Are Forensic Scientists Too Risk-Averse?

Willa M. Mannering[1], Macgregor D. Vogelsang[1], Thomas A. Busey[1], Fred L. Mannering[2]

[1]Department of Psychological and Brain Sciences, Indiana University

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

Abstract

The goal of this study is to determine if the moral values of society align with the current decision criteria in fingerprint comparisons, which currently exhibit a preference for preventing erroneous identification errors at the expense of erroneous exclusion errors. Subjects manipulated 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 are more risk-averse than members of the general public, although they align with error rate studies of fingerprint experimenters. Demographic data determined 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 society as applied to forensic decision making.

Keywords: *decision making, forensic science, signal detection theory*

Are Forensic Scientists Too Risk-Averse?

The tension between acquitting the guilty and convicting the innocent can be traced as far back as the 12th century scholar Moses Maimonides, who is quoted as: "*It is better and more satisfactory to acquit a thousand guilty persons than to put a single innocent one to death.*" In the West, this became known as Blackstone's Ratio: "*It is better that ten guilty persons escape than that one innocent suffer.*" Both of these philosophical statements express the idea that any decision in the judicial system has the potential for two kinds of error, and both attempt to make reasoned arguments for where the threshold should lie with respect to each kind of error. This tradeoff between errors does not lie solely at the level of acquittal or conviction at trial, but instead extends to *every decision made during a criminal investigation*. For example: What is the standard for a warrant? What percentage of cocaine in a substance qualifies it as illegal? What is the minimum gun barrel length for shotguns? Each of these examples has some statute or policy as determined by policy-makers or judges, who act as proxy for the greater society.

However, a glaring exception is found in the forensic pattern comparison disciplines, including fingerprint examinations, footwear comparisons, and firearms identification. Within the fingerprint discipline, there are no institutionally-mandated standards for what constitutes sufficient evidence to render a conclusion. Instead, the community of fingerprint examiners has collectively determined what constitutes sufficiency. Whether this de-facto sufficiency standard corresponds to the values of society is, as yet, untested. The present work introduces a novel paradigm that models the complex tradeoffs that occur during forensic decision making. We then use this paradigm to compare the values expressed by the public with those expressed by fingerprint examiners, as well as with the outcomes of fingerprint examination decisions. As we discuss below, the structure of the decisions made by fingerprint examiners may introduce an inherent risk aversion, which may not align with the values held by society.

A latent print examiner compares latent fingerprints from a crime scene to candidate exemplar fingerprints taken from a suspect to determine whether the two impressions may have originated from the same finger. To make this decision, they compare the amount of perceived detail in agreement between the two impressions and look for unexplainable differences. Most fingerprint comparisons are conducted by human examiners, not computers, and the lack of guidance from policymakers on what constitutes sufficient evidence for a conclusion leads to a morally ambiguous task for fingerprint examiners. As an alternative to a mandated, quantifiable standard, the forensic community has developed a set of procedures known as ACE-V, which stands for analysis, comparison, evaluation and verification (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[1].

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.,

[1] A point-based system also suffers from the ambiguity of what constitutes a point and whether it corresponds with a region on the comparison impression.

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 (Dror, 2016).

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

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

from non-mated pairs (which depends on both the quality of the images and the training of the examiners), (3) the prior probability of mated and non-mated pairs (essentially how good the detectives are at finding the correct suspect) and (4) how the evidence is used to ultimately achieve a result in a criminal trial. As we will see, there is no perfect placement of the thresholds that will eliminate all errors. However, there may be an optimal placement of the thresholds that reflects the values of society and therefore indirectly estimates the cost of various outcomes.

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, whether experts and members of the general public hold similar values, 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.

Background and Motivation

During normal casework, fingerprint examiners compare latent fingerprints processed from a crime scene and exemplar prints taken in a controlled environment with a known donor individual. Latent prints are often distorted, incomplete, and corrupted by visual noise. Exemplar prints are typically higher quality than the latent prints. The ground truth (typically unknown and unknowable by the examiner) can either be mated fingerprints or non-mated fingerprints. Mated fingerprints are pairs of latent and exemplar fingerprints that originate from the same finger. Non-mated fingerprints are pairs of latent and exemplar fingerprints that originate from different fingers. Because the status of a pair of fingerprints is rarely known outside of an experimental context, the examiner must make a conclusion that represents their expert opinion. To do this, an examiner conducts an analysis and comparison of the two prints to make one of three decisions about a pair of fingerprints: identification, exclusion, or inconclusive. An identification decision

means that the examiner believes there is enough perceived detail in agreement between two fingerprints to say the fingerprints came from the same finger. 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.

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 attention to what might be viewed as risk-aversion in latent fingerprint conclusions: examiners are much more likely to make an erroneous exclusion error than an erroneous identification error. This asymmetric error pattern has also been replicated: Tangen, Thompson, and McCarthy (2011) found a 7.8% erroneous exclusion rate and a 0.68% erroneous identification rate. These results reinforces the attention given to erroneous identifications as opposed to erroneous exclusions when considering the risk of different errors.

Ulery et al.'s (2011) study also revealed a large inconclusive rate, raising the possibility that examiners fall back on inconclusive decisions in order to prevent making (potentially serious)

errors. Of course, if examiners make a large number of inconclusive decisions then fewer crimes will be solved and we lose the opportunity to exonerate innocent individuals. This complicates the set of tradeoffs that occur when examiners must decide where to place their decision criteria, which then determines how much evidence is required before making an exclusion or identification conclusion. As we measure the values of the different outcomes of latent print examinations, we might find that the general public is less concerned with innocent people being put in 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.

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, including how easy it is to separate mated from non-mated pairs, as well as the proportion of mated pairs in the comparison set. We need an accurate representation of the quantitative nature of the tradeoffs that occur given these factors. To do this, we modeled the error rate data of Ulery et al. (2011) using signal detection theory (Macmillan and Creelman, 2005).

Modeling the Decision 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

Pair Type	Exclusion	Inconclusive	Identification	Total Mates or Non-mates
Mates	0.075	0.311	0.614	1
Non-mates	0.887	0.111	0.001	1

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 fit a model to the Table 1 data using the following assumptions. First, we assumed that the distribution of evidence scores is normally distributed and allowed the mated distribution to have a different variance than the non-mated distribution. The non-mated distribution is fixed with a mean of zero and a standard deviation of 1.0, which sets the scale of the evidence axis. Four free parameters were then adjusted: the location of the mated distribution along the evidence axis, the standard deviation of the mated distribution, and the two decision criteria (one that separates exclusion from inconclusive responses and the other that separates inconclusive from identification responses).

