Modeling the Alignment between Societal Values and Fingerprint Examiner Decisions

Willa M. Mannering
Indiana University, Bloomington

Macgregor Vogelsang
Indiana University, Bloomington

Thomas A. Busey
Indiana University, Bloomington

Fred L. Mannering
University of South Florida, Tampa
Abstract

Fingerprint comparisons are conducted by human examiners who compare two impressions to estimate the weight of evidence in favor of one of two hypotheses: that the impressions came from the same source or from different sources. They use the weight of evidence along with internal decision criteria to come to a conclusion. However, the origin and rational of this internal decision criteria are poorly defined. The goal of this study is to determine if the moral values of society align with the current decision criteria, which exhibit a bias toward erroneous exclusion errors over erroneous identification errors. Subjects were asked to manipulate a web-based visualization that reflects the tradeoffs between erroneous exclusion decisions and erroneous identification decisions. Data from fingerprint examiners and novices 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 have a more conservative identification criterion than members of the general public, which aligns with error rate studies of fingerprint experimenters but suggests a disconnect between practitioners and the public. Demographic data was then used to determine possible factors that contribute to the difference in decision criteria placement. This dataset represents a rich framework for measuring, interpreting, and responding to the values and beliefs of a just and moral society as applied to forensic decision making.
1. Introduction

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. How much detail in agreement is sufficient to determine that the two impressions originate from the same finger? Perhaps surprisingly, there is currently no fixed standard for what constitutes sufficiency for purposes of identification. Because most fingerprint comparisons are conducted by human examiners, not computers, this lack of guidance from policymakers leads to a morally ambiguous task for fingerprint examiners. By contrast, consider the example of the legal length of a shotgun. In the United States, the National Firearms Act of 1934 legislated that shotguns with a barrel length shorter than 18 inches are illegal unless registered with the Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF)—making a clear definition between legal and illegal barrel lengths. The law specifies a threshold (18 inches), and laboratories have developed procedures to apply the law (take a calibrated ruler, align it to the breach of the shotgun, and measure the barrel length). However, fingerprint comparisons have neither a legally defined threshold for what constitutes sufficiency, nor a legally specified procedure to determine whether a given comparison produces a result that exceeds that threshold. Instead, the forensic community has developed a set of procedures known as ACE-V, which stands for analysis, comparison, evaluation and verification (Ashbaugh, 1999; National Institute of Justice, 2011; Expert Working Group on Human Factors in Latent Print Analysis, 2012). 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 matches a region on the comparison impression.

In lieu of a fixed standard, 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 (Ulery et al., 2011). 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 (Snodgrass et al., 2004; Vanselst and Merikle, 1993). 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.

All of these procedures create an ambiguous and potentially concerning situation: examiners as a community 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 gun barrel length 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? Such a determination 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), and (3) the prior probability of mated and non-mated pairs (essentially how good the detectives are at finding the correct suspect).

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.

The goals of this study are to determine whether the current thresholds for identification and exclusion decisions within the fingerprint examination community are compatible with the values of society, and to determine which demographic attributes are relevant for predicting the location of the thresholds. 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.
1.1 Latent Print Comparisons

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. 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.

1.2 Error Rates in Fingerprint Comparisons

Ulery et al. (2011) 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. Five examiners made erroneous identification errors for an overall erroneous identification rate of 0.1%. Eighty-five percent of 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 to attention the current error bias in latent fingerprint examinations: examiners are much more likely to make an erroneous exclusion error than an erroneous identification error. That is, they are more likely to conclude that a pair of fingerprints do not match when in reality they do. This bias has also been found in several eye tracking studies which conclude that while examiners typically have fairly high accuracy, most of the errors they make are erroneous exclusions rather than erroneous identifications (Busey, Yu, Wyatte, & Vanderkolk, 2013; Busey, Nikolov, Yu, Emerick, & Vanderkolk, 2016). The bias may be present for several reasons. It is possible that society has placed more importance on making sure innocent people are not put in jail, which would pressure the examiners into making more erroneous exclusion errors over erroneous identification errors. In addition, Ulery et al.’s (2011) study 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 no 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.

While the error bias found by Ulery et al. (2011) reflects how examiners translate their own moral values and personal tradeoffs into the decision criteria, it may not accurately reflect the values that society holds. This is compounded by the fact that the consequences of changing the
decision criteria will typically have both positive and negative outcomes. In the case of making an identification decision, if an examiner requires less detail in agreement before making an identification, they could potentially contribute information that will increase the number of criminals in jail while also increase the number of innocent people in jail. Moving the decision criterion in the opposite direction would help keep more innocent people out of jail but could also let more criminals free. Because this tradeoff involves negative consequences regardless of the direction, it is considered a taboo tradeoff (Fiske and Tetlock, 1997).

1.3 Taboo Tradeoff

A taboo tradeoff is defined as a “mental comparison or social transaction that violates deeply-held normative intuitions about the integrity, even sanctity, of certain forms of relationship and of the moral-political values that derive from those relationships” (Fiske and Tetlock, 1997, pg. 256). We believe that asking people to explicitly state their values when it comes to incarceration and exoneration is a taboo tradeoff and might cause discomfort among participants, because they understand both the consequences of an innocent person in jail and a guilty person going free who might commit additional crimes. Despite these conflicts, an individual’s solution to the taboo tradeoff ultimately reflects their own attempt to resolve the tradeoff in a way that optimizes the outcomes that fit their personal values.

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 jail and are intolerant of the large percentage of inconclusive decisions found by Ulery et al. (2011) that might contribute to increases in crime rates. However, we cannot simply ask participants whether it is better to put innocent persons in jail or let guilty persons go free, because the tradeoff is quite complex and depends on multiple factors. We need an accurate representation of the quantitative nature of the
taboo tradeoff. To do this, we modeled the error rate data of Ulery et al. (2011) using signal detection theory (Macmillan and Creelman, 2005).

1.4 Modeling the Taboo Tradeoff with Signal Detection Theory

In order to estimate the tradeoffs that occur as an examiner adopts different decision criteria, we constructed a mathematical representation of the underlying distributions of mated and non-mated fingerprint pairs taken from the Ulery et al. (2011) error-rate data as reproduced in Table 1.

