead, the amount that each group would increase can be determined by asking how much more area under the non-mated and mated distributions falls to the right of the new location of the identification criterion.

To explore the decision tradeoff in forensics, we developed a web-based visualization (see Figure 2 and Methods for more detail) based on this model. The visualization quantifies the decision tradeoff in such a way that we can explore the values expressed by different participants. For example, if a subject is uncomfortable with a particular number of potential innocent people in prison, they can shift the decision criteria toward a more risk-averse value. However, this will simultaneously affect the number of potential criminals in prison, which will drop by an amount determined by the SDT model. The visualization provides both a graphical representation of the two distributions and immediate feedback for the consequences of different decision criterion choices. We also collected data from fingerprint examiners, and thus we can compare experts to members of the general public to determine whether the values expressed by examiners align with the values of society, and data from both groups can be compared to error rate studies to determine whether the values correspond to the *actual* thresholds used when conducting comparisons. We also use demographic data to determine the factors that are

associated with different resolutions of the decision tradeoff that exists in forensic decision-making to determine the possible sources of expert/general public differences.

Why might examiners differ from members of the public in terms of their beliefs of the sufficiency required to reach a decision in a forensic comparison? In Table 2 we consider a number of possible differences and discuss how each might affect the decision thresholds.

[Insert Table 2]

As noted by Dror (13), there are important expert/general public differences as experts gain skill and knowledge in their discipline. Dror distinguishes between the information observed and the conclusions made on the basis of this information, and how external contextual information might produce differences both between and within examiners for both observed information and conclusions derived. Whether changes in the decision threshold occur ('biasing') depends on the difficulty of the comparison, the direction of the biasing information, and the strength of the biasing information. Some elements of expertise may protect against such biasing, such as robustness to noise and inversion (16), configural processing of the images (17, 18). However, expertise also brings blind spots: these same abilities such as automaticity, schemas, selective attention, and configural processing may also serve to limit the information acquired by an examiner, which may lead to overconfidence (19-21). This might lead an examiner to become less risk averse), there are important expert/general public differences as experts gain skill and knowledge in their discipline. Dror distinguishes between the information observed and the conclusions made on the basis of this information, and how external contextual information might produce differences both between and within examiners for both observed information and conclusions derived. Whether changes in the decision threshold occur ('biasing') depends on the difficulty of the comparison, the direction of the biasing information,

and the strength of the biasing information. Some elements of expertise may protect against such biasing, such as robustness to noise and inversion, configural processing of the images (17, 18). However, expertise also brings blind spots: these same abilities such as automaticity, schemas, selective attention, and configural processing may also serve to limit the information acquired by an examiner, which may lead to overconfidence (19-21). This might lead an examiner to become less risk averse.

Several different factors related to knowledge of the role fingerprints play in the judicial system would likely lead to less risk aversion, because examiners are comforted by the fact that there are safeguards beyond their decision, and therefore presumably would be more willing to make an identification knowing that it likely would not be the sole evidence that convicts an individual.

Demographics, a greater understanding of the way that fingerprints are compared, an overall greater scientific background, and motivation may all differ between the general public and experts. However, we found it difficult to make a compelling *a priori* prediction for whether this should make experts more or less risk averse. However, we note these for completeness, and some of these can be addressed using our demographics.

Finally, there are two strong motivations that would lead an examiner to become more risk averse than the general public. The first is related to the issue of close non-mated impressions. These are typically recovered from database searches, and examiners with more experience are more likely to have encountered these and therefore become more cautious.

The second motivation is more personal for examiners. Erroneous identifications have the potential to put innocent persons in prison, which creates problems for the incarcerated individual, the victim and society (because the true perpetrator is still at large), the laboratory,

and the examiner. In this case the examiner may face both professional and personal (i.e. guilt) consequences if discovered to have been wrong in this way. However, erroneous exclusions do not garner the same attention, because a criminal who escapes prosecution is not likely to make that known. Thus, there is a strong set of incentives to avoid erroneous identifications at the potential costs of erroneous exclusions or inconclusive conclusions. This, we would argue, is a strong motivator that could make examiners risk averse when making identification decisions.

One final point before introducing the methods of our paradigm: We are modeling the tradeoffs that occur in any forensic decision-making task results in a conclusion drawn from a discrete set. We are not modeling the actual comparison itself, nor the mental processes that allow an examiner to produce a value along a unidimensional decision axis. The exact task is somewhat arbitrary in that our paradigm could be applied to footwear, firearms comparisons, or any discipline where examiners convert a continuous internal evidence value into one of a few discrete conclusions. We have chosen to ground our task in a fingerprint comparison domain because the error-rate data was available that allows us to quantify the nature of the decision tradeoff at different decision threshold values through the signal detection theory model. Any model will be imperfect, but by grounding our visualization in high-quality error rate data we argue that we can fairly represent the tradeoffs that occur as a decision threshold is varied, and do so in a way that is consistent with the strength of fingerprint evidence. Our goal is to identify differences between experts and the general public, as well as compare both against the data from error rate studies, to provide the kind of feedback that will allow fingerprint examiners to adjust their decision thresholds to be more in line with the values of the public at large. This approach also allows us to *directly* compare the values of examiners and the general public with

the error-rate data, because it places all of the data in the same format (the location of the identification decision criteria along the evidence axis).

5. Methods

A total of 455 subjects were considered for data analysis. The subjects used for analysis were split into two groups: examiners and the general public. There were 377 general public subjects were considered for analysis. 172 were undergraduates attending Indiana University, members of the Bloomington, Indiana community, or members of the surrounding community. The remaining 205 general public subjects were recruited from Amazon's Mechanical Turk. The undergraduate subjects performed the experiment in a computer lab and received course credit as compensation for participating. In an attempt to get qualified subjects from Mechanical Turk, the subjects were filtered based on location (in the US), whether they were eligible to serve on a jury, and their previous performance record. Mechanical Turk subjects who were selected performed the experiment remotely and were compensated 1 USD for their participation. The other general public subjects participated voluntarily and were not compensated. 78 examiner subjects participated in this study and were recruited from national forensic conferences such as the International Association for Identification and the Cogent Users Group International workshop. These subjects accessed the experiment through a web link and performed the experiment remotely as well.

All subjects watched a 6-minute instructional video before being directed to manipulate the web-based visualization (Figure 2). The transcription of the instructional video is available in Appendix A. After saving their exclusion criterion and identification criterion placements, subjects were asked to fill out demographic data. The experiment took approximately 15 minutes to finish.

Figure 2 illustrate the online visualization tool used for data collection. The actual visualization is still available, and can be accessed using the link below, and the reader is encouraged to visit the following site to quickly understand the interface and the nature of the decision tradeoff: <https://buseylab.sitehost.iu.edu/fingerprintvalues/?sandbox>.

The reader is encouraged to try this visualization in order to understand the dynamic nature of our task, which is difficult to convey in images alone. There is a ‘continue’ button that allows the video to be skipped (although this was not available to our participants who were required to watch the entire video before proceeding to the visualization).

[Insert Figure 2]

The horizontal axis in Figure 2 represents an evidence axis, which we described as the amount of perceived detail in agreement between two impressions. The two point-clouds represent the mated (top) and non-mated (bottom) pairs of impressions, and are jittered vertically to increase the visibility of individual points. During a comparison, the examiner collects evidence regarding the similarity and dissimilarity between two prints and weights each piece of evidence appropriately and internally to create the overall perceived detail in agreement. Another description of the horizontal axis is the weight of evidence in favor of one of two hypotheses: the two impressions come from the same source, and the two impressions come from different sources. The ‘perceived detail in agreement’ language assumes that the examiner appropriately weights the details in agreement and detail in disagreement to derive a location along the evidence axis that corresponds to a weight of evidence. This is necessary because some impressions recovered from database searches can share a great deal of similarity with a latent print, yet have one or two clear disagreements that make an identification decision inappropriate.

Each point in Figure 2 corresponds to one comparison between two impressions, and the horizontal location of each point represents the hypothetical amount of perceived detail in agreement an examiner observes over the course of a comparison as estimated from Signal Detection Theory. The precise location of each point along the horizontal axis was determined by sampling randomly from normal distributions as specified by the signal detection model for the mated and non-mated distributions. This random sampling was repeated for each subject, resulting in a new distribution of points.

The points in Figure 2 are partitioned by two decision criteria into three groups that represent the decisions made by the examiner. These decision criteria are controlled by the participant via interactive sliders. The first slider represents the exclusion criterion—all points to the left of this slider are exclusion decisions. The second slider represents the identification criterion—all points to the right of this slider are identification decisions. All points between the two sliders represent an inconclusive decision. The boxes below the graph explain the outcomes of placing the sliders in a specific position and are color coordinated to indicate good outcomes (green), bad outcomes (red), and inconclusive outcomes (yellow). The colors of the clouds match the colors of the boxes in the same positions; for example, the green cloud in the upper right-hand corner of both figures correspond to the outcomes in the green box in the upper right-hand corner. As the sliders are moved, the number of cases in each box changes to reflect the number of points that fall in each of the six outcomes.

