Journal of Data Science 22 (2), 314–332
July 2024
DOI: 10.6339/24-JDS1130
Education in Data Science

Demonstrative Evidence and the Use of Algorithms in Jury Trials

Rachel Rogers[1,*] and Susan VanderPlas[1]

[1]*Department of Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska, United States of America*

Abstract

We investigate how the use of bullet comparison algorithms and demonstrative evidence may affect juror perceptions of reliability, credibility, and understanding of expert witnesses and presented evidence. The use of statistical methods in forensic science is motivated by a lack of scientific validity and error rate issues present in many forensic analysis methods. We explore what our study says about how this type of forensic evidence is perceived in the courtroom – where individuals unfamiliar with advanced statistical methods are asked to evaluate results in order to assess guilt. In the course of our initial study, we found that individuals overwhelmingly provided high Likert scale ratings in reliability, credibility, and scientificity regardless of experimental condition. This discovery of scale compression - where responses are limited to a few values on a larger scale, despite experimental manipulations - limits statistical modeling but provides opportunities for new experimental manipulations which may improve future studies in this area.

Keywords *explainable machine learning; jury perception*

1 Introduction

The prevailing belief in bullet comparison in the forensic sciences is that guns can leave individualizing marks on bullets when fired, which can be used to identify the gun (President's Council of Advisors on Science and Technology, 2016). Current bullet matching methods rely on a subjective visual comparison of bullets completed by a forensic scientist in order to reach a conclusion (National Research Council (US), 2009). In order to improve upon the bullet matching method with increased scientific validity, President's Council of Advisors on Science and Technology (2016) urged the development of objective methods of analysis. These reports have spurred increased research, development, and assessment of statistical matching methods for firearms analysis, including (Hare et al., 2017; Vanderplas et al., 2020; Song et al., 2012).

As these algorithms are developed and validated, it becomes more important to understand how they may impact the evidentiary process - how will jurors react to algorithms used to match bullets?

In this factorial study, we examine the effect of algorithm use and demonstrative evidence (photos and data visualizations) in jurors' perception of examiner testimony. We assess the perception of the strength of evidence, guilt or innocence, examiner credibility, and the reliability and scientific validity of firearms examination. This study is intended to lay the groundwork for the use of algorithmic firearms comparisons in court. When presented with the same evidence,

*Corresponding author. Email: rachel.rogers@huskers.unl.edu.

Received July 31, 2023; Accepted April 6, 2024

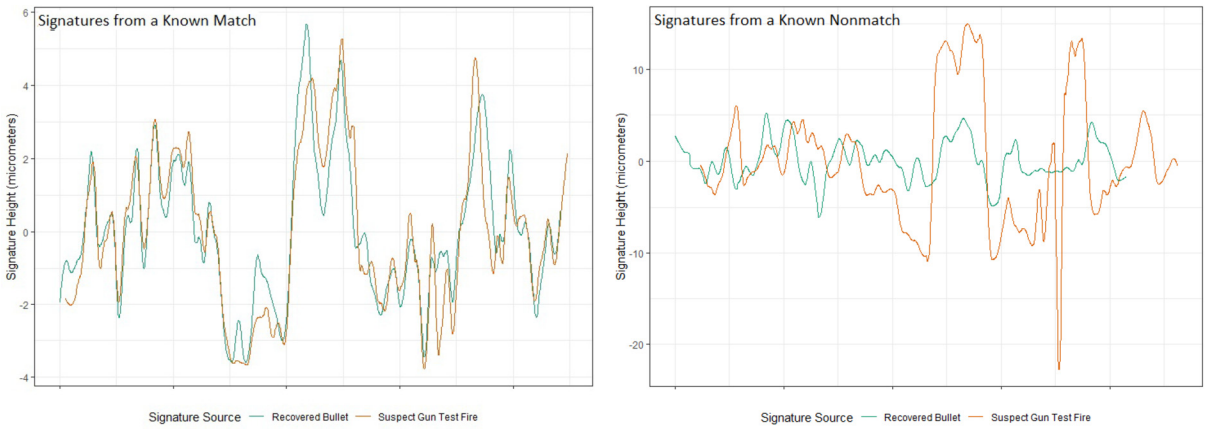

Figure 1: Bullet signatures for two lands. The left image indicates two matching lands, while the right image indicates two non-matching lands.

it is important to know whether or not images affect potential jurors' views on the reliability or credibility of the witness and the evidence that they present.

1.1 Bullet Matching Algorithm

The motivating idea of bullet matching is that there are unique scratches in a gun barrel resulting from the rifling process that make it possible to identify whether two bullets were fired from the same gun (National Research Council (US), 2009). This barrel rifling creates a spiral of raised areas, known as lands, and indented areas, known as grooves; these lands leave scratch marks on the bullet that can then be compared using an algorithm (Hare et al., 2017). The transition period between lands and grooves of the bullet are known as shoulders (Hare et al., 2017).

Firearms examiners visually examine the lands of bullets using a comparison microscope, which allows two bullets to be directly compared in the same viewfinder. If there is sufficient similarity (which is a subjective conclusion), the examiner will make an **identification**, suggesting that the two bullets are from the same source (SS) - they were fired from the same gun. If there is sufficient dissimilarity (again, a subjective decision), the examiner will make an **elimination**, concluding that the bullets were not fired from the same gun. If there is insufficient evidence in either direction (at least in theory, see (Hofmann et al., 2021)), then the examiner will return an **inconclusive** decision. The bullet matching algorithm developed by Hare et al. (2017) follows these steps: first, a 3D scan is taken of each bullet land, a stable cross-section is extracted, and shoulders (edges) are removed; then a smoothing function is applied twice in order to extract a representative profile, called the signature, which can be compared to signatures from other bullets (as shown in Figure 1); finally, traits derived from the signature are combined using a random forest classifier to produce a match score for each land, ranging in value from 0 to 1, where 0 is indicative of a different-source pair and 1 is indicative of a same-source pair. There are multiple lands per bullet, resulting in a grid of land-to-land match scores when comparing two bullets, as shown in Figure 2. Lands are ordered sequentially, so that bullets from the same source should produce high land-to-land match scores as the land number progresses (Vanderplas et al., 2020). For example, in the top left grid of Figure 2, Land 5 (L5) of B1 corresponds to Land 1 (L1) of B2, as can be seen by the higher random forest score. As the bullets are rotated, the lands continue to correspond (Land 6 of B1 to Land 2 of B2, and so forth) in a diagonal -