Table 1. Response proportions found in Ulery et al. (2011) data

<table>
<thead>
<tr>
<th>Pair Type</th>
<th>Exclusion</th>
<th>Inconclusive</th>
<th>Identification</th>
<th>Total Mates or Non-mates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mates</td>
<td>0.075</td>
<td>0.311</td>
<td>0.614</td>
<td>1</td>
</tr>
<tr>
<td>Non-mates</td>
<td>0.887</td>
<td>0.111</td>
<td>0.001</td>
<td>1</td>
</tr>
</tbody>
</table>

These response proportions are the result of both the abilities of examiners to separate mated from non-mated pairs, as well as the typical decision criteria 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 criteria. To model these underlying distributions of mated and non-mated pairs of impressions, we assume that a comparison results in an 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 very small amount of perceived detail in agreement, an examiner is likely to produce an exclusion 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 made 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. 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. Fig. 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.

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 taboo 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 taboo tradeoff despite the inability to directly assess the goodness of fit.
The summary of the Ulery et al. (2011) data by the signal detection theory model allows for the exploration of the consequences of implementing different decision criteria. 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 criminals in jail would increase and the number of innocent people in jail 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 taboo tradeoff in forensics, we developed a web-based visualization (see Fig.

Figure 1. Representation of the Ulery et al. (2011) data using signal detection theory

Signal Detection Model Fit of Ulery et al. (2011)

- Nonmated Distribution
- Mated Distribution
- Exclusion Criterion
- Identification Criterion

Amount of Perceived Detail in Agreement

Density

-4 -2 0 2 4 6 8 10
2, Fig. 3, and Method for more detail) based on this model. The visualization quantifies the taboo 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 innocent people in jail, they can shift the decision criteria toward a more conservative value. However, this will simultaneously affect the number of criminals in jail, 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. The data we collected allows us to compare experts to members of the general public to determine whether the decisions made by examiners aligns with the values of society. We also use demographic data to determine the factors that are associated with different resolutions of the taboo tradeoff that exists in forensic decision-making.

2. Method

2.1 Participants

A total of 454 subjects were considered for data analysis. The subjects used for analysis were split into two groups: examiners and novices. There were 376 novice subjects included in the analysis. 171 were undergraduates attending Indiana University, members of the Bloomington, Indiana community, or members of the surrounding community. The remaining 205 novice 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 novice 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 (Fig. 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.
Fig. 2 and Fig. 3 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 taboo tradeoff: https://www.iu.edu/~buseylab/fingerprintvalues/?sandbox.

The horizontal axis in each figure represents an evidence axis, which we described as the amount of perceived detail in agreement between two impressions. The two clouds of points represent the mated (top) and non-mated (bottom) pairs of impression. 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\(^1\). Each point in Fig. 2 and Fig. 3 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 vertical location of each point reflects the ground truth (mated or non-mated), and the points are further jittered vertically to make overlapping points visible. 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 Fig. 2 and Fig. 3 are partitioned by two decision criteria into three groups that represent the decisions made by the examiner. These decision thresholds 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

---

\(1\) 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.
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.

2.3 Subjective Utilities and Prior Probabilities

The taboo 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 jail and the number of guilty people potentially in jail. The exact nature of the taboo 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 taboo 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 Fig. 2 uses the same base rates as provided by the original Ulery et al. (2011) 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. Fig. 3 illustrates a 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 four comparisons conducted by examiners is on a mated impression. To assess whether participants are sensitive to the prior, we altered the number of mated pairs as shown in Fig. 3 for half of our participants while keeping the number of non-mated pairs unchanged. For comparison, Fig. 2 represents the high mated condition, and Fig. 3 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.

3. 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. (2011) 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 taboo 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 3 seconds to consider the web visualization and did not move the sliders from their starting positions. For a complete breakdown of subject inclusion please see Appendix C.

For the sake of succinctness and clarity, the six outcomes represented in the model will be hereby referred to as: innocents wrongly identified, criminals correctly identified, innocents correctly identified, criminals wrongly identified, mated inconclusive and non-mated inconclusive. Looking at the web visualizations (Fig. 2 and Fig. 3), the innocents wrongly identified outcome is located to the right of the identification slider in the non-mated distribution (bottom right, red cloud), the criminals correctly identified outcome is located to the right of the identification slider in the mated distribution (top right, green cloud), the innocents correctly identified outcome is located to the left of the exclusion slider in the non-mated distribution (bottom left, green cloud), the criminals wrongly identified 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). While this is the terminology we chose to use for sake of clarity, this does not necessarily mean that every case results in incarceration or exoneration. In addition to the assumptions underlying our signal detection theory representation of the Ulery et al. (2011) 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 Fig. 2 and Fig. 3. We assume
that all comparisons involve fingerprints that are relevant to a criminal case and are potentially inculpatory in nature. Examiners might include a fingerprint on a weapon left at the scene of a crime or a latent print recovered from a windowpane after a robbery. 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 jail. 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.

4. 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. (2011)?

3) Are the identification and exclusion criteria sensitive to the base rates of mated and non-mated pairs, as assess through the two mated conditions illustrated in the Fig. 2 and Fig. 3 visualizations?

4) What aspects of the demographic data predict a participant’s placement of the decision criteria and the number of innocents in jail 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 used 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).

4.1 Results of Criteria Placement Comparisons

Tables 2 and 3 show the mean criteria placement and standard deviation for the exclusion criterion and identification criterion, respectively, for examiners and novices in both the high mated and low mated conditions.

Table 4 shows the results of running a two-tailed, independent-samples t-test assuming unequal variance comparing the average placement of the exclusion and identification criterion for high and low mated groups between novices and examiners.