5.1 Subjective Utilities and Prior Probabilities

The decision tradeoff occurs as the identification or exclusion criterion is shifted along the horizontal axis, because such shifts change both the number of innocent people potentially in

prison and the number of guilty people potentially in prison. The exact nature of the decision tradeoff is determined by the model of the underlying mated and non-mated distributions as estimated by signal detection theory. The critical elements for this tradeoff include the amount of overlap between the two distributions (d' in signal detection terminology) as well as the variance of the mated distribution, both of which are provided by the signal detection theory model fit. However, a third factor also affects the decision tradeoff: the ratio of mated and non-mated distributions. This can be thought of as the prior probability of a detective providing a mated impression. Any attempt to estimate the utilities of the various outcomes of a fingerprint comparison will, in principle, depend in part on these priors. The visualization in Figure 2 uses the same base rates as provided by the original Ulery et al. (1) dataset: 5969 mated and 4083 non-mated impressions. However, these base rates were chosen based on criteria related to their experimental constraints and may or may not reflect the true priors. In fact, the true priors are likely situationally dependent, vary from jurisdiction to jurisdiction, and no current accurate estimate of these priors exists in the literature. Thus, the true value of the prior is difficult to determine.

Despite not having an accurate estimate of the prior probability of a mated pair, we can still assess whether examiners and members of the general public are sensitive to large changes in the priors. We did this by reducing the number of mated pairs while leaving the number of non-mated pairs the same. The bottom visualization in Figure 2 illustrates this dramatic shift in the ratio of mated to non-mated pairs, by reducing the number of mated points from 5969 to 1000. This value was arbitrarily chosen but reflects a situation in which one in five comparisons conducted by examiners is on a mated impression. To assess whether participants are sensitive to this prior, we altered the number of mated pairs as shown in the bottom visualization of Figure 2

for half of our participants while keeping the number of non-mated pairs unchanged. For comparison, the top visualization in Figure 2 represents the high mated condition, and the bottom visualization represents the low mated condition. If we observe differences between these two conditions, this would demonstrate that participants are sensitive to the prior probabilities when considering the placement of the identification and exclusion criteria.

5.2 Procedure

The subjects were randomly assigned one of the two conditions (high mated or low mated) for the web-based visualizations. In the high mated condition, the top cloud shown to the subjects was dense and consisted of 5969 individual points. This condition reflects the actual distribution of mated pairs found in the Ulery et al. (1) data. In the low mated condition, the top cloud was sparse and consisted of 1000 individual points. This manipulation was introduced to determine whether the number of cases in each category affects subjects' decision making processes, which indicates whether subjects are sensitive to the priors when considering the decision tradeoff.

The subjects were instructed to carefully read all of the outcomes in each box and move the sliders to a position where they were comfortable with the outcomes in the boxes. After the subjects were comfortable with the position of the sliders, they clicked the "Save Values" button below the boxes and proceeded to fill out demographic data. The demographic questions are listed in Appendix B. [note that some questions were not asked of all participants]. Participants in the Mechanical Turk phase of data collection also completed two knowledge/attention check questions, to determine their understanding of the experiment. Ultimately, we decided to include all Mechanical Turk subjects regardless of their score on the knowledge check and instead

filtered subjects by reaction time. As this experiment requires critical consideration of one's internal values, we disregarded subjects who took less than 10 seconds to consider the web visualization and did not move the sliders from their starting positions. This filtering process resulted in 9 Mechanical Turk subjects being removed for data analysis. For a complete breakdown of subject inclusion please see Appendix C.

The six outcomes represented in the model will be hereby referred to as: non-mated pairs incorrectly identified (i.e. potentially innocent people incorrectly identified as guilty), mated pairs correctly identified (i.e. potentially guilty people correctly identified as guilty), non-mated pairs correctly excluded (i.e. potentially innocent people correctly identified as innocent), mated pairs incorrectly excluded (i.e. potentially guilty people incorrectly identified as innocent), mated inconclusive and non-mated inconclusive. Looking at the web visualizations (Figure 2), the non-mated pairs incorrectly identified outcome is located to the right of the identification slider in the non-mated distribution (bottom right, red cloud), the mated pairs correctly identified outcome is located to the right of the identification slider in the mated distribution (top right, green cloud), the non-mated pairs correctly excluded outcome is located to the left of the exclusion slider in the non-mated distribution (bottom left, green cloud), the mated pairs incorrectly excluded outcome is located to the right of the exclusion slider in the mated distribution (top left, red cloud), the mated inconclusive outcome is located between the two sliders in the mated distribution (top yellow cloud), and the non-mated inconclusive outcome is located between the two sliders in the non-mated distribution (bottom yellow cloud). In addition to the assumptions underlying our signal detection theory representation of the Ulery et al. (1) data, our methods require one additional assumption that allows us to map the decision of the examiner (along with the base rates of mated and non-mated pairs) to outcomes as shown in the lower boxes of Figure

2. We assume that all comparisons involve fingerprints that are relevant to a criminal case and are potentially inculpatory in nature. There are other uses for fingerprint identification, such as victim remains processing and biometric identification, and under these circumstances an identification decision combined with a mated pair would not contribute to putting a criminal in prison. We believe that our instructions and examples (see Appendix A for transcript) made it clear to participants that we were referring only to impressions that had possible inculpatory or exculpatory implications in criminal proceedings. However, if members of the general public and fingerprint examiners make different assumptions about the utility of a comparison, this has implications for the interpretation of our results. We return to this point later in the discussion.

6. Results

This project has four central questions:

- 1) Do examiners and members of the general public differ in their placement of the identification and exclusion criterion, which may indicate differences in the utilities assigned to the outcomes?
- 2) Do the identification and exclusion criteria measured from examiners differ from the error rates reported by Ulery et al. (1)?
- 3) Are the identification and exclusion criteria sensitive to the base rates of mated and non-mated pairs, as assessed through the two conditions illustrated in the Figure 2 visualizations that vary the number of mated comparisons?
- 4) What aspects of the demographic data predict a participant's placement of the decision criteria and the number of innocents in prison a participant is willing to tolerate?

The answers to the first three questions come from an analysis of the placement of the identification and exclusion criteria between subject groups for the high and low mated conditions. To do this, we compared criteria placement data from the 455 subjects that participated in the experiment using traditional hypotheses testing. For the fourth question, we fit the data to a bivariate tobit model as well as a zero-inflated negative binomial regression model. For both of these models the same 455 subjects considered in the first analysis were considered, but only 369 subjects had complete demographic data. If the subject was missing data from any of the variables used for the tobit and negative binomial model estimation, they were discarded (a “Decline to Answer” option was included on all demographic questions).

6.1 Results of Criteria Placement Comparisons

Table 3 shows the mean criteria placement and standard deviation for the exclusion criterion and identification criterion, respectively, for examiners and the general public in both the high mated and low mated conditions. Additionally, it shows the results of performing Bartlett’s test (22) for population variance. The goal of this test is to determine whether our samples originate from a population with the same variance. We found that there was a significant difference in sample variation of identification criterion placement between the general public and examiners in the high mated condition; $\chi^2(1) = 10.47$, $p = <0.01$ and between the general public and examiners in the low mated condition; $\chi^2(1) = 9.98$, $p = <0.01$. There was not a significant difference in sample variation of exclusion criterion placement between the general public and examiners in the high mated condition; $\chi^2(1) = 3.05$, $p = 0.08$ or between the general public and examiners in the low mated condition; $\chi^2(1) = 3.57$, $p = 0.06$.

Table 3 also shows the results of running a two-tailed, independent-samples t-test assuming unequal variance comparing the average placement of the exclusion and identification criteria for high and low mated groups between the general public and examiners. There was a significant difference in identification criterion placement between the general public and examiners in the high mated condition; $t(265) = 4.71, p < 0.01, d = 0.61$ and between the general public and examiners in the low mated condition; $t(177) = 2.42, p = 0.02, d = 0.34$. There was not a significant difference in exclusion criterion placement between the general public and examiners in the high mated condition; $t(265) = -1.86, p = 0.07, d = -0.39$ or between the general public and examiners in the low mated condition; $t(177) = -0.08, p = 0.93, d = -0.02$.

[Insert Table 3]

To summarize these results, Figure 3 shows the average placement of the identification and exclusion criteria for both examiners and the general public in the high and low mated conditions. The horizontal axis represents the perceived detail in agreement between two fingerprints and the vertical axis represents the high mated and low mated conditions. The values on the horizontal axis are set relative to the standard deviation of the non-mated distribution, which is fixed at 1.0. The criterion placement as predicted by Ulery et al. (1) is represented on Figure 3 as black lines. This analysis shows that examiners tend to be more conservative than the general public when it comes to the placement of the identification criteria in both the high and low mated conditions.

[Insert Figure 3]

To visualize the surprisingly large difference between the two groups, Figure 4 illustrates the average criterion placement for the high mated condition for both groups and demonstrates that the identification criteria for examiners are much further to the right than for the general public.