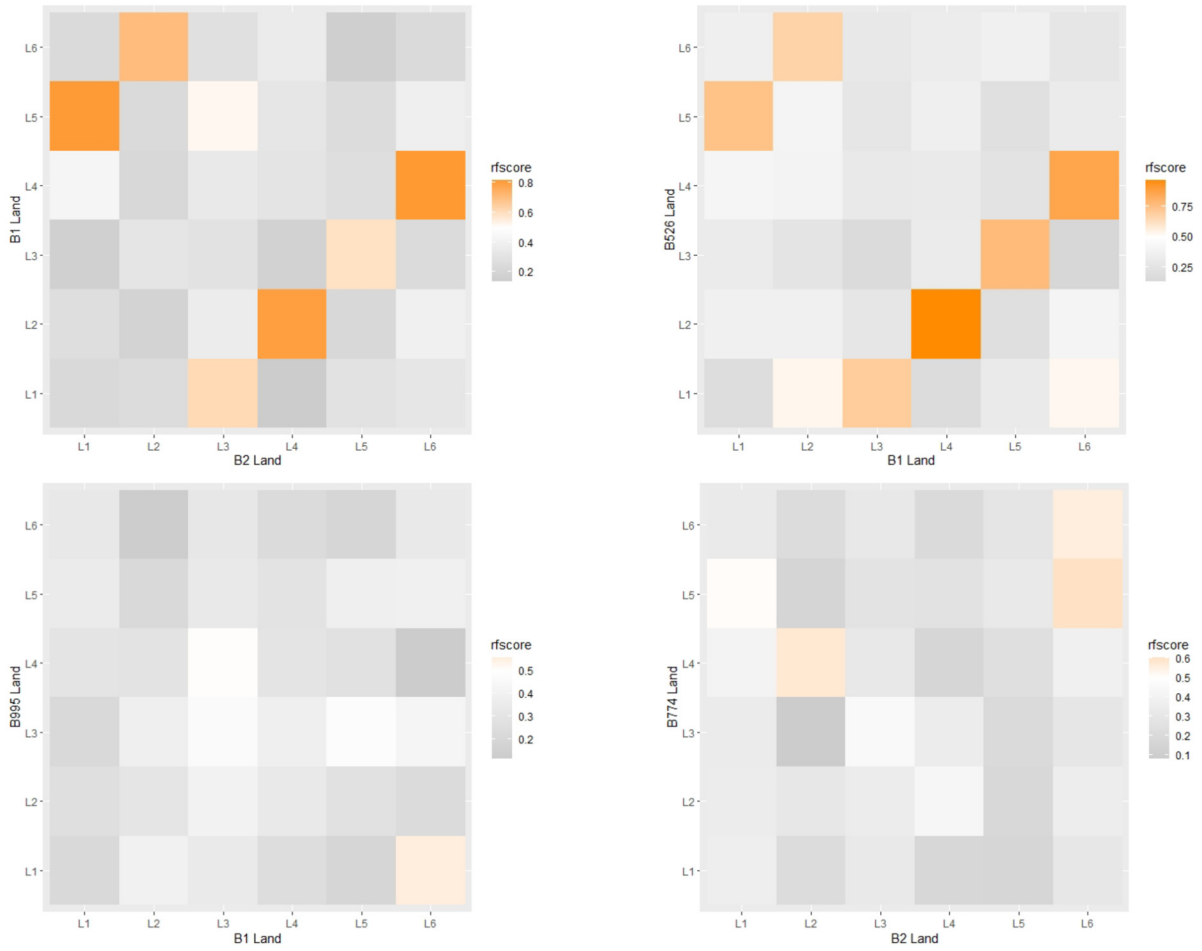

Figure 2: Comparison grids demonstrating various bullet comparisons. The top two images were from the same source, and were used as the test fire (left) and the algorithmic identification (right) in the sample testimony. The bottom two images are from different sources, and were used as the algorithmic elimination (left) and inconclusive (right) in the sample testimony.

indicating that the bullets match.

These individual scores for the maximum correspondence between the two bullets are averaged to create a bullet-level match score (Vanderplas et al., 2020). The match scores reported in the sample testimony for the corresponding image in Figure 2 are shown in Table 1. Vanderplas et al. (2020) validated the algorithm for use with guns outside models used in the training data. This is a critical step required before the algorithm could be used in forensic labs. For the algorithm to be used in practice, though, we must understand how jurors without statistical expertise understand and interpret the results from the algorithm when these results are presented during testimony. This problem of jury interpretation of statistical results has been encountered before, in disciplines such as DNA and fingerprints (Koehler, 2001; Garrett et al., 2018).

1.2 Explainable Machine Learning in Courts

Swofford and Champod (2022) interviewed judges, lawyers, scientists, and researchers in order to assess their feelings about the use of statistical methods and probabilistic language in court. They

Table 1: Match score and examiner conclusion language used for algorithm evidence in the sample testimony.

Bullet Comparison	Match Score	Language
Test Fire	0.976	
Identification	0.989	The match score indicates that there is substantial similarity between the two bullets, which suggests that they were most likely fired from the same barrel.
Elimination	0.034	The match score indicates that there is significant disagreement between the two bullets.
Inconclusive	0.34	The match score indicates that there is not sufficient agreement between the two bullets, which suggests that the results are inconclusive.

found that some individuals were concerned about jurors' ability to understand and properly interpret probabilistic language; some suggested that a mix of probabilistic language and match language may be more beneficial than strictly using one or the other.