These parameters were fit to the response proportions shown in Table 1 using maximum likelihood estimation[2]. 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.

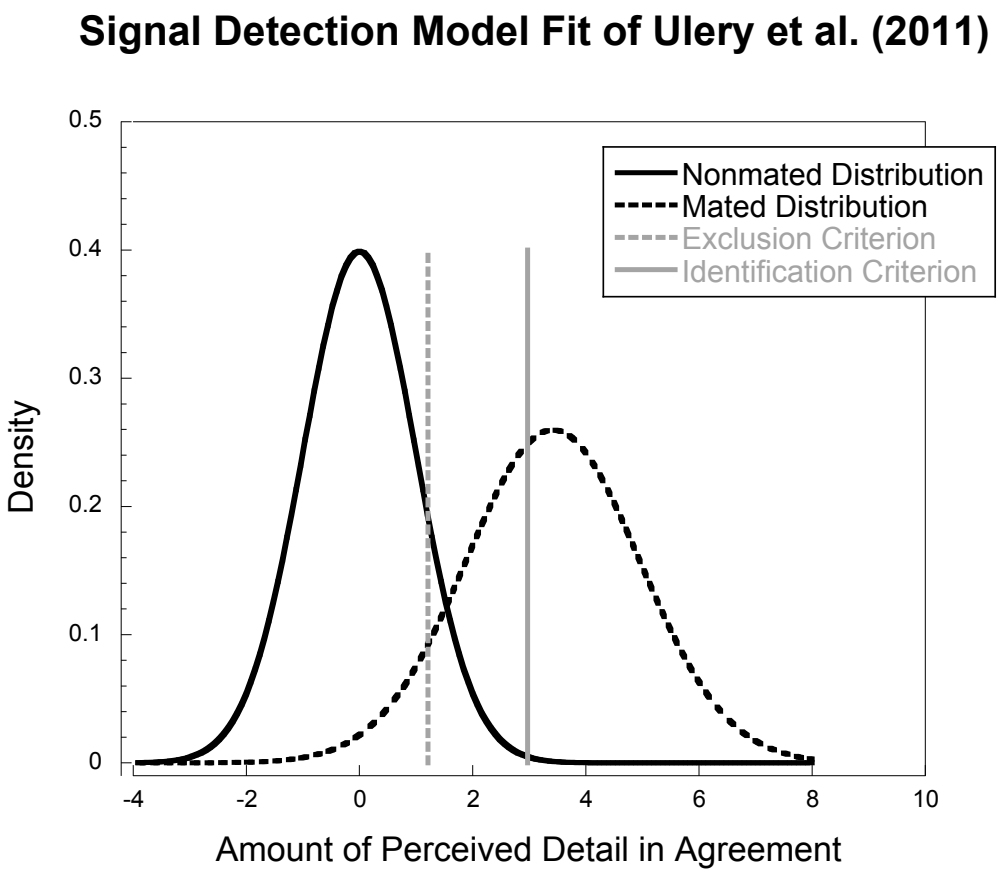
This close correspondence is of course expected, because there are four free parameters and four degrees of freedom in the data. However, we argue that the utility of this model derives from its ability to accurately capture the nature of the decision tradeoff near where estimates were actually derived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

The summary of the Ulery et al. (2011) data by the signal detection theory model allows for the exploration of the consequences of implementing different decision criteria.

[2] Note that typically one would not pool all participants together to produce a single dataset from which signal detection theory would be fit to. If participants differ in either decision threshold or sensitivity (and both are likely to happen), the resulting fits will tend to push the variance from different decision thresholds into the noise distribution, effectively reducing the sensitivity estimates. However, a group-wise fit is actually appropriate in this instance, because we are estimating the behavior of all fingerprint examiners. In any particular case, you do not know which examiner will conduct the comparison, and thus a system-wide estimate is more appropriate. Similarly, when asked to justify a forensic discipline on the stand, an examiner will typically cite error rate studies from the entire discipline, rather than a personal error rate.

Figure 1

Representation of the Ulery et al. (2011) Data Using Signal Detection Theory



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 decision tradeoff in forensics, we developed a web-based visualization (see Fig. 2, Fig. 3, and Method for more detail) based on this model. The visualization quantifies the

decision tradeoff in such a way that we can explore the values expressed by different participants. For example, if a subject is uncomfortable with a particular number of potential innocent people in jail, they can shift the decision criteria toward a more risk-averse value. However, this will simultaneously affect the number of potential 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. We will also collect data from fingerprint examiners, and thus the data we collect allows us to compare experts to members of the general public to determine whether the values expressed by examiners align with the values of society, and both can be compared to error rate studies to determine whether the values correspond to the *actual* thresholds used when conducting comparisons. We also use demographic data to determine the factors that are associated with different resolutions of the decision tradeoff that exists in forensic decision-making to determine the possible sources of expert/novice differences.

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

Table 2

Candidate differences between experts and novices, along with the likely shift in risk aversion. Examiners who are more risk-averse will have fewer erroneous identification outcomes (as well as fewer correct identification outcomes).

Candidate Difference between Examiners and Members of the General Public	**Likely outcome (risk-averse implies requiring more evidence in order to make an identification decision than less risk-averse individuals)**
Examiners understand that fingerprint evidence is not typically the only evidence in a criminal proceeding	Examiners should be **less risk-averse** than members of the public, because errors have less consequence than assumed by the lay public.