This result suggests that the general public are less concerned with avoiding erroneous identification errors and may be willing to accept more of these errors in exchange for more potential correctly identified criminals. Additionally, Figure 3 shows a large difference in the location of the exclusion criteria between our study and the Ulery et al. (1) study. In our study, examiners placed their exclusion criteria significantly lower along the horizontal axis. This could suggest that examiners were attempting to minimize the erroneous exclusion error by increasing the inconclusive section of the model.

[Insert Figure 4]

Surprisingly (at least to us), examiners in both conditions place their identification decision criterion on almost exactly the same location as was revealed by the Ulery et al. (1) error-rate study. Examiners believe in, and behave, in almost exactly the same risk-averse manner. The general public, however, are much less risk averse than the threshold revealed by the error-rate study. Note that both experts and the general public differed greatly from the error-rate exclusion threshold. Both groups were much more risk averse with respect to exclusion errors than was revealed by the error-rate data. When it comes to exclusions, it may be easier to make a decision in practice than to consider the abstract possibility of an erroneous exclusion. However, these differences are perhaps less compelling than the identification threshold results, because there may be little practical difference between an exclusion and an inconclusive decision.

6.2 Associations between Demographic Data and Criteria Placement

The resolution of the decision tradeoff involves a set of personal beliefs about the role of justice in society. For example, a ‘law-and-order’ politician may stress the need for locking up

criminals, at the possible risk of incarcerating innocent individuals as well. Organizations such as the Innocence Project may have concerns about the rights of wrongfully-convicted individuals and might argue for a more conservative decision threshold that requires stronger evidence before incarcerating a suspect. What factors might affect whether someone holds a decision threshold that is more conservative (i.e. moving to the right along the axis in Figure 2), or less conservative (i.e. moving to the left in Figure 2)?

To answer this question, we conducted an exploratory analysis that relates the demographic data provided by each participant and their decision criterion placement. We noted that the primary difference between examiners and the general public seems to be the weight that they place on the erroneous identification outcomes, because the base-rate manipulation had no effect on the examiners' criteria and very little effect on the general public's criteria. This suggests that there may be individuals who wish to avoid erroneous identifications at all costs, and those that are willing to consider some as a tradeoff with potentially putting more criminals in prison. In addition, there may be demographic variables that might be related to the choice of identification or exclusion criteria.

We developed two statistical models that address possible associations between our demographic survey data and the decision criterion placement provided by each participant. Both of these approaches are correlational in nature, but still provide meaningful description of the associations we observe between demographics and decision criterion placement.

The first model we estimated is a bivariate tobit model (23) which allows us to explore the influence of demographic factors on the final placement of the decision criteria. The bivariate tobit model is a censored regression that uses demographic data to predict the locations of the

two decision criteria for each participant. This model allows us to assess which demographic variables contribute to variation in the decision criterion placement in a systematic way.

The second model we used is a zero-inflated negative binomial model, which addresses the relation between the demographic variables and the existence of a philosophical belief: Should *any* innocent persons ever be in prison? If so, which factors increase or decrease the acceptable amount of people in prison for each participant? To do this, the model estimates the probability of participants falling into a zero-count state (no innocents should be incarcerated) or a count state (a number of innocents may be incarcerated in an effort to incarcerate more criminals).

6.3 Summary Statistics of Demographic Data

Looking at demographic characteristics of our subject sample, the average age (34.96 years) and average household size (2.97) are close to the national averages. The average subject household income (nearly \$87,000 per year) is higher than the U.S. average. The male/female split of 46.4/53.4 is close to what one might expect in a national sample. 54.2% of the subjects were single, and 50.2% had a bachelor's degree or above (indicating this subject sample is more educated than the U.S. population as a whole).

With regard to fingerprint examination experience, nearly 87% of subjects did not identify themselves as fingerprint examiners (no experience or trainee). Of the 13.1% of subjects who did identify themselves as fingerprint examiners, more than half (58%) had eight years of experience or more. Finally, 84.4% of subjects indicated that they did not have any experience with the justice system, whereas the remaining 15.6% were either fingerprint examiners, police, defense attorneys or judges.

[Insert Figure 5]

Figure 5 shows the percentage of subjects that tolerate non-mated pairs incorrectly identified. The data show that 89 of the 369 subjects used in the tobit and negative binomial models were unwilling to wrongly identify any non-mated pairs (roughly 24 percent of the sample). This includes 24 of the 47 fingerprint examiners (51 percent of fingerprint examiners) and 65 of the 322 non-fingerprint examiners (roughly 20 percent). This preponderance of zeros suggests that a two-state process may be at play in the data. That is, there may be one group of subjects who are unwilling to incorrectly identify any non-mated pairs (state 1) and another group who are willing to incorrectly identify some number of non-mated pairs in their quest to correctly identify as many mated pairs as possible (state 2). We attempt to account for this two-state possibility by fitting our data to the zero-inflated negative binomial model later in this paper.

6.4 Bivariate Tobit Model

Our initial analysis using the t-tests was restricted to grouping subjects by whether they were examiners or the general public and did not take advantage of the additional demographic data that were collected. To include full demographic data in the study, we use a tobit model to consider the ratio of non-mated pairs incorrectly identified to mated pairs correctly identified.

This ratio is representative of the position of the identification criterion, as the non-mated pairs incorrectly identified and mated pairs correctly identified outcomes are located to the right of the criterion. If the number of cases in both inconclusive outcomes are allowed to vary, then the location of the identification criterion is fully accounted for by considering the ratio of non-mated pairs incorrectly identified to mated pairs correctly identified. This ratio will have a minimum of zero (when no non-mated pairs are incorrectly identified) and will be a continuous variable when one or more non-mated pairs are incorrectly identified. With the 369 observations,

this ratio has a mean of 0.1002 (standard deviation of 0.242) with a minimum of 0.0 and a maximum of 2.525. The ratio of mated pairs incorrectly excluded and non-mated pairs correctly excluded, which is representative of the exclusion criterion placement, should also be considered. For this ratio, with the 369 observations, the mean is 0.0389 (standard deviation of 0.0680) with a minimum of 0.0 and a maximum of 0.739.

To statistically model these ratio data, a standard ordinary least squares regression would be inappropriate because the data are censored at zero (application of ordinary least squares regression would result in biased and inconsistent model-parameter estimates). Instead, a censored regression model, such as the tobit model, is appropriate (23). However, because there are two ratios that should be considered simultaneously, a bivariate tobit model is appropriate. This model allows for correlation between the two ratios discussed above because subjects select both their left and right slider positions to define these ratios when participating in the experiment. The bivariate tobit model takes the form (24),

$$\begin{aligned}
 Y_{nk}^* &= \boldsymbol{\beta}_k \mathbf{X}_{nk} + \varepsilon_{nk}, & n=1, 2, \dots, N, \quad k=1, 2 \\
 Y_{nk} &= Y_{nk}^* & \text{if } Y_{nk}^* > 0 \\
 &= 0 & \text{if } Y_{nk}^* \leq 0,
 \end{aligned} \tag{1}$$

where Y_{n1}^* (with $k=1$) is latent variable of the ratio of non-mated pairs incorrectly identified to mated pairs correctly identified for subject n and is observed only when positive, Y_{n2}^* (with $k=2$) is latent variable of the ratio mated pairs incorrectly excluded to non-mated pairs correctly excluded for subject n and is observed only when positive, N is the total number of subjects, Y_{n1} is the observed ratio of non-mated pairs incorrectly identified to mated pairs correctly identified dependent variable, Y_{n2} is the observed ratio of mated pairs incorrectly excluded and non-mated pairs correctly excluded dependent variable, \mathbf{X}_{n1} and \mathbf{X}_{n2} are vector of explanatory variables corresponding to ratio equations, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors of estimable parameters, and ε_{n1} and ε_{n2}

are bivariate normally and independently distributed error terms with zero means, variances $\sigma_{\varepsilon_{n1}}^2$ and $\sigma_{\varepsilon_{n2}}^2$, correlation ρ . This tobit model can be readily estimated by standard maximum likelihood methods (25).

6.4.1 Bivariate Tobit Model Estimation Results

The bivariate tobit model estimation results are presented in Table 4. There is a possibility that unobserved heterogeneity may be present in the data and that this may be affecting the estimation results. To test for this, a random-parameters bivariate tobit model was considered, which allows for the possibility that individual subjects may have their own unique parameter estimates based on a parameter-distributional assumption made by the analyst (25, 26). A wide variety of distributional assumptions were considered for each estimated parameter, but we were unable to find any statistically significant difference with the traditional fixed-parameter approach. Thus, all parameters are conventional fixed parameters meaning all subjects have the same β_1 and β_2 are vectors.

[Insert Table 4]

Turning first to the estimation results for the ratio of non-mated pairs incorrectly identified to mated pairs correctly identified, the fingerprint examiner indicator resulted in a negative parameter indicating a smaller ratio (fewer non-mated pairs incorrectly identified per mated pairs correctly identified). The number of children (if not a fingerprint examiner) variable produced a positive parameter indicating the more children that a household has the higher the ratio (more non-mated pairs incorrectly identified per mated pairs correctly identified). The older male variable (1 subject is a male 65 years old or older, 0 otherwise) produced a positive parameter indicating a higher ratio for such subjects. The white indicator (1 if subject identified

themselves as being white, 0 otherwise) produced a negative parameter indicating a smaller ratio (fewer non-mated pairs incorrectly identified per mated pairs correctly identified). Finally, the high mated pairs indicator (1 if subject faces a cluster of 5969 mated pairs, 0 if 1000 mated pairs) was understandably significant because changing the value of the denominator (number of mated pairs) clearly changes the ratio values, making the ratio smaller as indicated by the negative parameter value. It should be noted that variables that did not produce parameters that were statistically significant from zero were excluded from the estimation of the ratio of non-mated pairs incorrectly identified to mated pairs correctly identified.