FRStat (Friction Ridge Statistics), an analysis program which assigns statistical values to fingerprint analyses, uses likelihood ratios to characterize strength of evidence, with phrasing like "The probability of observing this amount of correspondence is approximately [X] times greater when the impressions are made by the same source rather than by different sources" (Swofford, 2017). Garrett et al. (2018) found that jurors did not provide significantly different likelihoods that the individual was the source of the prints when they were presented with a wide range of FRStat likelihood results, ranging from values of 10 times greater to 100,000 times greater.

In another study, Koehler (2001) investigated the perception of probabilities and frequencies in the case of DNA. They found that individuals were more likely to believe the subject was the source of the DNA when the same number was presented as a probability rather than a frequency. They also asked individuals to identify the number of people that would match DNA for their given match proportion in a population of 500,000; 60.7% of those given a frequency and 42.1% of those given a probability were able to correctly identify the number. These examples illustrate that there is (justifiable) concern for how statistical methods and results may be presented and interpreted in the courtroom. In the application of Hare et al. (2017)'s algorithm, the correspondence between bullets is described as a match score produced by a random forest, with values between 0 and 1. This machine learning application may add another hurdle in juror understanding.

1.3 Demonstrative Evidence

Demonstrative evidence, such as images, can serve as an aid in explaining results and methods used in the forensic sciences. However, there is the potential that the use of images can be biasing. In a study conducted by Cardwell et al. (2016), researchers found that topically related images may make a scenario more believable, even if the images provide no additional evidentiary value. Individuals were asked to 'give' food to animals represented as words, then were later presented with animals (either as words or accompanied with an image) and asked to identify if they had

given food to the animal (Cardwell et al., 2016). Participants were more likely to believe that they had given food to the animal if it was accompanied with an image, regardless of whether or not they had actually given food (Cardwell et al., 2016). Alternatively, In a series of studies asking jurors to evaluate the mental state of the defendant at the time of the crime, Schweitzer et al. (2011) found no effect of non-informative neuroimages on jurors' judgements.

In the courtroom, Kellermann (2013) describes the use of non probative images to elicit responses from juries in the form of "truthiness" (feelings that a statement is true) or "falsiness" (feelings that a statement is false), without introducing additional information through the images. As statistical graphics can improve our ability to understand data and model results, it is possible that the use of explanatory images may increase jurors' ability to understand the use of algorithms for evaluating forensic evidence. These graphics differ from those in Cardwell et al. (2016) as they are directly showing evidence that is also being presented and explained verbally. Despite these differences, it is still possible that these images may influence potential jurors' perceptions of the speaker, or their feelings of "truthiness" in the case.

2 Methods

2.1 Participants

Participants were recruited using Prolific, an online platform for scientific research. Prolific offers researchers the ability to obtain a representative sample of participants from a specific region (in this case, the United States) across age, race, and gender. Individuals were additionally asked to self-screen for jury eligibility (no past felony convictions, over the age of majority, not emergency response personnel, etc.). Participants were paid $8.40 for their participation in the study and completed the study with a median response time of around 18 minutes.

2.2 Online Jury Studies

While every attempt was made to use a representative sample in this study, there are certain unavoidable biases that are present in online jury research (Garrett et al., 2020), particularly when transcripts are used in place of videos. Individuals who participate in Prolific surveys may not be representative of eligible jurors in the United States. These individuals also do not undergo the jury selection process, and the jury selection process does not provide a representative sample of individuals in the United States (Abramson, 2018). In order to provide a study of reasonable length, testimony was limited to the relevant firearms evidence, excluding other witnesses and evidence that may have been presented in a real trial. Finally, jurors are unable to deliberate as a group; this may result in different conclusions than would be reached under the group dynamic present in deliberation (Bornstein and Greene, 2011; MacCoun and Kerr, 1988). While acknowledging the limitations of this study format, it is important to note that research on actual jury pools is nearly impossible to conduct for many different logistical and practical reasons, including privacy, cost, and access; even if these barriers were overcome in one jurisdiction, the results from that jurisdiction would not be nationally representative or even representative beyond the sampling area. Thus, online jury studies are an important tool to understand the effect of different manipulations of courtroom procedures, instructions, and admissible evidence.

Table 2: Conclusion language used in the firearms examiner testimony.

Conclusion	Language
Identification	I found that there were sufficient individualizing characteristics to make an identification, that is, that the two bullets were fired from the same barrel.
Inconclusive	I found that the class characteristics of the two bullets were the same, but there was not sufficient agreement among the individual characteristics. My comparison was inconclusive.
Elimination	I found that there was significant disagreement in individual characteristics.

2.3 Design

In order to assess the effect of evidence presentation and the use of algorithms on how jurors evaluated firearms evidence, we developed a factorial experiment, manipulating the examiner's conclusion (identification, inconclusive, or elimination), whether algorithm testimony was included, and whether testimony included demonstrative evidence (pictures and charts), for a $3 \times 2 \times 2$ factorial experiment. Participants were randomly assigned to one of the twelve experimental conditions.

Participants were presented with a trial scenario either with or without the use of a bullet matching algorithm. When the algorithm was absent, participants were provided with testimony from a bullet comparison conducted by a firearms examiner, including the comparison of two test fires from the questioned gun in order to establish a baseline. The wording of the examiner's conclusion is shown in Table 2. When the algorithm was present, participants read the same firearms examiner testimony, with an additional algorithmic comparison of the bullets (which supported the firearms examiner's conclusion). In the transcript, the firearms examiner explained their training in the bullet matching algorithm as well as that the algorithm produces a score between 0 and 1 (0.976 for the test fires from the gun, 0.034 for the elimination condition, 0.34 for the inconclusive condition, and 0.989 for the identification condition). The transcript then shows the examiner's interpretation of the algorithm results in reference to their own conclusion, as shown in Table 1. Following the firearms examiner's testimony, the transcript provided testimony from an individual involved in the development of the algorithm, describing the algorithm's process and limitations. This is consistent with the way DNA comparison algorithms were presented before these algorithms were ubiquitous: a representative from the company providing the algorithm would testify about its development. Similar situations often arise when investigators make use of algorithms for triangulating a phone's location or linking posts made under different accounts to the same person. Once an algorithm's use becomes commonplace, the algorithm expert is often not required to testify, but our goal was to assess the initial stage of the use of a bullet matching algorithm in practice.