Examiners know that most comparisons are verified by a second examiner	Examiners should be **less risk-averse** than members of the public, because errors have less consequence than assumed by the lay public.
Examiners know that not all fingerprint evidence leads to convictions, and that other uses such as victim elimination or non-probative matches are possible	Examiners should be **less risk-averse** than members of the public, because errors have less consequence than assumed by the lay public.
Examiners may differ in important ways demographically that may make them favor a more putative approach to law enforcement	Examiners may be **less risk-averse** than members of the public, because they may be attracted to law enforcement through an emphasis on solving crimes.
Examiners may have special perceptual abilities not shared by novices (e.g. configural processing, better contrast sensitivity for ridge details)	Examiners may be **less risk-averse** than members of the public. This may result from *expertise blind spots* or tunnel vision that restricts information flow and may hide critical information that could lead to overconfidence.
Experts may have a greater understanding of the comparison process.	**Unclear/no effect.** Experts may understand fingerprint comparisons more than our novice participants despite our educational materials. How this difference affects the decision thresholds is uncertain, although demographic details such as contact with the criminal justice system may bear on this question.
Members of the general public may believe fingerprints to be a solved problem and therefore errors are unlikely to occur.	**Unclear/no effect.** Our visualization makes it clear that errors will occur for most threshold settings.
Examiners and the general public may have different scientific or statistical backgrounds	**Unclear/depends on demographics**. Many older examiners may not have a scientific background having come from policing. Younger examiners are more likely to have a hard science or social science bachelor's degree. Demographic analysis of levels of education will clarify any differences and suggest a likely outcome.
Experts have a better grasp of our paradigm and can better connect our visualization to the outcomes	**Unclear/no effect.** This may just introduce noise into the novice data without affecting the central tendency. Tests on the variance of both groups may address this. We also screened out participants based on reaction time.
Experts may be more motivated to think deeply about the tradeoffs	**Unclear/no effect.** This may just introduce noise into the novice data without affecting the central tendency. Tests on the variance of both groups may address this. Some novices were tested in-person to ensure participants took the task seriously.
Experts may encounter close non-mated prints that make them realize the risks involved.	Examiners may be **more risk-averse** than members of the public, and this may be a function of years of experience. More experience allows for more time to encounter 'close calls' (non-mated prints that are very similar in appearance). This may also be a function of the size of the database searched: Modern databases are much larger than previous generations and will reveal more close non-matches.
Examiners may face severe personal and professional consequences for an erroneous identification, but not for an erroneous exclusion	Examiners may be **more risk-averse** than members of the public, because erroneous identifications may result in corrective actions or even loss of employment.

As noted by Dror (2016), there are important expert/novice differences as experts gain skill and knowledge in their discipline. He distinguishes between the information observed and the conclusions made on the basis of this information, and how external contextual information might produce differences both between and within examiners for both observed information and conclusions derived. Whether changes in the decision threshold occur ('biasing') depends on the difficulty of the comparison, the direction of the biasing information, and the strength of the biasing information. Some elements of expertise may protect against such biasing, such as robustness to noise and inversion (Thompson & Tangen, 2014), configural processing of the images (Busey & Vanderkolk, 2005; Vogelsang, Palmeri, & Busey, 2017), However, expertise also brings blind spots: these same abilities such as automaticity, schemas, selective attention, and configural processing may also serve to limit the information acquired by an examiner, which may lead to overconfidence (Dror, 2011; Fisher & Keil, 2015; Hinds, 1999). This might lead an examiner to become less risk-averse.

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

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

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

The second motivation is more personal. Erroneous identifications have the potential to put innocent persons in jail, which creates problems for the incarcerated individual, the victim and society (because the true perpetrator is still at large), the laboratory, and the examiner. However, erroneous exclusions do not garner the same attention, because a criminal who escapes prosecution is not likely to make that known. Thus, there is a strong set of incentives to avoid erroneous identifications at the potential costs of erroneous exclusions or inconclusive conclusions. This, we would argue, is a strong motivator that could make examiners risk-averse when making identification decisions.

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

experts and novices, as well as compare both against the data from error rate studies, to provide the kind of feedback that will allow fingerprint examiners to adjust their decision thresholds to be more in line with the values of the public at large. This approach also allows us to *directly* compare the values of examiners and the general public with the error-rate data, because it places all of the data in the same format (the location of the identification decision threshold along the evidence axis).

Method

Participants

A total of 455 subjects were considered for data analysis. The subjects used for analysis were split into two groups: examiners and novices. There were 377 novice subjects were considered for analysis. 172 were undergraduates attending Indiana University, members of the Bloomington, Indiana community, or members of the surrounding community. The remaining 205 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.

Figure 2

High-mated Web-based Visualization

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.

Web-Based Visualization

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 decision tradeoff: https://buseylab.sitehost.iu.edu/fingerprintvalues/?sandbox.

Figure 3

Low-mated Web-based Visualization

The reader is encouraged to try this visualization in order to understand the dynamic nature of our task, which is difficult to convey in images alone.

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 point clouds represent the mated (top) and non-mated (bottom) pairs of impression, and are jittered vertically to increase the visibility of individual points. During a comparison, the examiner collects evidence regarding the similarity and dissimilarity between two prints and weights each piece of evidence

appropriately and internally to create the overall perceived detail in agreement[3]. 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 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 is are exclusion decisions. The second slider represents the identification criterion—all points to the right of this slider are identification decisions. All points between the two sliders represent an inconclusive decision. The boxes below the graph explain the outcomes of placing the sliders in a specific position and are color coordinated to indicate good outcomes (green), bad outcomes (red), and inconclusive outcomes (yellow). The colors of the clouds match the colors of the boxes in the same positions; for example, the green cloud in the upper right-hand corner of both figures correspond to the outcomes in the green box in the upper right-hand corner. As the sliders are moved, the number of cases in each box changes to reflect the number of points that fall in each of the six outcomes.

---

[3] 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.

Subjective Utilities and Prior Probabilities

The decision tradeoff occurs as the identification or exclusion criterion is shifted along the horizontal axis, because such shifts change both the number of innocent people potentially in jail and the number of guilty people potentially in jail. The exact nature of the decision tradeoff is determined by the model of the underlying mated and non-mated distributions as estimated by signal detection theory. The critical elements for this tradeoff include the amount of overlap between the two distributions (d' in signal detection terminology) as well as the variance of the mated distribution, both of which are provided by the signal detection theory model fit. However, a third factor also affects the decision tradeoff: the ratio of mated and non-mated distributions. This can be thought of as the prior probability of a detective providing a mated impression. Any attempt to estimate the utilities of the various outcomes of a fingerprint comparison will, in principle, depend in part on these priors. The visualization in 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.

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 decision tradeoff.