There was a significant difference in identification criterion placement between novices (M = 2.1879, SD = 1.2915, N = 232) and examiners (M = 2.9472, SD = 0.8058, N = 37) in the high mated condition; t(267) = 4.777, p = 9.67e-6, Cohen’s d = 0.6137 and between novices (M = 2.5127, SD = 1.3376, N = 232) and examiners (M = 2.9708, SD = 0.8485, N = 37) in the low mated condition; t(184) = 2.626, p = 0.0099, d = 0.3672. There was not a significant difference in exclusion criterion placement between novices (M = -0.0082, SD = 1.0668, N = 145) and
Table 2. Placement of Identification Criteria

<table>
<thead>
<tr>
<th></th>
<th>High Mated</th>
<th>Low Mated</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Examiner</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>2.9472</td>
<td>2.9708</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.8058</td>
<td>0.8485</td>
</tr>
<tr>
<td>Sample Size</td>
<td>37</td>
<td>41</td>
</tr>
<tr>
<td><strong>Novice</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>2.1879</td>
<td>2.5127</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>1.2915</td>
<td>1.3376</td>
</tr>
<tr>
<td>Sample Size</td>
<td>232</td>
<td>145</td>
</tr>
</tbody>
</table>

Table 3. Placement of Exclusion Criteria

<table>
<thead>
<tr>
<th></th>
<th>High Mated</th>
<th>Low Mated</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Examiner</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>-0.4360</td>
<td>0.2143</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>1.3085</td>
<td>1.1964</td>
</tr>
<tr>
<td>Sample Size</td>
<td>37</td>
<td>41</td>
</tr>
<tr>
<td><strong>Novice</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>-0.0082</td>
<td>0.2146</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>1.0668</td>
<td>0.9393</td>
</tr>
<tr>
<td>Sample Size</td>
<td>232</td>
<td>145</td>
</tr>
</tbody>
</table>

Table 4. Results of t-test comparison

<table>
<thead>
<tr>
<th></th>
<th>Examiner Mean</th>
<th>Novice Mean</th>
<th>t</th>
<th>df</th>
<th>p</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exclusion Criteria</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Mated</td>
<td>-0.4360</td>
<td>-0.0082</td>
<td>-1.866</td>
<td>267</td>
<td>0.0686</td>
<td>0.3879</td>
</tr>
<tr>
<td>Low Mated</td>
<td>0.2143</td>
<td>0.2146</td>
<td>-0.002</td>
<td>184</td>
<td>0.9987</td>
<td>.0003</td>
</tr>
<tr>
<td><strong>Identification Criteria</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Mated</td>
<td>2.9472</td>
<td>2.1879</td>
<td>4.777</td>
<td>267</td>
<td>9.67e-6</td>
<td>0.6137</td>
</tr>
<tr>
<td>Low Mated</td>
<td>2.9708</td>
<td>2.5127</td>
<td>2.626</td>
<td>184</td>
<td>0.0099</td>
<td>0.3672</td>
</tr>
</tbody>
</table>
examiners (M = -0.4360, SD = 1.3085, N = 41) in the high mated condition; $t(267) = -1.866$, $p = 0.0686$, $d = 0.3879$ or between novices (M = 0.2146, SD = 0.9393, N = 145) and examiners (M = 0.2143, SD = 1.1964, N = 41) in the low mated condition; $t(184) = -0.0015$, $p = 0.9987$, $d = 0.0003$.

To summarize these results, Fig. 4 shows the average placement of the identification and exclusion criteria for both examiners and novices 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 criteria placement as predicted by Ulery et al. (2011) is represented on Fig. 4 as black lines.

This analysis shows that examiners tend to be more conservative than novices when it comes to the placement of the identification criteria in both the high and low mated conditions. This suggests that novices are less concerned with avoiding erroneous identification errors and may be willing to accept more of these errors in exchange for more correctly identified criminals. Additionally, Fig. 4 shows a large difference in the location of the exclusion criteria between our study and the Ulery et al. (2011) 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.
Figure 4. Average Criteria Placement. All 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 black bars represent the location of the ID and Exclusion decision criteria of examiners who participated in the Ulery et al. (2011) study.

4.2 Associations between Demographic Data and Criteria Placement

The resolution of the taboo 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 criteria that require stronger evidence before incarcerating a suspect. What factors might affect whether someone holds a decision criterion that is more conservative (i.e. moving to the right along the axis in Fig. 2), or less conservative (i.e.
moving to the left in Fig. 2)?

To answer this question, we developed two statistical models that address possible associations between our demographic survey data and the decision criterion set by each participant. The first model we estimated is a bivariate tobit model (Tobin, 1958) 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 criteria in a systematic way.

We suspect that many of the subjects are heavily basing their final decision criteria placement on the number of innocents in jail. Our suspicions are corroborated by the existing trend towards heavily punishing fingerprint examiners for making erroneous identifications while the punishment for erroneous exclusions is comparatively less severe. To address this, 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 jail? If so, which factors increase or decrease the acceptable amount of people in jail 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). This model is based on the assumption that a subject is either in one of two states, the zero-count state or the count state but not both (since as analysts we cannot determine which state a subject is in with certainty, a probabilistic splitting function is used). If a subject is in the zero-count state, under no circumstances would they ever allow any innocent persons to be put in jail. It is useful to imagine subjects placed in this category as being philosophically opposed to the idea of allowing any innocents to be incarcerated.
Unfortunately, due to the nature of the decision tradeoff, these subjects must accept the fact that less criminals will be incarcerated if no erroneous identification errors are allowed. On our web visualization, these subjects would move their identification slider far to the right to prevent all erroneous identifications. On the other hand, if a subject is in the count state then they may tolerate some innocents in jail. In this case, the model will estimate which demographic factors influence the number of innocents in jail for each subject. It is important to note that subjects in this state may also end up putting zero innocents in jail. Though it seems like the count state and zero-count state could overlap, it is useful to think of their distinction in terms of philosophy. Those in the zero-count state will never tolerate innocents in jail while those in the count state are not fundamentally against innocents in jail but may end up putting zero innocents in jail.