In demonstrative evidence conditions, images demonstrating barrel rifling (baku13, 2005), a fired bullet (Gremi-ch, 2009), and a comparison microscope with striation marks were included in the testimony. When both demonstrative evidence and the algorithm were used, the testimony also included an image of the land-to-land comparison grids. Grids generated by the algorithm were shown for test fires, which were fired from the same gun, and should result in an identification, and for the questioned bullet comparison, reflecting the conclusion of the firearms examiner, as shown in Figure 2. The testimony of the algorithm expert included images of the

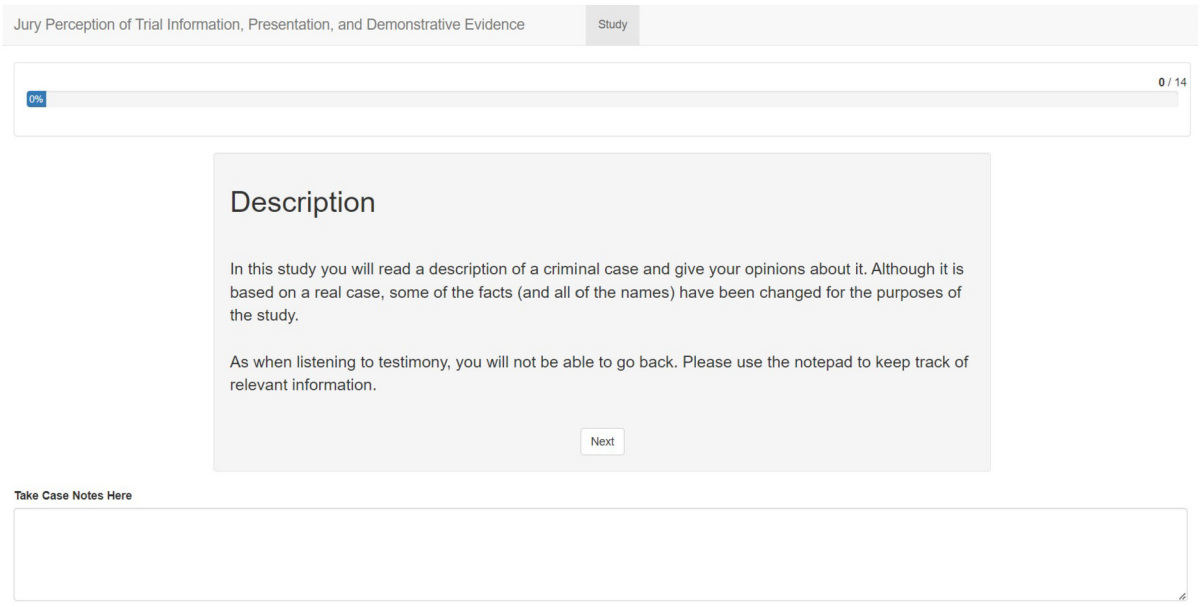

Figure 3: Screenshot of the study description on the first page of the Shiny app.

initial cross section scan with lines indicating shoulder removal (Hare et al., 2017), and signature comparison for matching and non-matching bullet lands, show in Figure 1. Presented algorithm match scores and demonstrative evidence were derived from bullet scans which were evaluated by trained firearms examiners as part of an unpublished study, to ensure that the additional information presented was properly calibrated to the design scenario.

2.4 Study Format

Participants were asked to read a short excerpt of court testimony with regards to an attempted robbery – a scenario based on Garrett et al. (2020). In this case, the only evidence linking the defendant, Richard Cole, to the crime scene is a comparison between a gun found in the car and a bullet recovered from the crime scene. The transcripts were based on testimony given in real trials, as were our edits creating language about algorithms and quantitative evidence. In order to facilitate participants' recall and provide insight into the portions of the testimony participants found to be important, participants were provided with a way to take notes throughout the presentation of testimony, as jurors are allowed to take notes during the trial for use as a memory aid during deliberation. At the end of the transcript(s), participants were asked to rate their impression of the evidence presented, as well as their impression of the expert witnesses using Likert scales. The survey was created using R Shiny (Chang et al., 2023). Figure 3 depicts a screenshot of the study description.

3 Results

Five hundred and ninety-one participants completed the survey; of these, 569 correctly answered both attention check questions, identifying the caliber of the gun used in the crime and selecting a response indicated by the question text. The attention check questions were present to ensure

Table 3: Number of participants per condition.

Algorithm	Picture	Elimination	Inconcl.	Identification
No	No	49	54	55
No	Yes	49	51	52
Yes	No	50	36	32
Yes	Yes	48	50	43

Table 4: Number of participants based on ethnicity.

Ethnicity	Count
Asian	36
Black	75
Mixed	12
Other	10
Unknown	4
White	463

that online participants were reading the testimony as well as the questions before selecting answers. All 569 participants correctly answering both attention check questions were included in the analyses reported below. The number of participants for each of the 12 conditions is shown in Table 3, and demographic information is shown in Table 4.

The average age of participants was 46.46 with a standard deviation of 16.33. 288 participants identified as male, while 308 identified as female, with 4 unknown.

Table 5 summarizes the Likert scale questions asked of participants. Participants were asked to rate the following according to a 7-point Likert scale: their views on the examiner's credibility as well as the evidence's reliability and scientificity (eg. "extremely unreliable" to "extremely reliable"). When the algorithm was absent, participants were asked to rate the reliability and scientificity of the examiner's firearm comparison and the field of firearm comparison as a whole. When the algorithm was present, participants were additionally asked to judge the reliability of the algorithm comparison and the overall firearm comparison (including both the algorithm and the examiner's comparison) in addition to the examiner's firearm comparison and the field of firearm comparison as a whole.