Turning to the second tobit equation (the observed ratio of mated pairs incorrectly excluded to non-mated pairs correctly excluded), Table 4 shows that none of the variables included in the model estimation produced parameter estimates that were significantly different from zero, except for the high mated pairs indicator (1 if respond faces a cluster of 5969 mated pairs, 0 if 1000 mated pairs) which again is significant because changing the value of the numerator, in this case, clearly changes the ratio values, making the ratio larger as indicated by the positive parameter value. Thus, for this second ratio, no explanatory variable from socio-demographic data produced a result that was significantly different from zero. Moreover, the correlation between these two ratios produced a t-statistic less than one, indicating that there is no significant correlation between these two ratios (the correlation is not significantly different from zero).

This finding has important implications because it suggests that the ratio of mated pairs incorrectly excluded to non-mated pairs correctly excluded cannot be predicted with our collected socio-demographic information, even though the ratio of non-mated pairs incorrectly identified to mated pairs correctly identified can be predicted (with several variables found to be

statistically significant from zero). From the perspective of this experiment, the bivariate tobit finding suggests the location of the identification criterion position can be predicted based on observable data but the location of the exclusion criterion cannot be predicted.

6.4.2 Summary of Bivariate Tobit Model Results

In summary, the results of fitting the tobit model to our data lead us to the following conclusions. First, there are several demographic factors that influence the placement of the identification criterion. These factors include whether a subject is an examiner, number of children per household, race, age, and gender. If a subject was an examiner or white, the model produced a negative parameter meaning these subjects tend to be more conservative, or less likely to tolerate non-mated pairs incorrectly identified. If a subject was an older male (over 65 years) they tended to be less conservative, or more likely to tolerate non-mated pairs incorrectly identified. Finally, as more children are added to a household, subjects become less conservative, thus tolerating more non-mated pairs incorrectly identified. The second conclusion is that no demographic factors were able to predict the placement of the exclusion criterion. That is, no demographic factors produced a significant parameter when attempting to predict the location of the exclusion criterion. Additionally, we found that the placement of the identification criterion and the exclusion criterion were not correlated. The results of fitting this model support our earlier suspicion that subjects are generally more concerned with the outcomes dictated by the identification criterion, specifically the erroneous identification error which has the potential to incriminate innocent people. To explore this issue further, the count data of individual pairs of prints (represented by the points in Figure 2) was fit to a zero-inflated negative binomial model

which focuses solely on the number of non-mated pairs incorrectly identified.

6.5 Zero-Inflated Negative Binomial Model

The goal of the zero-inflated negative binomial analysis is to identify subject characteristics that are statistically significant determinants of the number of non-mated pairs that individual subjects are willing to incorrectly identify. In each of the high and low mated conditions of the experiment, there are a total of 4000 non-mated pairs under consideration, and the number of non-mated pairs incorrectly identified as mated will be a non-negative integer ranging from 0 to 4000.

This makes the data well-suited to analysis by traditional count-data regression methods such as Poisson (see for example Hostetter (27), for an application to gesture analysis) and negative binomial regressions. For the Poisson regression model, the probability $P(i_n)$ of subject n incorrectly identifying i_n non-mated pairs,

$$P(i_n) = \frac{EXP(-\lambda_n)\lambda_n^{i_n}}{i_n!} \quad (2)$$

where λ_n is the Poisson parameter for subject n , which is the subjects expected number of erroneous identification errors. To incorporate explanatory variables, Poisson regression specifies the Poisson parameter λ_n as a log-linear function, $\lambda_n = EXP(\beta\mathbf{X}_n)$, where \mathbf{X}_n is a vector of explanatory variables that determine the number of non-mated pairs incorrectly identified by subject n and β is a vector of estimable parameters (25).

Depending on the nature of the data being modeled, the Poisson regression may not always be appropriate because the Poisson distribution restricts the mean and variance to be equal ($E[i_n] = VAR[i_n]$). In many datasets it is common for the variance of the counts to be much

greater than the means ($VAR[i_n] \gg E[i_n]$) in which case the data are considered to be overdispersed. To account for this possibility, the negative binomial regression model is derived by rewriting the expected number of non-mated pairs incorrectly identified as, $\lambda_n = EXP(\beta X_n + \varepsilon_n)$, where $EXP(\varepsilon_n)$ is a gamma-distributed error term with mean 1 and variance α . The addition of this gamma-distributed results in the negative binomial regression (also sometimes referred to as the Poisson-Gamma regression since a gamma function is added to account for overdispersion) and allows the variance to differ from the mean as $VAR[i_n] = E[i_n][1 + \alpha E[i_n]] = E[i_n] + \alpha E[i_n]^2$. The addition of this gamma term results in a negative binomial regression (also sometimes referred to as the Poisson-Gamma regression since a gamma function is added to account for overdispersion). The Poisson regression is a limiting model of this negative binomial regression as α approaches zero. Thus, if α is determined during estimation to be significantly different from zero, the negative binomial is appropriate and, if it is not, the Poisson model is appropriate because with α equal zero the negative binomial reduces to the Poisson model. Both Poisson and negative binomial models can be readily estimated with standard maximum likelihood methods (25).

However, for the case of non-mated pairs incorrectly identified, a count-data regression approach should also be considered that allows for the possibility of two count states; a zero-count state, and a count state. The idea is that a substantial portion of subjects may have a strict belief that non-mated pairs should never be incorrectly identified, which would put these subjects in a zero-count state. Other subjects may be willing to incorrectly identify non-mated pairs (to correctly identify more mated pairs), and these subjects would be in a count state (which would include non-negative integers as in a traditional count-data model). Because it is not known

which subjects are in the zero-count and count states, a model-estimation process that incorporates the state-splitting estimation must be considered.

Two popular models that account for this two-state possibility are the zero-inflated Poisson regression, which has a Poisson count-state function, and the zero-inflated negative binomial regression, which has a negative binomial count-state function (25, 28, 29). Details of this model are presented in the Appendix D.

Finally, to assess the effects of individual explanatory variables in the \mathbf{X}_n vector, on the mean number of non-mated pairs (λ_n) incorrectly identified by subject n , a marginal effect, which gives the effect that a one-unit change in subject n 's explanatory variable x_{nk} (one element of the \mathbf{X}_n vector) has the mean number of non-mated pairs incorrectly identified and is computed as,

$$ME_{x_{nk}}^{\lambda_n} = \frac{\partial \lambda_n}{\partial x_{nk}} = \beta_k EXP(\beta \mathbf{X}_n) \quad (3)$$

where k is the k^{th} element of the \mathbf{X}_n vector. Because each subject has their own marginal effect, the average marginal effect over the over the subject population N for each explanatory variable found to be statistically significant will be reported. Also, in two-state zero-inflated models, where the same variable may be in both the zero-state splitting function and the count state, the marginal effect will capture the total net effect.

6.5.1 Zero-Inflated Negative Binomial Model Estimation Results

Table 5 presents the model estimation results with corresponding marginal effects (the average effect that a one unit change in the explanatory variable will have on the number of non-mated pairs incorrectly identified). The value of the Vuong statistic (see Appendix D for description) shown in Table 5 is 3.30, which indicates more than 99% certainty that the two-state

zero-inflated model is preferred relative to a single state model. The negative binomial dispersion parameter of 2.373 with a t-statistic of 27.64 is significantly different from zero suggesting that the negative binomial is statistically preferred over the Poisson. Thus, these statistics strongly support the zero-inflated negative binomial model relative to the non-zero-inflated Poisson and negative binomial models, and the zero-inflated Poisson model. Also, as further evidence in addition to the Vuong test, a simple likelihood ratio test comparing the log-likelihood at convergence for the negative binomial (-1944.97) with the log-likelihood at convergence for the zero-inflated negative binomial (-1922.82) gives a χ^2 statistic of 44.30 $[-2(-1944.97-(-1922.82))]$ with 4 degrees of freedom, which implies more than 99.99% confidence that the simple negative binomial model and the zero-inflated negative binomial model are not equal.