The subjects were instructed to carefully read all of the outcomes in each box and move the sliders to a position where they were comfortable with the outcomes in the boxes. After the subjects were comfortable with the position of the sliders, they clicked the "Save Values" button below the boxes and proceeded to fill out demographic data. The demographic questions are listed in Appendix B. [note that some questions were not asked of all participants]. Participants in the Mechanical Turk phase of data collection also completed two knowledge/attention check questions, to determine their understanding of the experiment. Ultimately, we decided to include all Mechanical Turk subjects regardless of their score on the knowledge check and instead filtered subjects by reaction time. As this experiment requires critical consideration of one's internal values, we disregarded subjects who took less than 10 seconds to consider the web visualization and did not move the sliders from their starting positions. This filtering process resulted in 9

Mechanical Turk subjects being removed for data analysis. For a complete breakdown of subject inclusion please see Appendix C.

For the sake of succinctness and clarity, the six outcomes represented in the model will be hereby referred to as: potential innocents wrongly identified, potential criminals correctly identified, potential innocents correctly identified, potential criminals wrongly identified, mated inconclusive and non-mated inconclusive. Looking at the web visualizations (Fig. 2 and Fig. 3), the potential innocents wrongly identified outcome is located to the right of the identification slider in the non-mated distribution (bottom right, red cloud), the potential criminals correctly identified outcome is located to the right of the identification slider in the mated distribution (top right, green cloud), the potential innocents correctly identified outcome is located to the left of the exclusion slider in the non-mated distribution (bottom left, green cloud), the potential 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). 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. 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.

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 assessed through the two conditions illustrated in the Fig. 2 and Fig. 3 visualizations that vary the number of mated comparisons?
4) What aspects of the demographic data predict a participant's placement of the decision criteria and the number of innocents in 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 considered in the first analysis were considered, but only 369 subjects had complete demographic data. If the subject was missing data from any of the variables used for the tobit and negative binomial model estimation, they were discarded (a "Decline to Answer" option was included on all demographic questions).

Results of Criteria Placement Comparisons

Tables 3 and 4 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 5 shows the results of performing Bartlett's test (Bartlett, 1937) for population variance. The goal of this test is to determine whether our samples originate from a population with the same variance. We found that there was a significant difference in sample variation of identification criteria placement between novices (M = 2.1879, SD = 1.2932, N = 230) and examiners (M = 2.9472, SD = 0.8058, N = 37) in the high mated condition; $\chi^2(266) = 10.4651$, $p = 0.0012$ and between novices (M = 2.5127, SD = 1.3281, N = 138) and examiners (M = 2.9708, SD = 0.8485, N = 41) in the low mated condition; $\boldsymbol{\chi^2}(178) = 10.4651$, $p = 0.0016$. There was not a significant difference in sample variation of exclusion criteria placement between novices (M = -0.0082, SD = 1.0713, N = 230) and examiners (M= -0.4360, SD = 1.3085, N = 37) in the high mated condition; $\chi^2(266) = 3.0502$, $p = 0.0807$ or between novices (M = 0.2146, SD = 0.9573, N = 138) and examiners (M= 0.2143, SD = 1.1964, N = 41) in the low mated condition; $\chi^2(178) = 3.5642$, $p = 0.059$.

Table 6 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 = 1.2932, N = 230) and examiners (M = 2.9472, SD = 0.8058, N = 37) in the high mated condition; $t(265) = 4.7058$, $p =$ 1.2546e-5, d = 0.6052 and between novices (M = 2.5127, SD = 1.3281, N = 138) and examiners (M = 2.9708, SD = 0.8485, N = 41) in the low mated condition; $t(177) = 2.4163$, $p = 0.0174$, d = 0.3435. There was not a significant difference in exclusion criterion placement between novices

(M = -0.0082, SD = 1.0713, N = 230) and

Table 3

Placement of Identification Criteria

		High Mated	Low Mated
Examiner	Mean	2.9472	2.9708
	Standard Deviation	0.8058	0.8485
	Sample Size	37	41
Novice	Mean	2.1879	2.5127
	Standard Deviation	1.2932	1.3281
	Sample Size	230	138

Table 4

Placement of Exclusion Criteria

		High Mated	Low Mated
Examiner	Mean	-0.4360	0.2143
	Standard Deviation	1.3085	1.1964
	Sample Size	37	41
Novice	Mean	-0.0082	0.2146
	Standard Deviation	1.0713	0.9573
	Sample Size	230	138

Table 5

Results of Bartlett's test for population variance

Exclusion Criteria					
	Examiner Mean	**Novice Mean**	χ^2	**dF**	**p**
High Mated	-0.4360	-0.0082	3.0502	266	0.0807
Low Mated	0.2143	0.2146	3.5642	178	0.059
Identification Criteria					
High Mated	2.9472	2.1879	10.4651	266	0.0012
Low Mated	2.9708	2.5127	9.9751	178	0.0016

Table 6

Results of t-test comparison

Exclusion Criteria						
	Examiner Mean	**Novice Mean**	**t**	**dF**	**p**	**d**
High Mated	-0.4360	-0.0082	-1.8649	265	0.0688	-0.3864
Low Mated	0.2143	0.2146	-0.0842	177	0.9332	-.0171
Identification Criteria						
High Mated	2.9472	2.1879	4.7058	265	1.2546e-5	0.6052
Low Mated	2.9708	2.5127	2.4163	177	0.0174	0.3435

examiners (M= -0.4360, SD = 1.3085, N = 37) in the high mated condition; t (265) = -1.8649 , p = 0.0688, d = -0.3864 or between novices (M = 0.2146, SD = 0.9573, N = 138) and examiners (M= 0.2143, SD = 1.1964, N = 41) in the low mated condition; t(177) = -0.0842, p = 0.9332, d =-0.0171.

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 potential 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.