Strength of evidence (e.g. how much evidence there was to suggest the defendant was innocent or guilty) was measured on a 9-point Likert scale.

3.1 Scale Compression

Throughout this analysis, likert-style responses commonly display **scale compression** - most of the participants' responses fall into one or two bins at one extreme of the scale. This suggests that the scenario itself is not calibrated to be able to detect changes in participant views. Scale compression may occur in quantitative responses if, for example, an exam designed to assess student learning results in a class average of 95: the test is not designed to be able

Table 5: Likert scale questions asked of study participants.

Condition	Question
All	How strong would you say the case against the defendant is?
All	How strong is the evidence that the defendant's gun was used to fire the shot in the convenience store, in your opinion?
All	How credible did you find the testimony of Terry Smith (the firearm examiner)?
Algorithm	How credible did you find the testimony of Adrian Jones (the algorithm expert)?
Algorithm	How reliable do you think the firearm evidence in this case is?
All	How reliable do you think the firearm examiner's subjective opinion of the bullet comparison evidence is, in this case?
Algorithm	How reliable do you think the firearm algorithm evidence is, in this case?
All	How scientific do you think the firearm examiner's subjective opinion of the bullet comparison evidence is, in this case?
Algorithm	How scientific do you think the firearm evidence is in this case, overall?
Algorithm	How scientific do you think the firearm algorithm evidence is in this case?
All	Based on this testimony, how would you rate your understanding of the method described for the examiner's personal bullet comparison?
Algorithm	Based on this testimony, how would you rate your understanding of the method described for the bullet matching algorithm?
All	How often do firearm examiners make mistakes when determining whether bullets were fired through the same gun?
All	In general, how reliable do you think firearm evidence is?
All	In general, how scientific do you think firearm evidence is?

to separate the students who understand the material extremely well from the students who have only partially mastered the material. Our participants generally found that the examiner was credible and that the evidence presented was reliable and scientific. This compression is demonstrated in Figure 4, which shows participant selection for reliability categories of the examiner's comparison across all experimental conditions. Here, the vast majority of participants chose the top two categories (508 out of 569, or 89.28%). The other Likert categories have few observations - making it difficult to conduct comparisons across experimental conditions. This lack of variation in responses makes it difficult to use standard statistical approaches, such as linear models. In this paper, we use graphics to explore the data from this study in order to develop testable hypotheses for future iterations of studies using the Cole scenario. In general, we primarily see strong effects when examining questions relating the examiner's conclusion (Identification/Match, Inconclusive, Elimination/Non match) to the likelihood that the gun was used in the crime or that the defendant was guilty. This suggests that at the bare minimum, the scenario is well calibrated to assess the relationship between the examiner conclusion and the

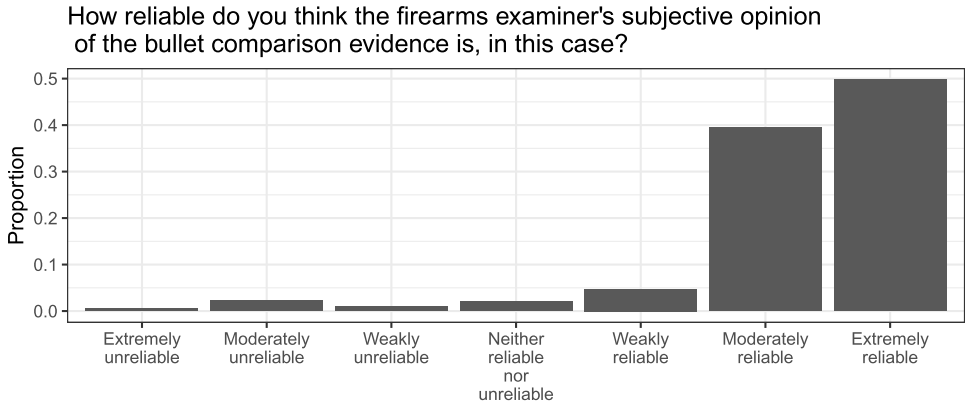

Figure 4: Histogram of perceived reliability of the exmaminer's firearm comparison across all study conditions. There is evidence of significant scale compression across conditions, suggesting that in order to be able to measure and statistically model differences in perception of examiner reliability, the transcripts must contain more information which might cause participants to question the reliability of the examiner and of firearms comparisons.

guilt or innocence of the suspect.

While we used a scenario which had been previously used in similar experiments (Garrett et al., 2018) and modified it to examine the use of algorithms, the original authors of this scenario are not statisticians and did not fully identify the underlying issues that may have contributed to an inability to detect differences between treatment groups in the original experiment. It is important to design and calibrate these types of user experiments so that presented evidence is "just right" - not too strong, not too weak. This calibration allows any manipulations, such as those in this study, to show up in the resulting data. When scale compression is present, however, it is hard to show increased confidence in the examiner's opinion when the base assessment is already at "extremely reliable". Thus, future iterations of this study need to do more to challenge the examiner's reliability and the perceived scientificity of the discipline - not because these things are necessarily in question (though there have been several successful legal challenges on the use of firearms evidence in court), but because in order to understand the effects of external manipulations, it is important to set up an experiment where these effects can be measured. Inclusion of jury instructions (reminding the jury that they are the triers of fact and must be convinced beyond a reasonable doubt) and stronger cross-examination which focuses on error rates of firearms examination and examples of false convictions related to firearms may help to reduce scale compression and provide a more nuanced view of the effect of other interventions.

3.2 Credibility

All participants (regardless of experimental condition) were asked to rate the credibility of the firearms examiner; results are provided in Figure 5. 535 individuals selected "moderately credible" and "extremely credible" in the scale, while 34 individuals selecting lower categories.

In addition, participants who were assigned to the algorithm condition were asked to rate the credibility of the algorithm expert. These results also show scale compression: 250 chose "extremely credible" or "moderately credible" and 9 chose a lower category.