[Insert Table 5]

Before turning to the specific parameter estimation results, it is important to mention other aspects of the model that were considered. First, as mentioned in the experimental design, the number of mated pairs presented to subjects was 1000 for some and 5969 for others. This suggests the possibility that subjects' "innocent" decisions may be influenced by the number of mated pairs presented in their experiment. To test for this, the sample was split in two; those subjects facing 1000 mated pairs and those facing 5969 mated pairs. Two separate models were estimated for each of these two sub-populations. The test statistic is $X^2 = -2[LL(\beta_{all}) - LL(\beta_{1000}) - LL(\beta_{5969})]$, where $LL(\beta_{all})$ is the log-likelihood at convergence of the model estimated with all data (as shown in Table 5), $LL(\beta_{1000})$ is the log-likelihood at convergence of the model using only data from subjects facing 1000 mated pairs, and $LL(\beta_{5969})$ is the log-likelihood at convergence of the model using only data from subjects facing 5969 mated pairs. In this test the same variables are used in all three models and this X^2 test statistic is χ^2 distributed with degrees

of freedom equal to the summation of the number of estimated parameters in the “1000” and “5969” models minus the number of estimated parameters in the “all” model. The test statistic indicates that there is only 18% confidence that the separate “1000” and “5969” models are statistically different from the “all” model, so this justifies the use of a single model for all subjects. Second, we estimated a random-parameters zero-inflated negative binomial, which allows for the possibility of individual subjects having unique parameter estimates. As was the case for the bivariate tobit model, we were unable to find any statistically significant difference with the traditional fixed-parameter approach shown in Table 5. Thus, the traditional assumption that there is one effect for explanatory variables across all subjects is statistically valid and unobserved heterogeneity does not seem to be playing a role in the model estimation results.

Turning first to the estimation results in the zero-state splitting function in Table 5 (where a positive parameter estimate increases the likelihood a subject will be in the zero state and a negative parameter estimate decreases the likelihood), the fingerprint-examiner indicator variable produced a positive parameter indicating that subjects identified as fingerprint examiners were more likely to be in the zero state (inherently unwilling to incorrectly identify any non-mated pairs) relative to non-fingerprint examiners. There could be a number of reasons for this. First, the hypothetical nature of our experiment may make non-fingerprint examiners less aware of the consequences of incorrectly identifying non-mated pairs. Second, the potential training of fingerprint examiners may contribute to their likelihood of being in the zero-state. Third, fingerprint examiners may be a self-selected group of individuals that are inherently less likely to believe that incorrectly identifying non-mated pairs is a tolerable option. These, and possibly other factors, could be playing a role.

Table 5 shows that subjects less than 25 years old were less likely to be in the zero state and thus more likely to be in the count state. In fact, the marginal effects shown in Table 5 suggest that the net effect is that individuals less than 25 years old were, on average, willing to incorrectly identify 37.90 more non-mated pairs than other age groups.

Finally, for the zero-state probabilities, Table 5 shows that, for those subjects who were not fingerprint examiners, the more children there were in the subject's household, the less likely the subject was to be in the zero state, and they were thus more likely to be in the count state (and thus more likely to incorrectly identify non-mated pairs).

Turning to variables found to be statistically significant in the count state (where a positive parameter estimate increases the number of non-mated pairs a subject is willing to incorrectly identify, and a negative parameter decreases this number), subjects with household incomes less than \$50,000 dollars per year were more willing to incorrectly identify non-mated pairs. Marginal effects in Table 5 show that subjects in this lower-income bracket were, on average, willing to incorrectly identify 82.14 more non-mated pairs than their higher-income counterparts.

In addition to lower income subjects, Table 5 shows males 65 years old or older were also more willing to incorrectly identify non-mated pairs, with marginal effects showing on average willing to incorrectly identify 191.55 more non-mated pairs than younger males and all females. The combination of lower incomes and older males seems to capture a demographic that is particularly hard on crime and more willing to incorrectly identify non-mated pairs in their quest for correctly identifying as many mated pairs as possible.

Interestingly, subjects who identified themselves as white (77.8% of the sample), were significantly less likely to incorrectly identify non-mated pairs relative to other ethnicities in the

sample (marginal effects in Table 5 show they were, on average, willing to incorrectly identify 126.93 fewer non-mated pairs).

Finally, having more children not only made subjects less likely to be in the zero-count state, but also made them significantly more likely to incorrectly identify non-mated pairs when in the count state. Marginal effects show the total effect of zero-state and count-state estimations is that for each additional child a subject has, the subject is willing to incorrectly identify on average 68.02 more non-mated pairs.

Please note that the fingerprint-examiner indicator variable, though significant in determining the probability that subject will be in the zero state, was not significant in determining the number of non-mated pairs the subject was willing to incorrectly identify. Still, the net effect of the zero and count-state marginal effects show (Table 5) that fingerprint examiners are willing to incorrectly identify 33.10 fewer non-mated pairs than their non-fingerprint-examiner counterparts with the same characteristics is in the zero state (note that this is so because the marginal effect equation is a partial derivative that holds other characteristics fixed).

6.5.2 Summary of Zero-Inflated Negative Binomial Model Results

Fitting the zero-inflated negative binomial model to our data provides us with several interesting results. The model estimation includes the determination of which demographic factors affect the likelihood that a subject will be in the zero-count state or the count state. Then, for the count state, we determine which demographic factors affect how many incorrectly-identified non-mated pairs the subject is willing to tolerate. The factors found to be significant for predicting whether a subject falls into the zero-count or count state are: whether they are an

examiner, age, and children. If subjects are examiners or if they are less than 25 years old, then they are much more likely to be in the zero-count state. However, the more children a household has the less likely subjects are to be in the zero-count state. Moving on to the count state, the factors that affect the number of non-mated pairs incorrectly identified are: income, age, gender, race, and number of children. Subjects with yearly income less than \$50,000, older (greater than 65) males, and households with more children were likely to tolerate more non-mated pairs incorrectly identified. On the other hand, subjects who identified as white were likely to tolerate fewer non-mated pairs incorrectly identified.

The result we find particularly interesting is the fact that fingerprint examiners relative to the general public seem to be more likely to be in the zero-state—meaning, they are philosophically opposed to putting innocents in prison (by incorrectly identifying non-mated pairs). There could be a number of reasons for this as discussed in Table 2. First, the hypothetical nature of our experiment may make the general public less aware of the consequences of incorrectly identifying non-mated pairs. Second, the potential training of fingerprint examiners may contribute to their likelihood of being in the zero-state. Third, fingerprint examiners may be a self-selected group of individuals that are inherently less likely to believe that innocents in prison is a tolerable option. These, and potentially other factors, could be playing a role.

7. Discussion

The results of this study support the idea that the general public are 1) less biased against erroneous identification errors than examiners and 2) are less tolerant of a large number of inconclusive decisions. According to the results shown in Table 3, there is a significant difference between the general public and examiner placement of the identification criterion. As

shown in Table 3, the mean placement of the identification criterion by the general public is lower than the examiners for both the high mated and low mated conditions. This suggests that the general public subjects are more willing to accept an erroneous identification error in exchange for fewer inconclusive decisions. Note that the manipulation of the number of mated pairs has separate participants, and therefore represents a conceptual replication of the result that examiners are more conservative than members of the general public.

This finding is also supported by the results of the negative binomial model showing that fingerprint examiners are more likely to be in the zero state than the general public. This means that fingerprint examiners are more likely to be philosophically opposed to ever convicting an innocent person, at the cost of less criminal convictions, while non-fingerprint examiners may have a threshold higher than zero.

In the low mated condition, the size of the mated distribution was decreased to a value approximately a fifth of the high mated condition. Despite this fairly dramatic change, the results show that there is no significant difference between criterion placement in either fingerprint examiners or the general public as a result of this manipulation. This suggests that the subjects are paying less attention to the overall ratio of mated to non-mated pairs in each outcome and instead are paying attention to a specific number of outcomes, namely the number of potential innocent people in prison. This idea is supported by the lack of change between the low and high mated conditions and the results of the tobit and negative binomial models which suggest more indicators are available to predict the number of potential innocent people in prison than any other outcome.

It is possible that the difference in values between examiners and the general public is caused by a gap in knowledge as described in Table 2. This study asks for the values people

place on the outcomes of fingerprint examinations. Naturally, fingerprint examiners are more familiar with these outcomes and have had more time than the course of this study to form an opinion. Fingerprint examiners may also have stronger opinions than the general public because the outcomes of these decisions directly affect their livelihood where as a person from the general public would most likely be removed from the consequences. While all of our participants were informed about the possible outcomes and ramifications of these decisions, the outcomes are more nuanced than can be completely explained in an instructional video. For example, the relative proportion of mated to non-mated pairs is not actually known and can be defined in multiple ways. We used the same proportions that Ulery et al. (1) used in their study, but it is not certain that these proportions reflect reality. Additionally, not all identification decisions necessarily lead to convictions. This information was included in the instructional video and in the boxes explaining the outcomes on the web-based visualization. However, it is possible that the examiners and the general public may have interpreted this information differently. The instructions in the video focus on criminal cases, but examiners may have been considering other types of cases as a result of their personal experience (i.e. victim identification which places victims at a crime scene). If this were the case, we would expect examiners to have a less conservative identification decision criterion (farther to the left) in order to compensate for possible non-criminal cases. However, we did not find this to be true. As the examiners were significantly more risk averse in their placement of the identification criterion than the general public, we suspect that both subject groups were appropriately considering the outcomes of criminal cases as intended.

Our finding that examiners were more risk averse than the members of the general public rules out many of the possible differences between groups as posited in Table 2. In addition, years

of experience was not significant, although it is correlated with age, and that multi-collinearity makes it difficult to tease apart age and years of experience. Thus, it is possible that older examiners are more risk averse.