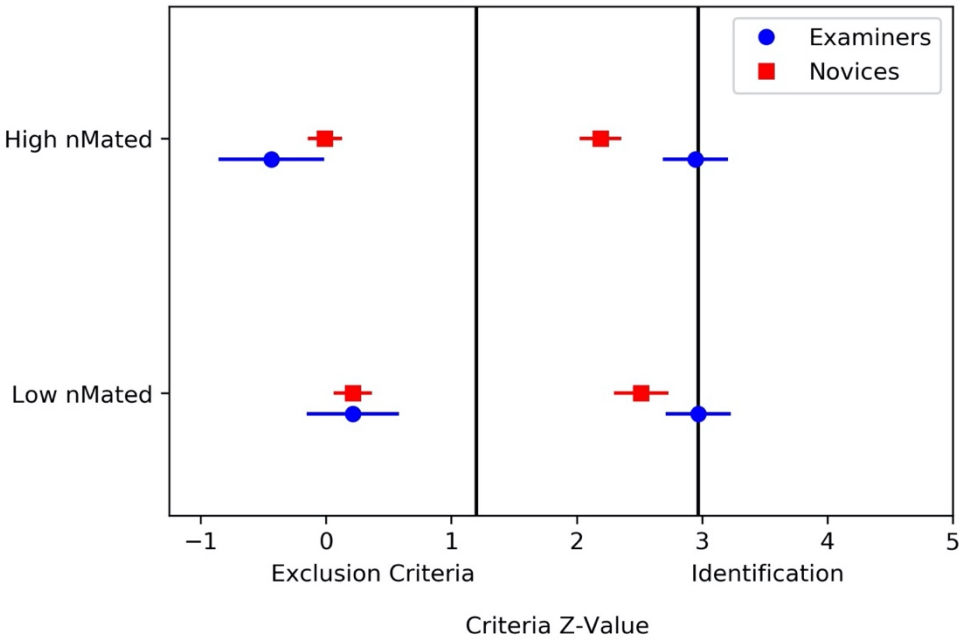
Surprisingly (at least to us), examiners in both conditions place their identification decision criterion on almost exactly the same location as was revealed by the Ulery et al (2011) error-rate study. Thus, examiners behave and believe in almost exactly the same risk-averse manner. Novices, however, are much less risk-averse than the threshold revealed by the error-rate study. Note that both experts and novices differed greatly from the error-rate exclusion threshold. Both groups were much more risk-averse with respect to exclusion errors than was revealed by the error-rate data. When it comes to exclusions, it may be easier to make a decision in practice than to consider the abstract possibility of an erroneous exclusion. However, these differences are perhaps less compelling than the identification threshold results, because there may be little practical difference between an exclusion and an inconclusive decision.

Associations between Demographic Data and Criteria Placement

The resolution of the decision tradeoff involves a set of personal beliefs about the role of justice in society. For example, a 'law-and-order' politician may stress the need for locking up criminals, at the possible risk of incarcerating innocent individuals as well. Organizations such as the Innocence Project may have concerns about the rights of wrongfully-convicted individuals and might argue for a more conservative decision 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)?

Figure 4

Average Criteria Placement



Note. 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 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.

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 potential 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.

Summary Statistics of Demographic Data

Table 7 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 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 potential 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 potential innocents (24.12 percent of the sample as shown in Table 7). 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.

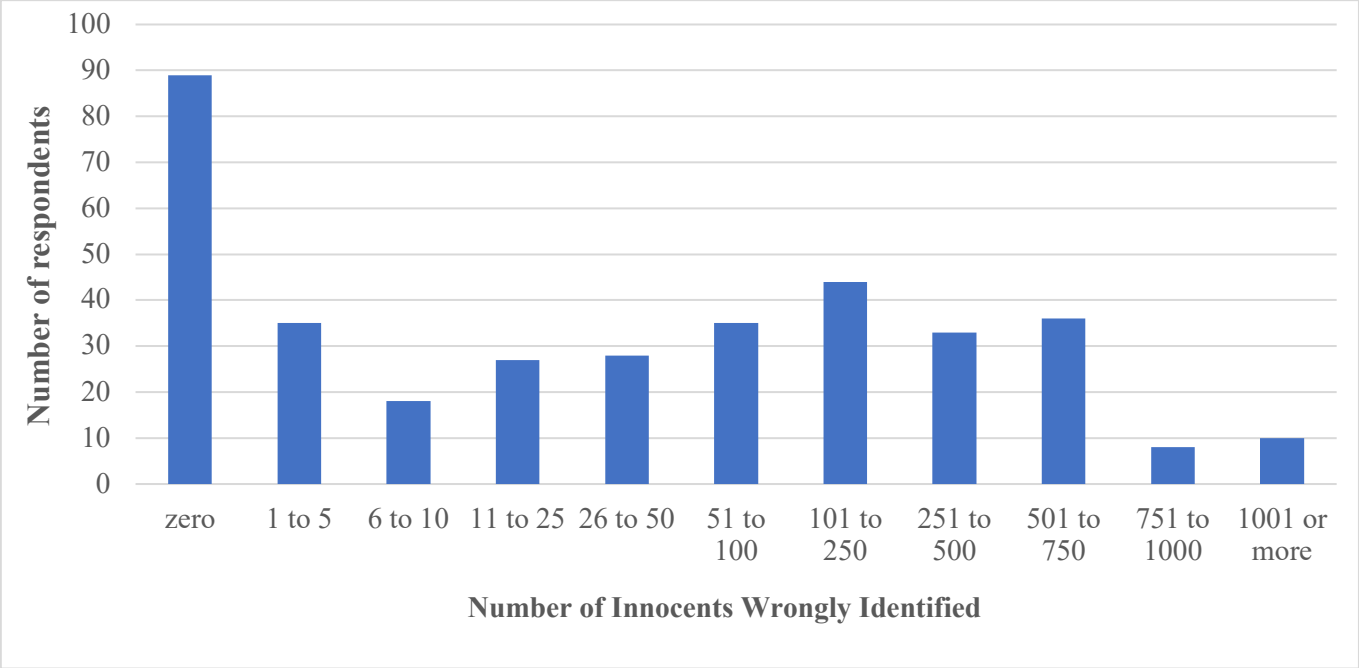
Table 7

Summary statistics of survey responses

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

Figure 5

Number of subjects wrongly identifying innocents by innocent-frequency category



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 potential innocents wrongly identified to potential criminals correctly identified. This ratio is representative of the position of the identification criterion, as the potential innocents wrongly identified and potential 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 potential innocents wrongly identified to potential criminals correctly identified. This ratio will have a minimum of zero (when no potential innocents are wrongly identified) and will be a continuous

variable when one or more potential 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 potential criminals wrongly identified and potential 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),

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

where Y^*_{n1} (with k=1) is latent variable of the ratio of potential innocents wrongfully identified to potential criminals correctly identified for subject n and is observed only when positive, Y^*_{n2} (with k=2) is latent variable of the ratio potential criminals wrongly identified to potential 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 potential innocents wrongfully identified to potential criminals correctly identified dependent variable, Y_{n2} is the observed ratio of potential criminals wrongly fully identified and potential innocents correctly identified dependent variable, $\mathbf{X}_{n1}$ and $\mathbf{X}_{n2}$ are

vector of explanatory variables corresponding to ratio equations, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors of estimable parameters, and ε_{n1} and ε_{n2} are bivariate normally and independently distributed error terms with zero means, variances $\sigma^2_{\varepsilon_{n1}}$ and $\sigma^2_{\varepsilon_{n2}}$, correlation ρ. This tobit model can be readily estimated by standard maximum likelihood methods (Washington et al., 2011).