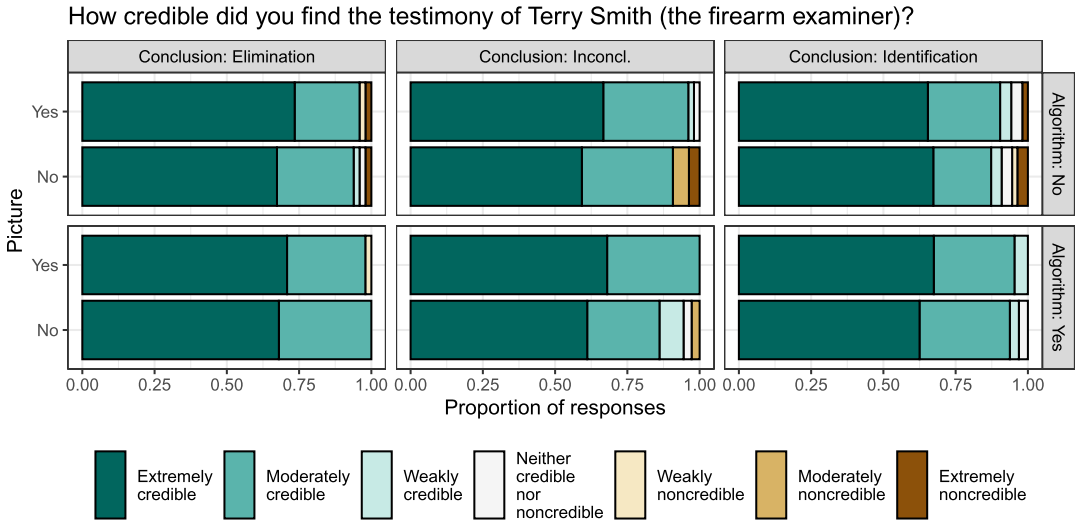

Figure 5: Histogram of examiner credibility by conclusion, demonstrative evidence, and algorithm conditions. Inconclusive conclusions have slightly lower credibility (particularly in the absence of demonstrative evidence), but overall, the primary observation when considering this data is that there is significant scale compression.

3.3 Reliability

Figure 6 displays the results for the perception of firearm reliability as a field; a question that was also asked of all participants. This chart is similar to the graphic for credibility in that the top two categories ("moderately reliable" and "extremely reliable") contain many of the responses (490 observations), while other lower categories are sparsely populated (79 observations). In this case, however, there is some correlation between the conclusion and the ratings of reliability - specifically, both in the presence and absence of the algorithm, "moderately reliable" was more popular than "extremely reliable" for inconclusive decisions (for a cumulative total of 96 and 55 observations, respectively). This is the opposite of the trend seen in the elimination and identification conditions, where "moderately reliable" contains similar or less observations than "extremely reliable" (93 and 85 respectively in the elimination condition; 76 and 85 respectively in the identification condition).

3.4 Scientificity

Figure 7 shows the opinion of the scientificity of the examiner's comparison, divided by algorithm condition. As before, most responses are at the high end of the scale, with 491 observations in the two highest categories, and 78 observations in the remaining categories. When evaluating the top two categories of scientificity, the inconclusive category demonstrates a different trend than the other categories when the algorithm is present. While other categories either favor "extremely scientific" over "moderately scientific" or have approximately equal results, when the algorithm is present and there is an inconclusive decision, participants tended toward "moderately scientific" over "extremely scientific" (45 observations and 26 observations, respectively).

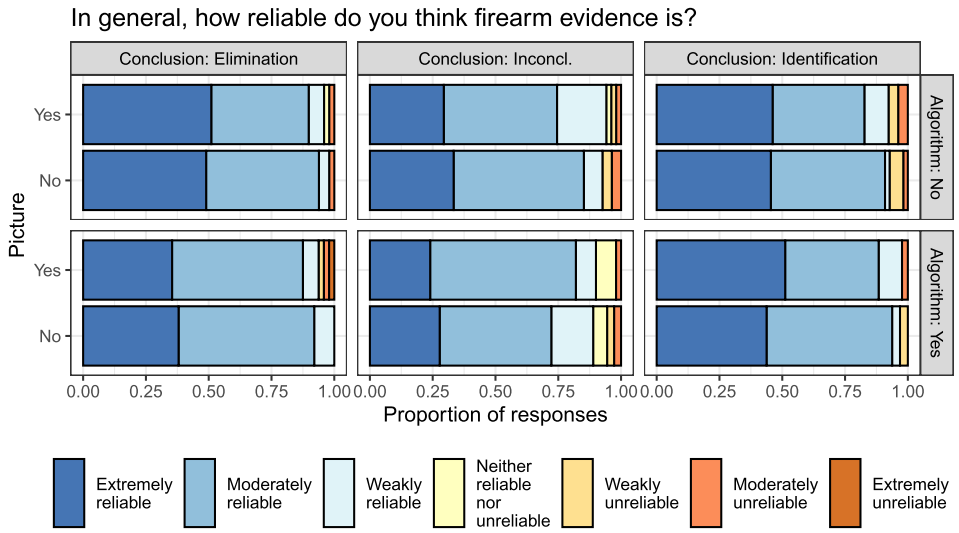

Figure 6: Perception of reliability of the field of firearms examination when manipulating demonstrative evidence (pictures) and use of the algorithm. Participants who read testimony about the algorithm were slightly more likely to say the field was moderately reliable and less likely to say the field was extremely reliable. Similarly, demonstrative evidence may be associated with a small reduction in perception of the reliability of the field.