4.3 Summary Statistics of Demographic Data

Table 5 shows the summary statistics of all survey responses used in the tobit and negative binomial models discussed below. 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
Table 5. Summary statistics of survey responses

<table>
<thead>
<tr>
<th>Variable Description</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of criminals wrongly identified as innocent</td>
<td>115.18</td>
<td>271.99</td>
<td>0</td>
<td>2982</td>
</tr>
<tr>
<td>Number of criminals inconclusive</td>
<td>984.14</td>
<td>1142.29</td>
<td>0</td>
<td>5955</td>
</tr>
<tr>
<td>Number of criminals correctly identified as criminals</td>
<td>2688.16</td>
<td>2043.58</td>
<td>13</td>
<td>5909</td>
</tr>
<tr>
<td>Number of innocents wrongly identified as criminals</td>
<td>188.62</td>
<td>329.78</td>
<td>0</td>
<td>2495</td>
</tr>
<tr>
<td>Number of innocents inconclusive</td>
<td>1763.86</td>
<td>1090.78</td>
<td>0</td>
<td>4000</td>
</tr>
<tr>
<td>Number of innocents correctly identified as innocent</td>
<td>2091.14</td>
<td>1082.26</td>
<td>0</td>
<td>4000</td>
</tr>
<tr>
<td>Age of respondent in years</td>
<td>34.96</td>
<td>13.41</td>
<td>18</td>
<td>80</td>
</tr>
<tr>
<td>Annual income of respondent (in dollars)</td>
<td>86,972</td>
<td>68,192</td>
<td>0</td>
<td>320,000</td>
</tr>
<tr>
<td>Number of people in household</td>
<td>2.97</td>
<td>1.51</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Number of children in household</td>
<td>0.97</td>
<td>1.26</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

**Percent-response data**

- Percent of respondents unwilling to wrongly identify any innocents: 24.12
- Percent male/ female: 46.6/ 53.4
- Percent of respondents by relationship status: single/ married/ divorced/ separated/ widowed: 54.2/ 39.6/ 4.6/ 0.8/ 0.8
- Percent of respondents by educational level: high school/ college student/ some college/ bachelor's degree/ master's degree/ professional degree/ doctorate: 8.4/ 26.0/ 14.4/ 32.3/ 14.1/ 0.8/ 4.1
- Percent of respondents by experience with fingerprint examinations: no experience/ trainee/ less than 2 years/ 2-4 years/ 5-7 years/ 8-12 years/ 13-20 years/ 20+ years: 84.7/ 2.2/ 1.6/ 2.6/ 1.3/ 3.2/ 2.3/ 2.1
- Percent of respondents by association with justice system: no association/fingerprint examiner/ prosecutor/ police/ defense attorney/ judge: 84.4/ 13.1/ 0.0/ 1.9/ 0.3/ 0.3
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.

Fig. 5 is a frequency analysis of how many innocents in jail the subjects tolerate. This figure shows that nearly 90 of the 369 subjects used in the tobit and negative binomial models were unwilling to wrongly identify any innocents (24.12 percent of the sample as shown in Table 5). 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 wrongly identify any innocents (state 1) and another group who are willing to wrongly identify some number of innocents in their quest to correctly identify as many criminals as possible (state 2). We attempt to account for this two-state possibility by fitting our data to the negative binomial model later in this paper.

4.4 Bivariate Tobit Model

Our initial analysis using the t-tests was restricted to grouping subjects by whether they were examiners or novices 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 innocents wrongly identified to criminals correctly identified. This ratio is representative of the position of the identification criterion, as the innocents wrongly identified and criminals 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 innocents wrongly identified to criminals correctly identified. This ratio will have a minimum of zero (when no innocents are wrongly identified) and will be a continuous variable when one or more innocents are wrongfully 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 criminals wrongly identified and innocents correctly identified, 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 (Tobin, 1958). 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 (Anastasopoulos et al., 2012),
MODELING SOCIETAL VALUES OF EXAMINER DECISIONS

\[ Y_{nk}^* = \beta_{nk} X_{nk} + \varepsilon_{nk}, \quad n = 1, 2, \ldots, N, \quad k = 1, 2 \]

\[ Y_{nk}^* = Y_{nk} \quad \text{if} \quad Y_{nk}^* > 0 \]

\[ = 0 \quad \text{if} \quad Y_{nk}^* \leq 0, \]

where \( Y_{n1}^* \) (with \( k=1 \)) is latent variable of the ratio of innocents wrongfully identified to criminals correctly identified for subject \( n \) and is observed only when positive, \( Y_{n2}^* \) (with \( k=2 \)) is latent variable of the ratio criminals wrongly identified to innocents correctly identified for subject \( n \) and is observed only when positive, \( N \) is the total number of subjects, \( Y_{n1} \) is the observed ratio of innocents wrongfully identified to criminals correctly identified dependent variable, \( Y_{n2} \) is the observed ratio of criminals wrongly identified and innocents correctly identified dependent variable, \( X_{n1} \) and \( X_{n2} \) are vector of explanatory variables corresponding to ratio equations, \( \beta_1 \) and \( \beta_2 \) are vectors of estimable parameters, and \( \varepsilon_{n1} \) and \( \varepsilon_{n2} \) are bivariate normally and independently distributed error terms with zero means, variances \( \sigma_{\varepsilon_{n1}}^2 \) and \( \sigma_{\varepsilon_{n2}}^2 \), correlation \( \rho \), and correlation matrix,

\[
\Sigma_{\varepsilon_{nk}} = \begin{pmatrix}
\sigma_{\varepsilon_{n1}}^2 & \rho_{\varepsilon_{n1}\varepsilon_{n2}} \sigma_{\varepsilon_{n1}} \sigma_{\varepsilon_{n2}} \\
\rho_{\varepsilon_{n1}\varepsilon_{n2}} \sigma_{\varepsilon_{n1}} \sigma_{\varepsilon_{n2}} & \sigma_{\varepsilon_{n2}}^2
\end{pmatrix}
\]

This tobit model can be readily estimated by standard maximum likelihood methods (Washington et al., 2011).

4.5 Bivariate Tobit Model Estimation Results

The bivariate tobit model estimation results are presented in Table 6. 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 (Washington et al., 2011, Mannering
et al., 2016). 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 $\beta_1$ and $\beta_2$ are vectors.

Turning first to the estimation results for the ratio of innocents wrongfully identified to criminals correctly identified, the fingerprint examiner indicator resulted in a negative parameter indicating a smaller ratio (fewer innocents wrongly identified per criminals 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 innocents wrongly identified per criminals 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 innocents wrongly identified per criminals correctly identified). Finally, the high-criminals indicator (1 if subject faces a criminal cluster of 5969 criminals, 0 if 1000 criminals) was understandably significant because changing the value of the denominator (number of criminals) 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 innocents wrongfully identified to criminals correctly identified.