The difference in identification criterion placement between the two groups and the results of the negative binomial model may indicate that the decisions of examiners in latent print examinations do not accurately reflect the values of society. Examiners face potentially severe personal and professional costs to an erroneous identification, even with the safeguards built into the system such as verification and the fact that fingerprint evidence may not be the only factor in a case. The results of this study demonstrate that members of the general public may be more comfortable increasing the number of erroneous identification decisions in exchange for more correct identification of criminals. This could provide the feedback to examiners that they are too risk averse.

7.1 Implications for Fingerprint Comparisons

We return to the fundamental question motivating this research: In a democratic society, who decides how much evidence is sufficient to draw a forensic conclusion? In the US, judges act as gatekeepers for who is allowed to testify (30). However, the interpretation of that testimony is left to jurors (or defendants, in the 95% of cases that get plead out), and the placement of the decision criteria affects the outcomes, with the potential to scare a defendant into taking a plea deal. One goal of the present work is to identify differences between forensic examiners and members of the general public with respect to the desirable error rate (as inferred from the location of the decision criteria), and the results fairly clearly demonstrate that the general public is willing to accept a higher erroneous identification error rate as a trade-off for

additional criminal identifications. Whether we should trust the general public or forensic examiners (who may face personal penalties for erroneous identification errors) is a policy question rather than a scientific question, and therefore perhaps outside the domain of the current work. *However, the present work demonstrates a fairly significant disconnect between the two groups.*

One practical implication of the present work is the formulation of the forensic decision-making process as a decision tradeoff. This allows examiners to communicate several key points to managers. First, no forensic decision-making process can be free of errors. Therefore, we argue that one response is to place additional safeguards in place but to treat errors as learning opportunities rather than responding punitively. Second, we hope that this research will prompt discussions among forensic practitioners and stakeholders about where their decision thresholds come from, and whether they are appropriate. Given that we are unlikely to directly get guidance from policymakers, internal discussions among leaders in the forensic community are likely to provide the best path forward for setting appropriate decision thresholds. We hope that such discussions will arise from this and related research.

These issues are complex, and one reason that examiners are placed in this position is that they actually make *decisions*. They are essentially taking the strength of evidence in a particular comparison and distilling that down into one of three conclusions. They are implicitly including the prior probability of a mated pair into this calculation, despite the fact that this is typically unknown. An alternative would be to shift to a strength-of-evidence report rather than a definitive conclusion, as is done in Europe. A middle ground would be to expand the set of conclusions that examiners use, from three to five, to allow for investigative leads. Some agencies such as the Houston Forensic Science Center already allow for this for some

comparisons. One caveat with this approach is that while investigative leads may end up being helpful, additional categories may also make examiners even more risk averse (31). Each of these policy changes should be made with the support of empirical studies that demonstrate how best to communicate the strength of evidence in a particular case to an eventual consumer such as a detective, prosecutor, or defense attorney.

References

1. Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. Accuracy and reliability of forensic latent fingerprint decisions. *Proc Natl Acad Sci USA*. 2011 May 10;108(19):7733-8.
doi:10.1073/Pnas.1018707108
2. Maimonides M. *Neg. Comm. 290. The Commandments*: Soncino Press; 1967;269-71.
3. Blackstone W. *Commentaries on the Laws of England*, 1893; 358.
4. Standards for Examining Friction Ridge Impressions and Resulting Conclusions, Version 2.0 [database on the Internet]. 2013 [cited. Available from:
http://www.swgfast.org/documents/examinations-conclusions/130427_Examinations-Conclusions_2.0.pdf.
5. Dror I. The error in “error rate”: why error rates are so needed, yet so elusive. *J Forensic Sci*. 2020;65(4):1034-9. doi:10.2139/ssrn.3593309
6. Christensen AM, Crowder CM, Ousley SD, Houck MM. Error and its meaning in forensic science. *J Forensic Sci*. 2014;59(1):123-6. doi:10.1111/1556-4029.12275
7. Charlton D, Fraser-Mackenzie PA, Dror IE. Emotional experiences and motivating factors associated with fingerprint analysis. *J Forensic Sci*. 2010;55(2):385-93.
doi:10.1111/j.1556-4029.2009.01295.x
8. Ashbaugh D. *Quantitative-qualitative friction ridge analysis: an introduction to basic and advanced ridgeology*. Boca Raton, Fla: CRC Press, 1999; 108-148.

9. Scientific Working Group on Friction Ridge Analysis SaT, and United States of America. Fingerprint Sourcebook, 2011; 53.
10. Analysis. EWGoHFiLP. Latent print examination and human factors: improving the practice through a systems approach: the report of the Expert Working Group on Human Factors in Latent Print Analysis. . In: NIST NIJ NIOJ, editor. Washington, D.C.; 2012.
11. Snodgrass M, Bernat E, Shevrin H. Unconscious perception at the objective detection threshold exists. *Percept Psychophys*. 2004;66(5):888-95. doi:10.3758/bf03194982
12. Vanselst M, Merikle P. Perception below the objective threshold? *Consious Cogn*. 1993;2(3):194-203. doi:10.1006/ccog.1993.1018
13. Dror I. A hierarchy of expert performance. *Journal of Applied Research in Memory and Cognition*. 2016;5(2):121-7. doi:10.1016/j.jarmac.2016.03.001
14. Tangen JM, Thompson MB, McCarthy DJ. Identifying fingerprint expertise. *Psychol Sci*. 2011;22(8):995-7. doi:10.1177/0956797611414729
15. Macmillan NA, Creelman CD. *Detection theory : a user's guide*. 2nd ed. Mahwah, N.J.: Lawrence Erlbaum Associates, 2005; 24-42.
16. Thompson MB, Tangen JM. The nature of expertise in fingerprint matching: Experts can do a lot with a little. *Plos One*. 2014;9(12):e114759. doi:10.1371/journal.pone.0114759
17. Busey TA, Vanderkolk JR. Behavioral and electrophysiological evidence for configural processing in fingerprint experts. *Vision Res*. 2005;45(4):431-48. doi:10.1016/j.visres.2004.08.021
18. Vogelsang MD, Palmeri TJ, Busey TA. Holistic processing of fingerprints by expert forensic examiners. *Cognitive research: principles and implications*. 2017;2(1):1-12. doi:10.1186/s41235-017-0051-x

19. Dror I. The paradox of human expertise: Why experts get it wrong. In: Kapur N, Pascual-Leone A, Ramachandran V, Sergio Della Sala JC, Manly T, Mayes A, editors. *The Paradoxical Brain*: Cambridge University Press; 2011;177-88.
20. Fisher M, Keil F. The curse of expertise: when more knowledge leads to miscalibrated explanatory insight. *Cognitive Science*. 2015;40(5):1251-69. doi:10.1111/cogs.12280
21. Hinds PJ. The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance. *J Exp Psychol Appl*. 1999;5(2):205. doi:10.1037/1076-898x.5.2.205
22. Bartlett MS. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London Series A-Mathematical and Physical Sciences*. 1937;160(901):268-82. doi:10.1098/rspa.1937.0109
23. Tobin J. Estimation of Relationships for Limited Dependent-Variables. *Econometrica*. 1958;26(1):24-36. doi:10.2307/1907382
24. Anastasopoulos PC, Shankar VN, Haddock JE, Mannering FL. A multivariate tobit analysis of highway accident-injury-severity rates. *Accident Analysis and Prevention*. 2012 Mar;45:110-9. doi:10.1016/j.aap.2011.11.006
25. Washington S, Karlaftis M, Mannering FL, Anastasopoulos PC. *Statistical and econometric methods for transportation data analysis*. Third Edition ed. New York, NY: CRC Press, 2020; 1-477.
26. Mannering FL, Shankar V, Bhat CR. Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research*. 2016 Sep;11:1-16. doi:10.1016/j.amar.2016.04.001
27. Hostetter AB. Action attenuates the effect of visibility on gesture rates. *Cognitive*

Science. 2014;38(7):1468-81. doi:10.1111/cogs.12113

28. Lambert D. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*. 1992 Feb;34(1):1-14. doi:10.2307/1269547

29. Malyshkina NV, Mannering FL. Zero-state Markov switching count-data models: An empirical assessment. *Accident Analysis and Prevention*. 2010 Jan;42(1):122-30.

doi:10.1016/j.aap.2009.07.012

30. *Daubert v. Merrel Dow Pharmaceuticals, Inc.* U.S.; 1993;579.

31. Carter KE, Vogelsang MD, Busey TA. The utility of expanded conclusion scales during latent print examinations. *J Forensic Sci*. 2020;65(4):1141-54. doi:10.1111/1556-4029.14298

Table 1. Response proportions found in Ulery et al. (1) data. Raw frequencies are in parentheses.

Pair Type	Exclusion	Inconclusive	Identification	Total Mates or Non-mates
Mates	0.075 (450)	0.311 (1856)	0.614 (3663)	1 (5969)
Non-mates	0.887 (3622)	0.111 (455)	0.001 (6)	1 (4083)

Table 2. Candidate differences between examiners and members of the general public, along with the likely shift in risk aversion. Examiners who are more risk averse will have fewer erroneous identification outcomes (as well as fewer correct identification outcomes).