Bivariate Tobit Model Estimation Results

The bivariate tobit model estimation results are presented in Table 8. 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 $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors.

Turning first to the estimation results for the ratio of potential innocents wrongfully identified to potential criminals correctly identified, the fingerprint examiner indicator resulted in a negative parameter indicating a smaller ratio (fewer potential innocents wrongly identified per potential 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 potential innocents wrongly identified per potential 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 potential innocents wrongly identified per potential 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 (potential 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 potential criminals wrongly identified to potential innocents correctly identified), Table 8 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 potential criminals wrongly identified to potential innocents correctly identified cannot be predicted with our collected socio-demographic information, even though the ratio of potential innocents wrongfully identified to potential 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.

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 potential innocents wrongly identified. If a subject was an older male (over 65 years) they tended to be less conservative, or more likely to tolerate potential innocents wrongly identified. Finally, as more children are added to a household, subjects become less conservative, thus tolerating more potential 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 potential innocents incorrectly identified as guilty.

Table 8

Bivariate Tobit Model of Ratio of Innocents Wrongfully Identified to Criminals Correctly Identified and of the Criminals Wrongly Identified to Innocents Correctly Identified.

Variable Description	Estimated Parameter	*t* statistic[a]
Ratio Innocents Wrongfully Identified to Criminals Correctly Identified		
Constant	0.218	5.12***
Fingerprint examiner indicator (1 if respondent is a fingerprint examiner, 0 otherwise)	–0.160	–3.62***
Number of children in respondent's household if not a fingerprint examiner	0.029	1.86*
Older male indicator (1 respondent is a male 65 years old or older, 0 otherwise)	0.217	4.15***
White indicator (1 if respondent identified themselves as being white, 0 otherwise)	–0.114	–2.79***
High criminals indicator (1 if respond faces a criminal cluster of 5969 criminals, 0 if 1000 criminals)	–0.165	–3.42***
Ratio of Criminals Wrongly Identified to Innocents Correctly Identified		
Constant	0.00051	0.02
Fingerprint examiner indicator (1 if respondent is a fingerprint examiner, 0 otherwise)	0.00364	0.35
Number of children in respondent's household if not a fingerprint examiner	0.00187	0.37
Older male indicator (1 respondent is a male 65 years old or older, 0 otherwise)	0.01247	0.47
White indicator (1 if respondent identified themselves as being white, 0 otherwise)	0.00615	0.49
High criminals indicator (1 if respond faces a criminal cluster of 5969 criminals, 0 if 1000 criminals)	0.05013	2.80***
Correlation coefficient between equations, $\rho_{\varepsilon_{n1}\varepsilon_{n2}}$	–0.121	–0.87
Number of observations	369	

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

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 potential 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,

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

where λ_n is the Poisson parameter for subject n, which is the subjects expected number of erroneous identification errors. To incorporate explanatory variables, Poisson regression specifies the Poisson parameter λ_n as a log-linear function, $\lambda_n = EXP(\boldsymbol{\beta}\mathbf{X}_n)$, where $\mathbf{X}_n$ is a vector of explanatory variables that determine the number of potential innocents wrongly identified by subject n and $\boldsymbol{\beta}$ 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 potential innocents wrongly identified as, $\lambda_n = EXP(\boldsymbol{\beta}\mathbf{X}_n + \varepsilon_n)$, where $EXP(\varepsilon_n)$

is a gamma-distributed error term with mean 1 and variance α. The addition of this gamma-distributed results in the negative binomial regression (also sometimes referred to as the Poisson-Gamma regression since a gamma function is added to account for overdispersion) and allows the variance to differ from the mean as $VAR[i_n] = E[i_n][1+ \alpha E[i_n]] = E[i_n]+ \alpha E[i_n]^2$. The addition of this gamma term results in a negative binomial regression (also sometimes referred to as the Poisson-Gamma regression since a gamma function is added to account for overdispersion). The Poisson regression is a limiting model of this negative binomial regression as α approaches zero. Thus, if α is determined during estimation to be significantly different from zero, the negative binomial is appropriate and, if it is not, the Poisson model is appropriate because with α equal zero the negative binomial reduces to the Poisson model. Both Poisson and negative binomial models can be readily estimated with standard maximum likelihood methods (Washington et al., 2011). However, for the case of potential 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 potential innocents should never be wrongly identified, which would put these subjects in a zero-count state. Other subjects may be willing to wrongly identify potential 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). Details of this model are presented in the Appendix D.

Finally, to assess the effects of individual explanatory variables in the $\mathbf{X}_n$ vector, on the mean number of potential innocents (λ_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 $\mathbf{X}_n$ vector) has the mean number of potential innocents wrongly identified and is computed as,

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

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

Zero-Inflated Negative Binomial Model Estimation Results

Table 9 presents the summary statistics for variables found to be statistically significant in the zero-inflated negative binomial estimation, and Table 10 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 potential innocents wrongly identified). The value of the Vuong statistic (see Appendix D for description) shown in Table 10 is 3.30, which indicates more than 99% certainty that the two-state zero-inflated model is preferred relative to a single state model. The negative binomial dispersion parameter of 2.373 with a t-statistic of 27.64 is significantly different from zero suggesting that the negative binomial is statistically preferred over the Poisson. Thus, these statistics strongly support the zero-inflated negative binomial model relative to the non-zero-inflated Poisson and negative binomial models, and the zero-inflated

Poisson model. Also, as further evidence in addition to the Vuong test, a simple likelihood ratio test comparing the log-likelihood at convergence for the negative binomial (-1944.97) with the log-likelihood at convergence for the zero-inflated negative binomial (-1922.82) gives a χ^2 statistic of 44.30 [-2(-1944.97-(-1922.82))] with 4 degrees of freedom, which implies more than 99.99% confidence that the simple negative binomial model and the zero-inflated negative binomial model are not equal.