Turning to the second tobit equation (the observed ratio of criminals wrongly identified to innocents correctly identified), Table 6 shows that none of the variables included in the model estimation produced parameter estimates that were significantly different from zero, except for the high-criminals indicator (1 if respond faces a criminal cluster of 5969 criminals, 0 if 1000
criminals) 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 criminals wrongly identified to innocents correctly identified cannot be predicted with our collected socio-demographic information, even though the ratio of innocents wrongfully identified to criminals 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.

4.6 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 innocents wrongly identified. If a subject was an older male (over 65 years) they tended to be less conservative, or more likely to tolerate innocents wrongly identified. Finally, as more children are
Table 6. Bivariate Tobit Model of Ratio of Innocents Wrongfully Identified to Criminals Correctly Identified and of the Criminals Wrongly Identified to Innocents Correctly Identified.

<table>
<thead>
<tr>
<th>Variable Description</th>
<th>Estimated Parameter</th>
<th>( t ) statistic&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ratio Innocents Wrongfully Identified to Criminals Correctly Identified</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.218</td>
<td>5.12***</td>
</tr>
<tr>
<td>Fingerprint examiner indicator (1 if respondent is a fingerprint examiner, 0 otherwise)</td>
<td>–0.160</td>
<td>–3.62***</td>
</tr>
<tr>
<td>Number of children in respondent’s household if not a fingerprint examiner</td>
<td>0.029</td>
<td>1.86*</td>
</tr>
<tr>
<td>Older male indicator (1 respondent is a male 65 years old or older, 0 otherwise)</td>
<td>0.217</td>
<td>4.15***</td>
</tr>
<tr>
<td>White indicator (1 if respondent identified themselves as being white, 0 otherwise)</td>
<td>–0.114</td>
<td>–2.79***</td>
</tr>
<tr>
<td>High criminals indicator (1 if respond faces a criminal cluster of 5969 criminals, 0 if 1000 criminals)</td>
<td>–0.165</td>
<td>–3.42***</td>
</tr>
<tr>
<td><strong>Ratio of Criminals Wrongly Identified to Innocents Correctly Identified</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.00051</td>
<td>0.02</td>
</tr>
<tr>
<td>Fingerprint examiner indicator (1 if respondent is a fingerprint examiner, 0 otherwise)</td>
<td>0.00364</td>
<td>0.35</td>
</tr>
<tr>
<td>Number of children in respondent’s household if not a fingerprint examiner</td>
<td>0.00187</td>
<td>0.37</td>
</tr>
<tr>
<td>Older male indicator (1 respondent is a male 65 years old or older, 0 otherwise)</td>
<td>0.01247</td>
<td>0.47</td>
</tr>
<tr>
<td>White indicator (1 if respondent identified themselves as being white, 0 otherwise)</td>
<td>0.00615</td>
<td>0.49</td>
</tr>
<tr>
<td>High criminals indicator (1 if respond faces a criminal cluster of 5969 criminals, 0 if 1000 criminals)</td>
<td>0.05013</td>
<td>2.80***</td>
</tr>
<tr>
<td>Correlation coefficient between equations, ( \rho_{\varepsilon_1, \varepsilon_2} )</td>
<td>–0.121</td>
<td>–0.87</td>
</tr>
<tr>
<td>Number of observations</td>
<td>369</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Confidence level (two-tailed test): * greater than 90%; ** greater than 95%; ***greater than 99%
added to a household, subjects become less conservative, thus tolerating more innocents wrongly 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 Fig. 2 and Fig. 3) was fit to a zero-inflated negative binomial model which focuses solely on the number of innocents incorrectly identified as guilty.

4.7 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 innocents that individual subjects are willing to falsely identify as being at the crime scene. 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 wrongly 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, 2014, 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 $i_n$ innocents as guilty,

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

(3)
where $\lambda_n$ is the Poisson parameter for subject $n$, which is the subject's expected number of erroneous identification errors. To incorporate explanatory variables, Poisson regression specifies the Poisson parameter $\lambda_n$ as a log-linear function,

$$
\lambda_n = EXP(\beta X_n)
$$

where $X_n$ is a vector of explanatory variables that determine the number of innocents wrongly identified by subject $n$ and $\beta$ is a vector of estimable parameters (Washington et al., 2011).

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 ($VAR[i_n] >> E[i_n]$) 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 innocents wrongly 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 $\alpha$. The addition of this gamma-distributed error term 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 gives a function that follows a negative binomial distribution, and the resulting negative binomial regression (also sometimes referred to as the Poisson-Gamma regression since a gamma function is added to account for overdispersion) can be written as:

$$
P(i_n) = \left( \frac{1/\alpha}{(1/\alpha) + \lambda_n} \right)^{i_n} \frac{\Gamma[(1/\alpha) + i_n]}{\Gamma(1/\alpha) i_n !} \left( \frac{\lambda_n}{(1/\alpha) + \lambda_n} \right)^{i_n}
$$

where $\Gamma(.)$ is a gamma function. Please note that in Equation 6 the Poisson regression is a limiting model of the negative binomial regression as $\alpha$ approaches zero. Thus, if $\alpha$ is significantly
different from zero, the negative binomial is appropriate and, if it is not, the Poisson model is appropriate because with \( \alpha \) 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 (Washington et al., 2011).

However, for the case of innocents wrongly 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 innocents should never be wrongly identified, which would put these subjects in a zero-count state. Other subjects may be willing to wrongly identify innocents (to correctly identify more guilty people), 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 (Lambert, 1992; Malyshkina and Mannering, 2009; Washington et al., 2011). The zero-inflated Poisson model is set up as (with \( i_n \) being innocents wrongly identified),

\[
i_n = 0 \text{ with probability } p_n + \left(1 - p_n\right) \text{EXP}(-\lambda_n)
\]

\[
i_n = i \text{ with probability } \frac{\left(1 - p_n\right) \text{EXP}(-\lambda_n) \lambda_n^i}{i_n!}
\]

(7)

where \( p_n \) is the probability of subject \( n \) being in the zero state, and \( i \) is the number of innocents wrongly identified. Note that the upper portion of equation 7 is a combination of a having a zero because of being in the zero state (\( p_n \)), and having a zero because of being in the count state.
\[(1-p_n)\exp(-\lambda_n)\] with \(\exp(-\lambda_n)\) being the probability of zero in the count state (that is, with \(i_n = 0\) in equation 1).