Candidate Difference between Examiners and Members of the General Public	Likely outcome (risk averse implies requiring more evidence in order to make an identification decision than less risk averse individuals)		
	Examiners Less Risk Averse	Unclear/no effect	Examiners More Risk Averse
Examiners understand that fingerprint evidence is not typically the only evidence in a criminal proceeding	Errors have less consequence than assumed by the lay public.		
Examiners know that most comparisons are verified by a second examiner	Errors have less consequence than assumed by the lay public.		
Examiners know that not all fingerprint evidence leads to convictions, and that other uses such as victim elimination or non-probative matches are possible	Errors have less consequence than assumed by the lay public.		
Examiners may differ in important ways demographically that may make them favor a more punitive approach to law enforcement	Examiners may be attracted to law enforcement through an emphasis on solving crimes.		
Examiners may have special perceptual abilities not shared by the general public (e.g. configural processing, better contrast sensitivity for ridge details)	Examiners may have <i>expertise blind spots</i> or tunnel vision that restricts information flow and may hide critical information that could lead to overconfidence.		
Experts may have a greater understanding of the comparison process.		Experts may understand fingerprint comparisons more than our general public participants despite our educational materials. How this difference affects the decision thresholds is uncertain, although demographic details such as contact with the criminal justice system may bear on this question.	
Members of the general public may believe fingerprints to be a solved problem and therefore errors are unlikely to occur.		Our visualization makes it clear that errors will occur for most threshold settings.	

<p>Examiners and the general public may have different scientific or statistical backgrounds</p>		<p>Depends on demographics. Many older examiners may not have a scientific background having come from policing. Younger examiners are more likely to have a hard science or social science bachelor's degree. Demographic analysis of levels of education will clarify any differences and suggest a likely outcome.</p>	
<p>Experts have a better grasp of our paradigm and can better connect our visualization to the outcomes</p>		<p>This may just introduce noise into the general public data without affecting the central tendency. Tests on the variance of both groups may address this. We also screened out participants based on reaction time.</p>	
<p>Experts may be more motivated to think deeply about the tradeoffs</p>		<p>This may just introduce noise into the general public data without affecting the central tendency. Tests on the variance of both groups may address this. Some of the general public were tested in-person to ensure participants took the task seriously.</p>	
<p>Experts may encounter close non-mated prints that make them realize the risks involved.</p>			<p>This may be a function of years of experience. More experience allows for more time to encounter 'close calls' (non-mated prints that are very similar in appearance). This may also be a function of the size of the database searched: Modern databases are much larger than previous generations and will reveal more close non-matches.</p>
<p>Examiners may face severe personal and professional consequences for an erroneous identification, but not for an erroneous exclusion</p>			<p>Erroneous identifications may result in corrective actions or even loss of employment.</p>

Table 3. Results of Bartlett's population variance test and t-test comparison of exclusion criteria and identification criteria placement.

	Exclusion Criteria		Identification Criteria	
	High Mated	Low Mated	High Mated	Low Mated
Examiner				
Mean Placement	-0.44	0.21	2.95	2.97
Standard Deviation	1.31	1.20	0.81	0.85
Sample Size	37	41	37	41
General Public				
Mean Placement	-0.01	0.21	2.19	2.51
Standard Deviation	1.07	0.96	1.29	1.33
Sample Size	230	138	230	138
Tests for Population Variance				
χ^2	3.05	3.56	10.47	9.98
dF	1	1	1	1
p	0.08	0.06	<0.01	<0.01
Differences in Placement				
t	-1.86	-0.08	4.71	2.42
dF	1	1	1	1
p	0.07	0.93	<0.01	0.02
d	-0.39	-0.02	0.61	0.34

Table 4. Bivariate tobit model of ratio of non-mated pairs incorrectly identified to mated pairs correctly identified and of the mated pairs incorrectly excluded to non-mated pairs correctly excluded.

Variable Description	Estimated Parameter	t statistic^a
<i>Ratio non-mated pairs incorrectly identified to mated pairs correctly identified</i>		
Constant	0.218	5.12***
Fingerprint examiner indicator (1 if respondent is a fingerprint examiner, 0 otherwise)	-0.160	-3.62***
Number of children in respondent's household if not a fingerprint examiner	0.029	1.86*
Older male indicator (1 respondent is a male 65 years old or older, 0 otherwise)	0.217	4.15***
White indicator (1 if respondent identified themselves as being white, 0 otherwise)	-0.114	-2.79***
High mated pairs indicator (1 if respondent faces a cluster of 5969 mated pairs, 0 if 1000 mated pairs)	-0.165	-3.42***
<i>Ratio of mated pairs incorrectly excluded to non-mated pairs correctly excluded</i>		
Constant	0.00051	0.02
Fingerprint examiner indicator (1 if respondent is a fingerprint examiner, 0 otherwise)	0.00364	0.35
Number of children in respondent's household if not a fingerprint examiner	0.00187	0.37
Older male indicator (1 respondent is a male 65 years old or older, 0 otherwise)	0.01247	0.47
White indicator (1 if respondent identified themselves as being white, 0 otherwise)	0.00615	0.49
High mated pairs indicator (1 if respondent faces a cluster of 5969 mated pairs, 0 if 1000 mated pairs)	0.05013	2.80***
Correlation coefficient between equations,	-0.121	-0.87
Number of observations	369	

^a Confidence level (two-tailed test): * greater than 90%; ** greater than 95%; ***greater than 99%

Table 5. Zero-inflated negative binomial of the number of non-mated pairs incorrectly identified.

Variable Description	Estimated Parameter	<i>t</i> statistic^a	Marginal Effect
<i>Zero-State Splitting Function</i>			
Constant	-0.787	-7.35***	
Fingerprint examiner indicator (1 if respondent is a fingerprint examiner, 0 otherwise)	0.723	6.67***	-33.10
Young-age indicator (1 if respondent is less than 25 years old, 0 otherwise)	-0.828	-6.62***	37.90
Number of children in respondent's household if not a fingerprint examiner	-0.139	-2.02**	68.02
<i>Count State (Number of Non-mated Pairs Incorrectly Identified)</i>			
Constant	5.315	18.52***	
Lower income indicator (1 respondent's household income is less than \$50,000 per year, 0 otherwise)	0.430	1.84*	82.14
Older male indicator (1 respondent is a male 65 years old or older, 0 otherwise)	1.003	1.68*	191.55
White indicator (1 if respondent identified themselves as being white, 0 otherwise)	-0.665	-2.48**	-126.93
Number of children in respondent's household if not a fingerprint examiner	0.323	3.46***	68.02
<i>Negative binomial dispersion parameter</i>	2.373	27.64***	
Number of observations		369	
Log-likelihood at convergence (negative binomial)		-1944.97	
Log-likelihood at convergence (zero-inflated negative binomial)		-1922.82	
Vuong statistic for testing zero-inflated negative binomial versus the standard negative binomial model		3.30	

^a Confidence level (two-tailed test): * greater than 90%; ** greater than 95%; ***greater than 99%

Figure Legends

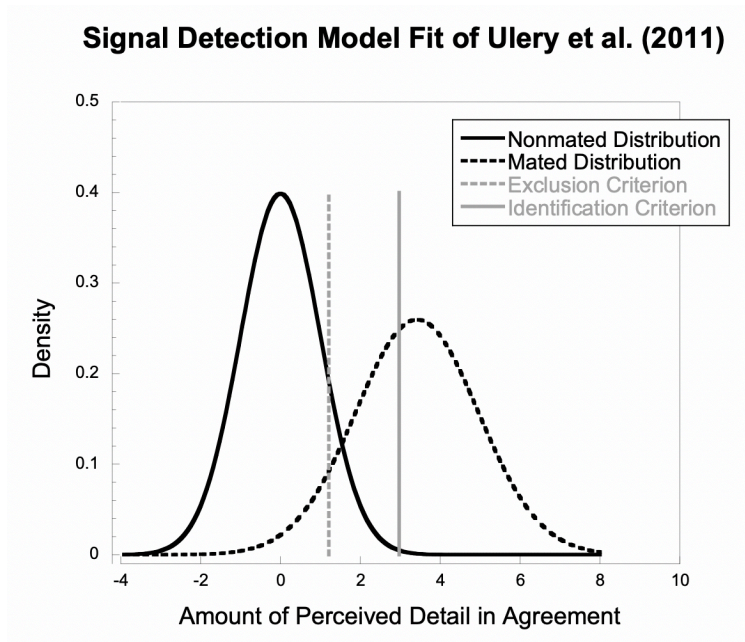


FIG. 1—Representation of the Ulery et al. (1) data using signal detection theory.