Table 9

Summary statistics for negative binomial variables

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

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 potential criminals presented to subjects was 1000 for some and 5969 for others. This suggests the possibility that subjects' "innocent" decisions may be influenced by the number of potential criminals presented in their experiment. To test for this, the sample was split in two; those subjects facing 1000 potential criminals and those facing 5969 potential 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 8), $LL(\beta_{1000})$ is the log-likelihood at convergence of the model using only data from subjects facing 1000 potential criminals, and $LL(\beta_{5696})$ is the log-likelihood at convergence of the model using only data from subjects facing 5969 potential criminals. In this test the same variables are used in all three models and this X^2 test statistic is χ^2 distributed with degrees of freedom equal to the summation of the number of estimated parameters in the "1000" and "5969" models minus the number of estimated parameters in the "all" model. The test statistic indicates that there is only 18% confidence that the separate "1000" and "5969" models are statistically different from the "all" model, so this justifies the 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 10. 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 10 (where a positive parameter estimate increases the likelihood a subject will be in the zero sate 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 potential 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 possibly other factors, could be playing a role.

Table 10 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 10 suggest that the net effect is that individuals less than 25 years old were, on average, willing to wrongly identify 37.90 more potential innocents than other age groups.

Finally, for the zero-state probabilities, Table 10 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 potential innocents).

Turning to variables found to be statistically significant in the count state (where a positive parameter estimate increases the number of potential 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 potential innocent. Marginal effects in Table 10 show that subjects in this lower-income bracket were, on average, willing to wrongly identify 82.14 more potential innocents than their higher-income counterparts.

In addition to lower income subjects, Table 10 shows males 65 years old or older were also more willing to wrongly identify potential innocents, with marginal effects showing on average willing to wrongly identify 191.55 more potential 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 potential innocents in their quest for correctly identifying as many potential criminals as possible.

Table 10

Zero-Inflated Negative Binomial of the Number of Innocents Wrongly Identified.

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

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

Interestingly, subjects who identified themselves as white (77.8% of the sample as indicated in Table 9), were significantly less likely to wrongly identify potential innocents relative to other ethnicities in the sample (marginal effects in Table 10 show they were, on average, willing to wrongly identify 126.93 fewer potential innocents).

Finally, having more children not only made subjects less likely to be in the in the zero-count state, but also made them significantly more likely to wrongly identify potential 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 on average 68.02 more potential innocents.

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 potential innocents the subject was willing to wrongly identify. Still, the net effect of the zero and count-state marginal effects show (Table 10) that fingerprint examiners are willing to wrongly convict 33.10 fewer potential 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).

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

Discussion

The results of this study support the idea that the general public are 1) less biased against erroneous identification errors than examiners and 2) are less tolerant of a large number of inconclusive decisions. According to the results shown in Table 6, there is a significant difference between the novice and examiner placement of the identification criterion. As shown in Table 6, 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 potential 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 potential 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 true. As the examiners were significantly more risk-averse 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.

Our finding that examiners were more risk-averse than the members of the general public rules out many of the possible differences between groups as posited in Table 2. In addition, years of experience was not significant, although it is correlated with age, and that multi-collinearity makes it difficult to tease apart age and years of experience. Thus, it is possible that older examiners are more risk-averse.

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

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 decision tradeoff. This allows examiners to communicate several key points to managers. First, no forensic decision-making process can be free of errors. Therefore, we argue that one response is to place additional safeguards in place but to treat errors as learning opportunities rather than responding 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.

These issues are complex, and one reason that examiners are placed in this position is that they actually make *decisions.* They are essentially taking the strength of evidence in a particular comparison and distilling that down into one of three conclusions. They are implicitly including the prior probability of a mated pair into this calculation, despite the fact that this is typically unknown. An alternative would be to shift to a strength-of-evidence report rather than a definitive conclusion, as is done in Europe. A middle ground would be to expand the set of conclusions that examiners use, from three to five, to allow for investigative leads. Some agencies such as the Houston Forensic Science Center already allow for this for some comparisons. Each of these policy changes should be made with the support of empirical studies that demonstrate how best to communicate the strength of evidence in a particular case to an eventual consumer such as a detective, prosecutor, or defense attorney.

References

Anastasopoulos, P., Shankar, V., Haddock, J., & Mannering, F. (2012). A multivariate tobit analysis of highway accident-injury-severity rates. *Accident Analysis and Prevention* 45(1), 110-119.

Ashbaugh, D. (1999). Quantitative-qualitative friction ridge analysis: an introduction to basic and advanced ridgeology. Boca Raton, Fla.: CRC Press.

Bartlett, M. S. (1937). "Properties of sufficiency and statistical tests". *Proceedings of the Royal Statistical Society*, Series A 160, 268–282

Busey, T. A., & Vanderkolk, J. R. (2005). Behavioral and electrophysiological evidence for configural processing in fingerprint experts. *Vision Res, 45*(4), 431-448. doi:S0042-6989(04)00436-5 [pii] 10.1016/j.visres.2004.08.021

"Commentaries on the laws of England". (1893). J.B. Lippincott Co., Philadelphia.

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993)

Dror, I. (2011). The paradox of human expertise: Why experts get it wrong. In A. Pascual-Leone, V. Ramachandran, J. Cole, S. Della Sala, T. Manly, A. Mayes, et al. (Authors) & N. Kapur (Ed.), The Paradoxical Brain (pp. 177-188). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511978098.011

Expert Working Group on Human Factors in Latent Print Analysis. (2012). Latent print examination and human factors: improving the practice through a systems approach: the report of the Expert Working Group on Human Factors in Latent Print Analysis. Washington, D.C.: NIST NIJ, National Institute of Justice.

Fiske, A., & Tetlock, P. (1997). Taboo tradeoffs: Reactions to transactions that transgress the spheres of justice. *Political Psychology* 18(2), 255-297.

Fisher, Matthew & Keil, Frank. (2015). The Curse of Expertise: When More Knowledge Leads to

Miscalibrated Explanatory Insight. Cognitive Science. 40. 10.1111/cogs.12280.

Hinds, P.J. (1999). The curse of expertise : The effects of expertise and debiasing methods on predictions of novice performance.