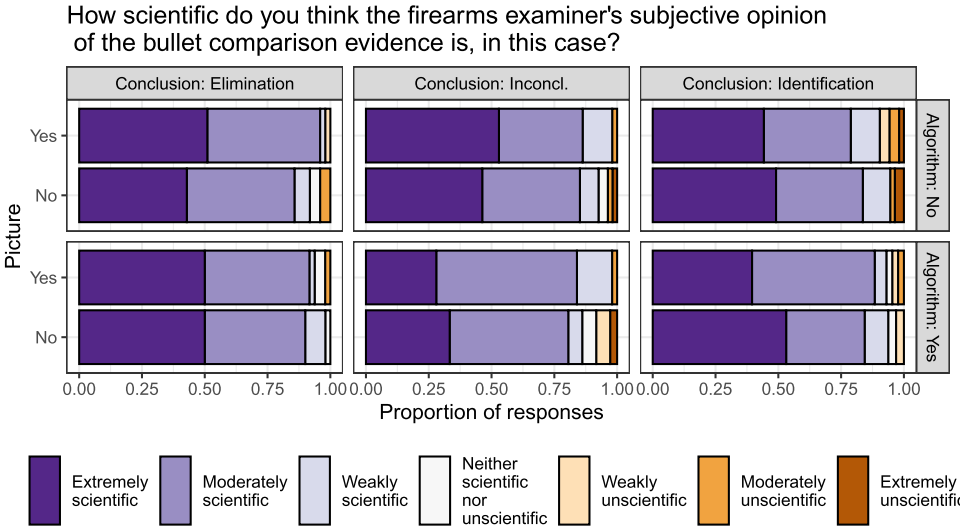

Figure 7: Perceived scientificity of the examiner's opinion by examiner conclusion, use of demonstrative evidence, and use of the algorithm. There are relatively few differences in perceived scientificity of the examiner's opinion across conditions, though inconclusive opinions seem to reduce perceptions of scientificity in the presence of the algorithm.

3.5 Understanding

Participants were asked to rate their understanding of the firearms examiner's comparison, as well as their understanding of the algorithm (when present). These results are shown in Figure 8

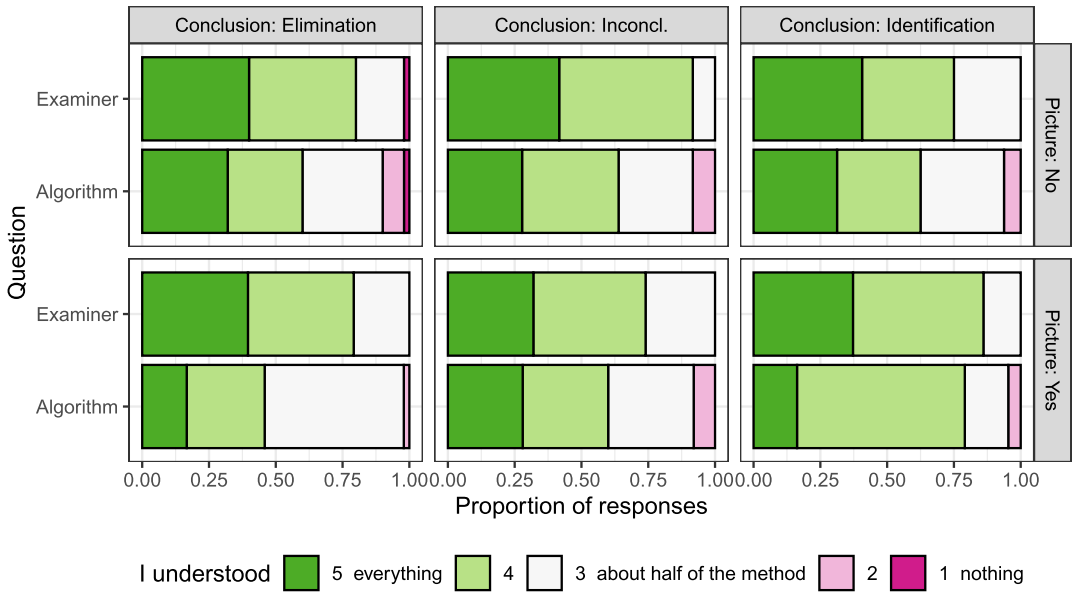

Figure 8: For the algorithm condition, participant understanding of algorithm and examiner testimony by conclusion and picture inclusion. Across conditions, participants were less confident in their understanding of the algorithm testimony compared to the examiner testimony.

Table 6: Table of participant understanding for participants in the algorithm condition, where 1 corresponds to understanding nothing and 5 corresponds to understanding everything.

Question	Conclusion	1	2	3	4	5
Algorithm	Elimination	1	5	40	28	24
Algorithm	Inconcl.	0	7	26	29	24
Algorithm	Identification	0	4	17	37	17
Examiner	Elimination	1	0	19	39	39
Examiner	Inconcl.	0	0	16	39	31
Examiner	Identification	0	0	14	32	29

and Table 6 for those given the algorithm condition - comparing their rated understanding of the firearms examiner's comparison and the algorithm comparison when both are present. Here, individuals mostly selected the three highest categories (242 vs 17 for the understanding of the algorithm and 258 vs 1 for the understanding of the forensics examiner's comparison), and there appears to be a difference in participants' rating of the algorithm and examiner when the algorithm is present. The algorithm was generally assigned lower values of understanding than the examiner. It should be noted that participants' rating of their own understanding may not truly indicate the understanding of the participants. Dunning et al. (2004) found that an individual's rating of their own knowledge may not directly correspond to their actual learning level. Figure 9 shows the perceived understanding of the firearms examiner's comparison across all categories.