The zero-inflated negative binomial regression model follows a similar formulation with,

\[
i_n = 0 \text{ with probability } p_n + (1 - p_n)\left[\frac{1/\alpha}{(1/\alpha) + \lambda_n}\right]^{\lambda_n/\alpha}
\]

\[
i_n = i \text{ with probability } (1 - p_n)\left[\frac{\Gamma\left((1/\alpha) + i_n\right)u_n^{1/\alpha}(1 - u_n)^i}{\Gamma(1/\alpha)i_n!}\right],
\]

where \(\mu_i = (1/\alpha)\left[(1/\alpha)+\lambda_i\right]\). Maximum likelihood methods are again used to estimate the parameters of these two zero-inflated modeling alternatives. Also, note that the splitting regime used in zero-inflated models (to determine \(p_n\) in Equations 7 and 8) is typically assumed to follow a probit (normal) probability process, where the probability of being in the zero state \((p_n)\) is:

\[
p_n = \Phi(\beta \cdot X_z)
\]

where \(\Phi(.)\) is the normal cumulative distribution function, \(X_z\) is a vector of characteristics that determine the probability of being in the zero state and \(\beta_z\) is vector of estimable parameters.

To determine if the dual state zero-inflated models should be chosen over traditional single-state count models, Vuong (1989) proposed a test statistic for non-nested models that is well suited for situations where the distributions (Poisson or negative binomial) are specified. The statistic is calculated as (for each subject \(n\)),

\[
m_n = LN\left\{\frac{f_1(i_n|X_n)}{f_2(i_n|X_n)}\right\}
\]

where \(f_1(i_n|X_n)\) is the probability density function of model 1, and \(f_2(i_n|X_n)\) is the probability density function of model 2. Using this, Vuongs' statistic for testing the non-nested hypothesis of model 1 versus model 2, with total number of subjects \(N\), is (Washington et al., 2011),
where $\bar{m}$ is the mean $\left(1/n \sum_{n=1}^{N} m_n\right)$ and $S_m$ is standard deviation, and $N$. This Vuong statistic is asymptotically standard normal distributed (to be compared to z-values). Using a 97.5% confidence level, if $V$ in Equation 10 is less than 1.96 ($z$-value for the 97.5% confidence level, one-tailed test), the test does not support the selection of one model over another. Large positive values of $V$ (greater than 1.96 at the 97.5% confidence level) favor model 1 over model 2, whereas large negative values support model 2. For example, if comparing a standard negative binomial regression and a two-state zero-inflated negative binomial regression, in Equation 9, $f_1(.)$ would be the density function of the zero-inflated negative binomial and $f_2(.)$ would be the density function of the negative binomial model. In this case, assuming a 97.5% critical confidence level, if $V > 1.96$ the statistic would favor the two-state zero-inflated negative binomial regression and a value of $V < –1.96$ would favor a standard the negative binomial regression (values in between would mean that the test was inconclusive).

Finally, to assess the effects of individual explanatory variables in the $X_n$ vector, on the mean number of innocents ($\lambda_n$) wrongly 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 $X_n$ vector) has the mean number of innocents wrongly identified and is computed as,

$$\text{ME}_{\lambda_n} = \frac{\partial \lambda_n}{\partial x_{nk}} = \beta_k \text{EXP}(\beta X_n)$$

(12)

where $k$ is the $k^{th}$ element of the $X_n$ 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.

4.8 Zero-Inflated Negative Binomial Model Estimation Results

Table 7 presents the summary statistics for variables found to be statistically significant in the
zero-inflated negative binomial estimation, and Table 8 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 innocents wrongly identified). The value of the Vuong statistic
shown in Table 8 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 $\chi^2$ 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.

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 criminals presented to subjects was 1000 for some and 5969 for others. This suggests
Table 7. Summary statistics for negative binomial variables

<table>
<thead>
<tr>
<th>Variable Description</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fingerprint examiner indicator (1 if respondent is a fingerprint examiner, 0 otherwise)</td>
<td>0.131</td>
<td>0.334</td>
</tr>
<tr>
<td>Lower income indicator (1 respondent’s household income is less than $50,000 per year, 0 otherwise)</td>
<td>0.374</td>
<td>0.485</td>
</tr>
<tr>
<td>Older male indicator (1 respondent is a male 65 years old or older, 0 otherwise)</td>
<td>0.033</td>
<td>0.178</td>
</tr>
<tr>
<td>White indicator (1 if respondent identified themselves as being white, 0 otherwise)</td>
<td>0.778</td>
<td>0.416</td>
</tr>
<tr>
<td>Young-age indicator (1 if respondent is less than 25 years old, 0 otherwise)</td>
<td>0.290</td>
<td>0.454</td>
</tr>
<tr>
<td>Number of children in respondent’s household if not a fingerprint examiner</td>
<td>0.873</td>
<td>1.25</td>
</tr>
</tbody>
</table>

the possibility that subjects’ “innocent” decisions may be influenced by the number of criminals presented in their experiment. To test for this, the sample was split in two; those subjects facing 1000 criminals and those facing 5969 criminals. 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 7), $LL(\beta_{1000})$ is the log-likelihood at convergence of the model using only data from subjects facing 1000 criminals, and $LL(\beta_{5969})$ is the log-likelihood at convergence of the model using only data from subjects facing 5969 criminals. In this test the same variables are used in all three models and this $X^2$ test statistic is $\chi^2$ 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 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 8. 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 8 (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 wrongly identify any innocents) 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 wrongly identifying innocents. 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 wrongly identifying innocents is a tolerable option. These, and potentially other factors, could be playing a role.

Table 8 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 8 suggest that the net effect is that individuals less than 25 years old were, on average, willing to wrongly identify 37.90 more innocents than other age groups.
Table 8. Zero-Inflated Negative Binomial of the Number of Innocents Wrongly Identified.