Hostetter, A. (2014). Action attenuates the effect of visibility on gesture rates. *Cognitive Science*, 38, 1468-1481.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34(1), 1-14.

Macmillan, N., & Creelman, C. (2005). Detection Theory: A User's Guide. Lawrence Erlbaum Associates, Publishers. Mahwah, New Jersey.

Moses Maimonides, *The Commandments, Neg. Comm. 290*, at 269-271 (Charles B. Chavel trans., 1967).

Malyshkina, N., & Mannering, F. (2010). Zero-state Markov switching count-data models: An empirical assessment. *Accident Analysis and Prevention* 42(1), 122-130.

Mannering, F., Shankar, V., & Bhat, C. (2016). Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research* 11, 1-16.

National Institute of Justice (2011) *Fingerprint Sourcebook*. (https://www.ncjrs.gov/pdffiles1/nij/225320.pdf)

Snodgrass, M., Bernat, E., & Shevrin, H. (2004). Unconscious perception at the objective detection threshold exists. *Perception and Psychophysics, 66*(5), 888-895.

Tangen J.M., Thompson M.B., McCarthy D.J. (2011) Identifying fingerprint expertise. *Psychological Science,* 22(8), 995–7.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* 26(1), 24-36.

Ulery, B., Hicklin, R., Buscaglia, J., & Roberts, M. (2011). Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences* 108(19), 7733-7738.

Vanselst, M., & Merikle, P. (1993). Perception below the objective threshold. *Consciousness and Cognition, 2*(3), 194-203.

Vogelsang, M. D., Palmeri, T. J., & Busey, T. A. (2017). Holistic processing of fingerprints by expert forensic examiners. *Cogn Res Princ Implic, 2*(1), 15. doi:10.1186/s41235-017-0051-x

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57(2), 307-334.

Washington, S., Karlaftis, M., & Mannering, F. (2011). Statistical and econometric methods for transportation data analysis. Chapman and Hall/CRC, Boca Raton, FL, Second Edition.

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 theses 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

Demographic Information Requested	Possible Answers
Age	18-24, 25-34, 35-44, 45-54, 55-65, 65-74, 75 and older, Decline to Answer
Gender Identity	Male, Female, Decline to Answer
Ethnicity origin (or Race)	White, Hispanic or Latino, Black or African American, Native American or American Indian, Asian/Pacific Islander, Other, Decline to Answer
Level of Education	High School Diploma, Currently in college, Some college or Associates degree, Bachelor's Degree, Master's Degree, Professional Degree, Doctorate Degree, Decline to Answer
Experience with fingerprint examinations	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
Association with the justice system	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
Personal interactions with the justice system (check all that apply)	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

	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

User ID	User Category	Number of Subjects Considered for Analysis	Number of Subjects Disqualified for Data Analysis
201	examiners, random mated condition	1	0
202	examiners, random mated condition	1	0
204	examiners, random mated condition	1	0
206	examiners, random mated condition	1	0
208	examiners, random mated condition	1	0
211	examiners, random mated condition	1	0
214	examiners, random mated condition	1	0
218	examiners, random mated condition	1	0
219	examiners, random mated condition	1	0
220	examiners, random mated condition	1	0
313	Mechanical Turk	205	9
400	novice, random mated condition	36	0
425	novice, high mated condition	68	0
500	examiners, random high or low	39	0
600	novice, high mated condition	1	0
700	novice, high mated condition	25	0
888	examiners, high mated condition	10	0
889	examiners, low mated condition	10	0
900	novice, early random mated condition	42	0
950	examiners	5	0
999	examiners	4	0

Appendix D: Zero-Inflated Count-Data Models

Zero-inflated count data models allow for a zero-state (where the count is zero) and a count state where the count is a non-negative integer. Poisson and negative binomial models are commonly used models for count data and the zero-inflated Poisson model is set up as (with i_n being innocents wrongly identified),

$$\begin{aligned} i_n &= 0 \text{ with probability } p_n + (1-p_n)\,EXP(-\lambda_n) \\ i_n &= i \text{ with probability } \frac{(1-p_n)\,EXP(-\lambda_n)\lambda_n^{i_n}}{i_n!} \end{aligned} \tag{D.1}$$

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 $[(1\text{-}p_n)EXP(-\lambda_n)]$ 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,

$$\begin{aligned} i_n &= 0 \text{ with probability } p_n + (1-p_n)\left[\frac{1/\alpha}{(1/\alpha)+\lambda_i}\right]^{1/\alpha} \\ i_n &= i \text{ with probability } (1-p_n)\left[\frac{\Gamma\left((1/\alpha)+i_n\right)\,u_n^{1/\alpha}(1-u_n)^{i_n}}{\Gamma(1/\alpha)\,i_n!}\right], \end{aligned} \tag{D.2}$$

where $\mu_i = (1/\alpha)/[(1/\alpha)+\lambda_i]$. 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 D.1 and D.2) is typically assumed to follow a probit (normal) probability process, where the probability of being in the zero state (p_n) is:

$$p_n = \Phi(\boldsymbol{\beta}_z \mathbf{X}_z) \tag{D.3}$$

where $\Phi(.)$ is the normal cumulative distribution function, $\mathbf{X}_z$ is a vector of characteristics that determine the probability of being in the zero state and $\boldsymbol{\beta}_z$ 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),

$$m_n = LN\left(\frac{f_1\left(i_n \mid \mathbf{X}_n\right)}{f_2\left(i_n \mid \mathbf{X}_n\right)}\right) \tag{D.4}$$

where $f_1(i_n|\mathbf{X}_n)$ is the probability density function of model 1, and $f_2(i_n|\mathbf{X}_n)$ 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),

$$V = \frac{\sqrt{N}\left[\left(1/N\right)\sum_{n=1}^{N} m_n\right]}{\sqrt{\left(1/N\right)\sum_{n=1}^{N}\left(m_n - \overline{m}\right)^2}} = \frac{\sqrt{N}\left(\overline{m}\right)}{S_m} \tag{D.5}$$

where $\overline{m}$ is the mean $1/N\sum_{n=1}^{N} m_n$ and S_m is standard deviation, and N is the sample size. This Vuong statistic is asymptotically standard normal distributed (to be compared to z-values). Using a 97.5% confidence level, if 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).