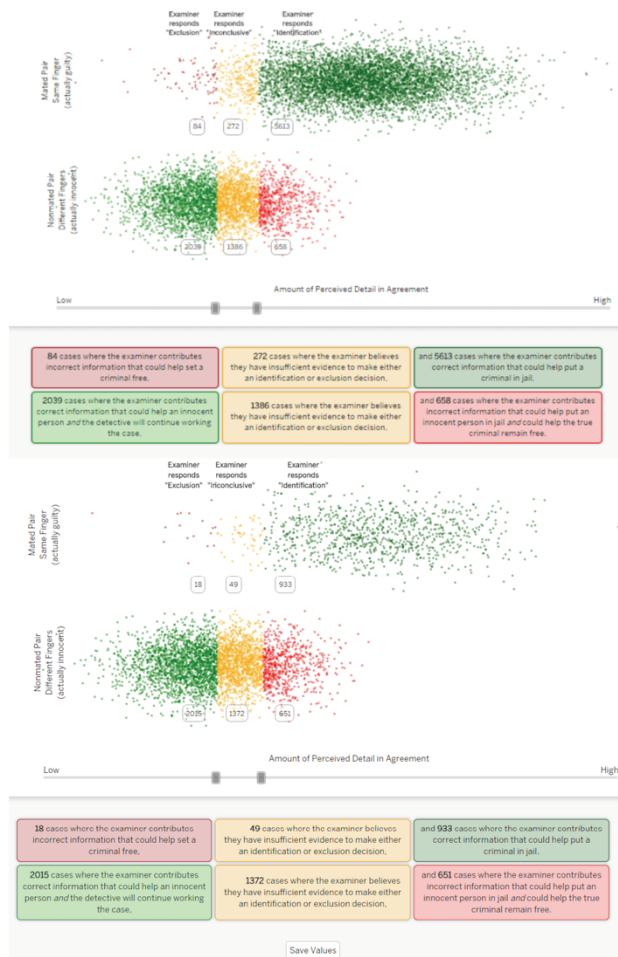


FIG. 2—Web-based Visualization. The horizontal axis is an evidence axis that characterizes the final value at the end of a comparison. Each comparison is represented by a single point, with the upper cloud corresponding to mated pairs, and the lower cloud corresponding to non-mated pairs (the points are jittered vertically for visibility). The two sliders correspond to the exclusion (left) and identification (right) decision criterion). The colors correspond to the outcomes in the table below, which change dynamically in real time as the participant drags either slider. This accurately captures the nature of the changing outcomes as either decision criteria is moved, and reveals that the frequencies change in complex ways that are difficult to intuit but easy to visualize. The top visualization illustrates the condition where the number of mated pairs is high (High

Mated condition) and the bottom visualization illustrates the condition where the number of mated pairs has been reduced to approximately 1/5th of the number (Low Mated condition).

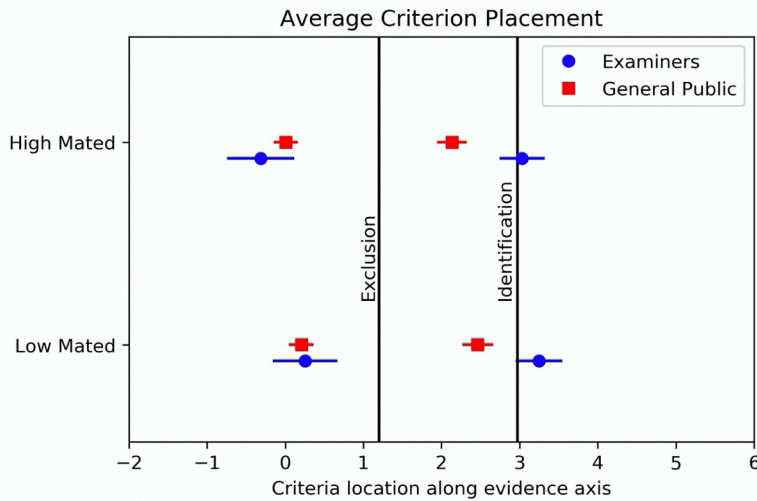


FIG. 3—Average Criterion Placement. The horizontal axis corresponds to the location along the horizontal axis in Figure 2 expressed in units of the standard deviation of the non-mated distribution. High mated and low mated indicate the values associated with the two base-rate conditions as illustrated in Figure 2. Error bars in this figure and following figures are a 95% confidence interval based on the standard error of the mean multiplied by 1.96. The vertical black lines represent the location of the identification and exclusion decision criteria of examiners who participated in the Ulery et al. (1) study.

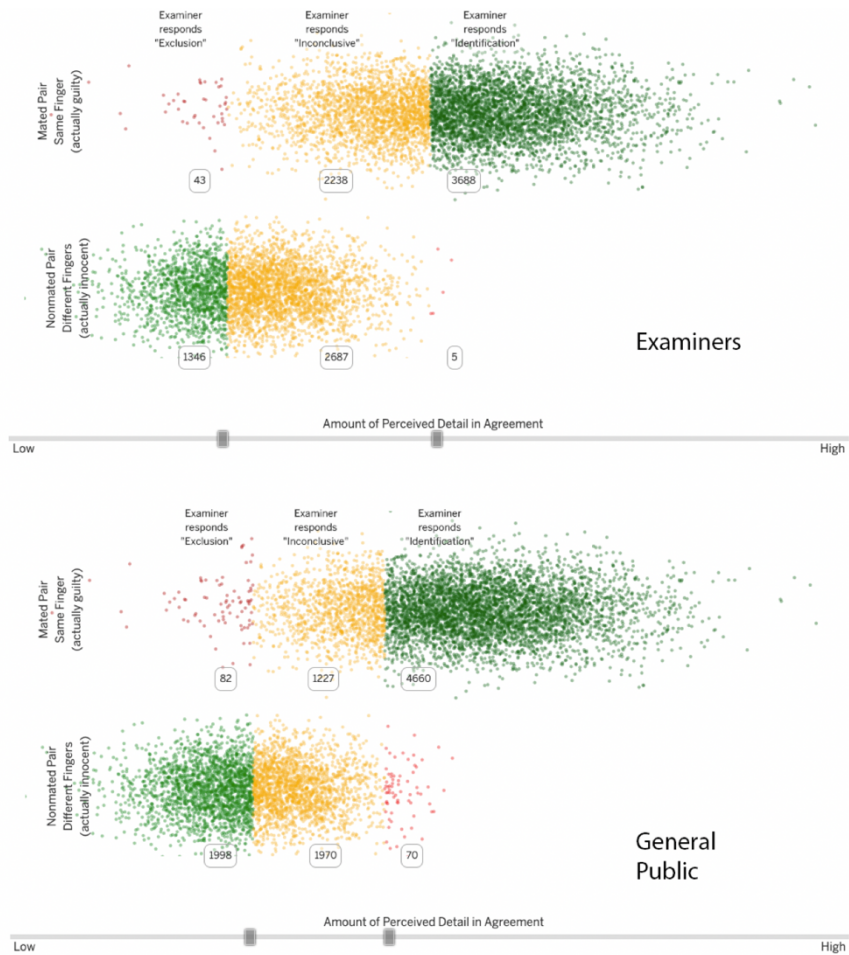


FIG. 4—Visualization of the differences of criteria placement by examiners and the general public for the high mated condition based on the data from Table 3. Note the large difference between the two groups in the identification criterion placement.

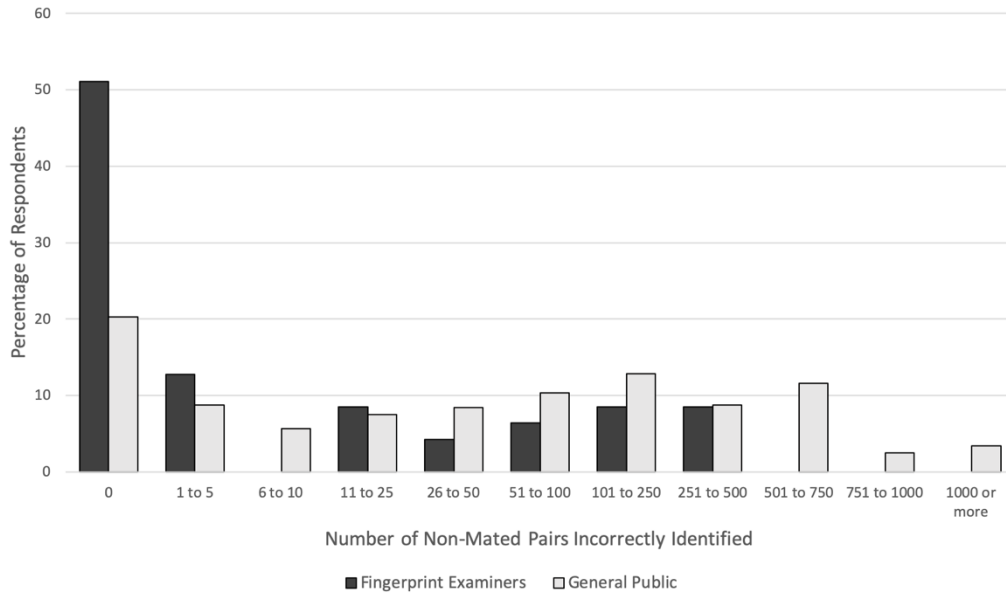


FIG. 5—

Percentage of examiners and members of the general public that incorrectly identify non-mated pairs by frequency category.