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

$^a$ Confidence level (two-tailed test): * greater than 90%; ** greater than 95%; ***greater than 99%
Finally, for the zero-state probabilities, Table 8 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 wrongly identify innocents).

Turning to variables found to be statistically significant in the count state (where a positive parameter estimate increases the number of innocents a subject is willing to wrongly convict, and a negative parameter decreases this number), subjects with household incomes less than $50,000 dollars per year were more willing to wrongly identify innocents (in their quest to identify as many criminals as possible). Marginal effects in Table 8 show that subjects in this lower-income bracket were, on average, willing to wrongly identify 82.14 more innocents than their higher-income counterparts.

In addition to lower income subjects, Table 8 shows males 65 years old or older were also more willing to wrongly identify innocents, with marginal effects showing on average willing to wrongly identify 191.55 more innocents 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 wrongly identify innocents in their quest for correctly identifying as many criminals as possible.

Interestingly, subjects who identified themselves as white (77.8% of the sample as indicated in Table 7), were significantly less likely to wrongly identify innocents relative to other ethnicities in the sample (marginal effects in Table 8 show they were, on average, willing to wrongly identify 126.93 fewer innocents).

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 wrongly identify innocents 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 wrongly identify 68.02 more innocents (on average) in their quest to make sure that as many criminals as possible were correctly identified.

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 innocents the subject was willing to wrongly identify. Still, the net effect of the zero and count-state marginal effects show (Table 8) that fingerprint examiners are willing to wrongly convict 33.10 fewer innocents 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, Equation 9, is a partial derivative that holds other characteristics fixed).

4.9 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 wrongly-identified innocents 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 innocents in jail 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 innocents in jail. On the other hand, subjects who identified as white were likely to tolerate fewer innocents in jail.

The result we find particularly interesting is the fact that fingerprint examiners relative to novices seem to be more likely to be in the zero-state—meaning, they are philosophically opposed to putting innocents in jail. There could be a number of reasons for this. First, the hypothetical nature of our experiment may make novices less aware of the consequences of wrongly identifying innocents. 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 wrongly identifying innocents is a tolerable option. These, and potentially other factors, could be playing a role.

5. General 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 amount of inconclusive decisions. According to the results shown in Table 4, there is a significant difference between the novice and examiner placement of the identification criterion. As shown in Table 4, the mean placement of the identification criterion by the novice subjects is lower than the examiner subjects for both the high mated and low mated conditions. This suggests that the novice subjects are more willing to accept an erroneous identification error in exchange for fewer inconclusive decisions. 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 novices. 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 slider positions in either fingerprint examiners or novices 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 innocent people in jail. 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 innocent people in jail than any other outcome.

It is possible that the difference in values between examiners and novices is caused by a gap in knowledge. 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 novices because the outcomes of these decisions directly affect their livelihood where as a novice 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 totally 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. (2011) 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 novices 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 the case. As the examiners were significantly more conservative in their placement of the identification criterion than novices, we suspect that both subject groups were appropriately considering the outcomes of criminal cases as intended.

The difference in 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. Currently, there is a bias against making erroneous identification errors in an examination relative to making erroneous identifications. If examiners do make an erroneous identification error, there is a possibility of losing their job or receiving a corrective action. However, the results of this study demonstrate that members of the general public could be more comfortable increasing the number of erroneous identification decisions in exchange for more correct identification of criminals.

5.1 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, and typically only an error rate is necessary to establish a discipline as scientifically valid (Daubert v. Merrell Dow Pharmaceuticals, 1993). However, the interpretation of that error rate is left to jurors (or defendants, in the 95% of cases that get plead out). 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, 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.

One practical implication of the present work is the formulation of the forensic decision-making process as a taboo 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 putatively. 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.
References


Appendix A: Instructional Video Transcription

The goal of this experiment is to measure the values of decisions in forensic examinations.

Unlike on TV, fingerprints are not matched by computers, they are matched by humans. Fingerprints collected from crime scenes are called *latent prints*, and the source is typically unknown. Fingerprints taken from suspects or victims are called *exemplar prints*, and the source is typically known. Latent prints tend to be distorted copies of the pattern on the finger, and they can also be degraded by visual noise. This often makes it difficult for a computer to analyze latent prints for crime scenes. Instead, fingerprint examiners visually compare the latent print against exemplar prints. Examiners use the amount of perceived detail in agreement between the two prints to decide whether they came from the same finger.

In reality there are two possible origins of print pairs. Mated pairs are pairs of fingerprints that actually came from the same finger. Non-mated pairs are pairs of fingerprints that came from different fingers. Of course, during normal casework we never know if a print pair is truly mated or truly non-mated. We can only conduct an examination to see if there is evidence that a print pair is mated or non-mated.

Once the examiner has conducted an analysis and comparison of the two prints, they can make one of three decisions. In an identification decision, in the opinion of the examiner the prints came from the same finger. In an exclusion decision, in the opinion of the examiner the prints came from different fingers. With an inconclusive decision, there is neither sufficient detail in agreement or disagreement to make a decision.

Note that there is a conceptual difference between the actual origin of a print (mated or non-mated) and the decision that the examiner makes (identification, exclusion, or inconclusive). For example, there can be a great deal of incidental similarity between non-mated pairs, which could lead to an error called an *erroneous identification*. On the other hand, noise and distortion can sometimes make mated pairs look very different, which could lead to an error called an *erroneous exclusion*. 
On the next slide I will illustrate the possible outcomes that can occur when an examiner makes a decision. It will be important to understand the consequences of the different outcomes, because we will see that these can sometimes depend on each other and trade off in complicated ways.

In an examination, there are two origins of prints: mated pairs and non-mated pairs. And the examiner makes these three decisions: exclusion, inconclusive, and identification. And depending on the origin and decision, there are different outcomes with different values. For example, if a pair is mated but the examiner said exclusion this is bad because the examiner contributes incorrect information that could help a criminal go free. This is what we call an erroneous exclusion. However, if an examiner says exclusion to a non-mated pair, this is a good outcome—the examiner contributes correct information that could help an innocent person and the detective will continue working the case. If they say identification to a mated pair, this is a good outcome because the examiner contributes correct information that could help put a criminal in jail. But if they say identification to a non-mated pair, this is a bad outcome because the examiner contributes incorrect information that could help put an innocent person in jail and could help the true criminal remain free. This is an erroneous identification. Finally, with inconclusive, it’s less clear the value that they provide. In this case, the examiner believes they have insufficient evidence to make either an identification or exclusion decision. And that’s true for both mated and non-mated pairs.

How does an examiner decide when to make an identification, inconclusive, or an exclusion decision? In part, it depends on their training and experience, looking at thousands of prints that are known to come from the same or different sources. However, their conclusions may also depend on the values of society. How important is it to provide evidence that puts guilty persons in jail and keeps innocent persons out of jail?

In the next section of the video, I will explain your part in the experiment, which will allow you to express your own values for different outcomes.
In a moment, you’ll see a visualization that looks like this. In the x-axis of this graph, this represents the amount of perceived detail in agreement as observed by a latent print examiner—which ranges from low, which is a low amount of perceived detail in agreement between two impressions, or high, which is high amount of perceived detail in agreement. The vertical axis separates non-mated pairs, which come from different fingers, from mated pairs, which are impressions from the same finger. An examiner makes one of three decisions: exclusion, inconclusive, or identification. And in this visualization, these are separated by the locations of two decision criteria, which are sliders that can move back and forth. In the table at the bottom, we represent the six possible outcomes that can occur. Note that as I move a slider back and forth, four of the six numbers will change in the tables. We would like you to carefully consider the number of cases that fall into each of these cells, which are color coded by their outcomes—with red is bad, green is good, and yellow is inconclusive. As you move these sliders back and forth, carefully read the number of cases and the outcomes that occur in each of these different cells and come up with a location of the two sliders that correspond to your own personal values. Once you’ve done that, you’ll click the save values button and then you’ll fill out some demographic data and then you’ll be done.
## Appendix B: Demographic Questions

<table>
<thead>
<tr>
<th>Demographic Information Requested</th>
<th>Possible Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>18-24, 25-34, 35-44, 45-54, 55-65, 65-74, 75 and older, Decline to Answer</td>
</tr>
<tr>
<td>Gender Identity</td>
<td>Male, Female, Decline to Answer</td>
</tr>
<tr>
<td>Ethnicity origin (or Race)</td>
<td>White, Hispanic or Latino, Black or African American, Native American or American Indian, Asian/Pacific Islander, Other, Decline to Answer</td>
</tr>
<tr>
<td>Level of Education</td>
<td>High School Diploma, Currently in college, Some college or Associates degree, Bachelor’s Degree, Master’s Degree, Professional Degree, Doctorate Degree, Decline to Answer</td>
</tr>
<tr>
<td>Experience with fingerprint examinations</td>
<td>Not a fingerprint examiner, Trainee (supervised comparisons), Less than 2 years (of unsupervised casework), 2-4 years, 5-7 years, 8-12 years, 13-20 years, Over 20 years, Decline to Answer</td>
</tr>
<tr>
<td>Association with the justice system</td>
<td>Not directly associated with the justice system; Involved with evidence gathering, interpretation, or analysis; Police officer, detective, or other public safety officer; Associated with a prosecutor’s office; Associated with criminal defense; Judge; Advocate for incarcerated individuals; Decline to Answer</td>
</tr>
<tr>
<td>Personal interactions with the justice system (check all that apply)</td>
<td>You have personally been a defendant in an arrest or trial, You have a family member or close friend who has been a defendant in an arrest or trial, You have served as a jury member at a trial, You have been called to testify as a witness, You have a family member associated with the justice system (e.g. police officer, judge, parole officer), You are associated with or support organizations that promote fairness in the justice</td>
</tr>
</tbody>
</table>
system (e.g. The Innocence Project, legal rights groups),
You support organizations that assist with law
enforcement (e.g. Fraternal Order of Police), Other,
None of these apply to me, Decline to Answer

Annual household income
Less than $20,000; $20,000 - $34,999; $35,000 -
$49,999; $50,000 - $74,999; $75,000 - $99,999;
$100,000 - $149,999; $150,000 - $199,999; $200,000
or more; Decline to Answer

Marital Status
Single, never married; Married or domestic partnership,
Widowed, Divorced, Separated, Decline to Answer

Number of People in household
1, 2, 3, 4, 5, 6 and over, Decline to Answer

Number of children in household
0, 1, 2, 3, 4, 5 and over, Decline to Answer

Zip code (optional)
Write in

Please provide any comments about this experiment that
you would like to share with the researchers. (optional)
Write in
### Appendix C: Subject Inclusion

Table C. Subjects considered and used for data analysis broken down by User ID

<table>
<thead>
<tr>
<th>User ID</th>
<th>User Category</th>
<th>Number of Subjects Considered for Analysis</th>
<th>Number of Subjects Disqualified for Data Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>201</td>
<td>examiners, random mated condition</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>202</td>
<td>examiners, random mated condition</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>204</td>
<td>examiners, random mated condition</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>206</td>
<td>examiners, random mated condition</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>208</td>
<td>examiners, random mated condition</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>211</td>
<td>examiners, random mated condition</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>214</td>
<td>examiners, random mated condition</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>218</td>
<td>examiners, random mated condition</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>219</td>
<td>examiners, random mated condition</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>220</td>
<td>examiners, random mated condition</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>313</td>
<td>Mechanical Turk</td>
<td>205</td>
<td>5</td>
</tr>
<tr>
<td>400</td>
<td>novice, random mated condition</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>425</td>
<td>novice, high mated condition</td>
<td>68</td>
<td>0</td>
</tr>
<tr>
<td>500</td>
<td>examiners, random high or low</td>
<td>39</td>
<td>0</td>
</tr>
<tr>
<td>700</td>
<td>novice, high mated condition</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>888</td>
<td>examiners, high mated condition</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>889</td>
<td>examiners, low mated condition</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>900</td>
<td>novice, early random mated condition</td>
<td>42</td>
<td>0</td>
</tr>
<tr>
<td>950</td>
<td>examiners</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>999</td>
<td>examiners</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>