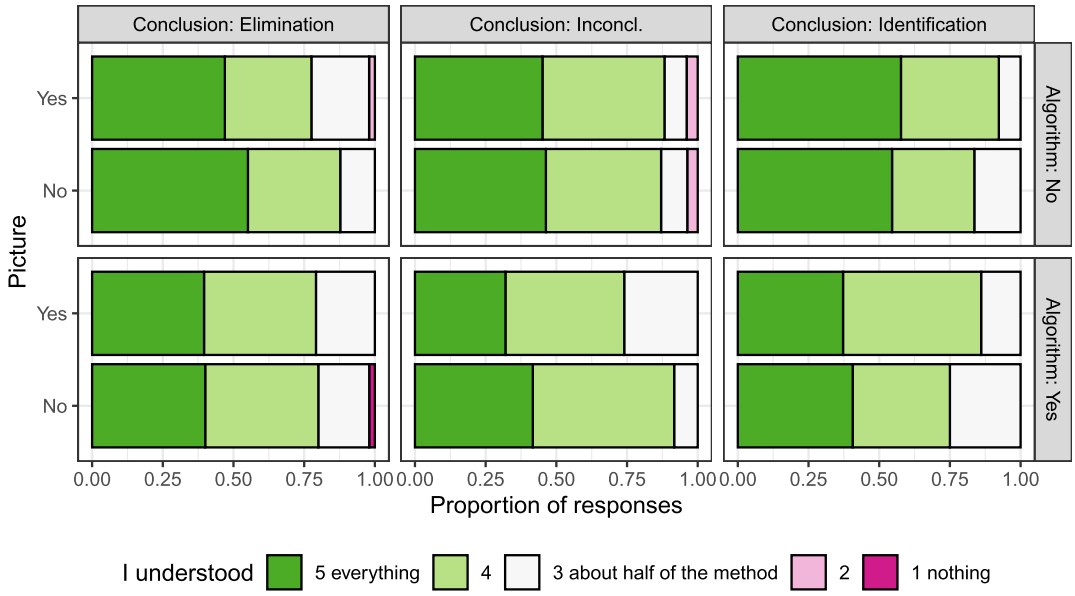

Figure 9: Histogram of perceived understanding of the firearms examiner's bullet comparison.

3.6 Probability

Linear models are used for questions regarding the probability that Cole committed the crime and the probability that the gun was used in the crime. These probabilities are modeled with a beta generalized linear model that considers the interaction between conditions; results can be found in the supplementary material. Figure 10 displays participant reported probabilities that Cole committed the crime or Cole's gun was involved in the crime. This figure indicates that there is a difference between conclusions (as expected, due to differences in the strength of evidence presented), but not much of a difference for cases when the algorithm is included. Here, in the case of an elimination, probabilities assigned by participants appear to be extremely low, while in the case of an identification, probabilities appear to be relatively high. In the case of an inconclusive decision, participants also tended to favor a lower probability, although this trend is not as extreme as in the case of an elimination. Because bullet match scores, resulting from the use of the algorithm, range from 0 to 1, there was concern that participants may incorrectly interpret this value as the probability of involvement in the crime. The vertical lines in Figure 10 indicate this match score for each condition. There does not appear to be evidence that those who received the algorithm condition anchored to the match score values, when comparing the distribution of those who did not receive the match score (the non-algorithm condition).

Participants were also asked whether or not they would choose to convict Cole, based on the criminal trial standard of "beyond a reasonable doubt". The conviction decision of participants and their assigned probabilities of involvement in the crime are shown in Figure 11, grouped by the conclusion of the firearms examiner. 10 out of 196 (5%) who received the elimination condition, 13 out of 191 (7%) who received the inconclusive decision, and 112 out of 182 (62%) who received the identification condition chose to convict. The few individuals who chose to convict in the elimination and inconclusive conditions assigned higher probabilities than their counterparts, indicating that individuals thought Cole was guilty or that the gun was used in the crime. This discrepancy between conclusion and participant selections may be due to faulty

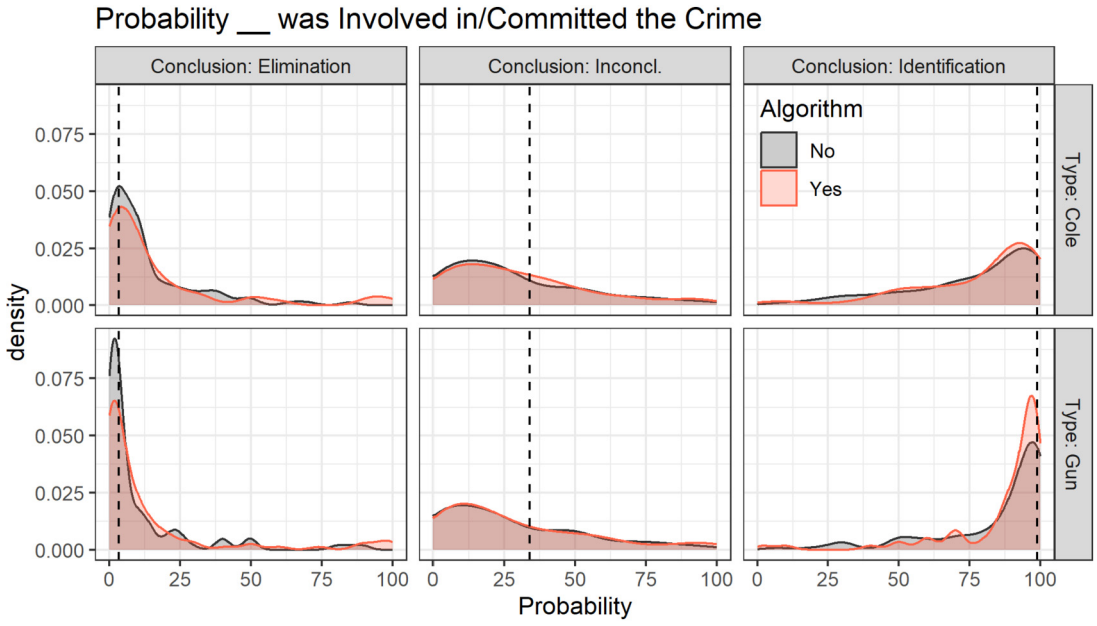

Figure 10: Probability the gun was used in the crime, or that Cole committed the crime. Dashed lines indicate bullet match scores for the algorithm. Participants were less committed to Cole's guilt or innocence than to the gun's involvement (or not) in the crime across all conditions, suggesting that at least some participants correctly discerned that the gun was not necessarily an indication of Cole's involvement, e.g. that the evidence linking him to the crime was circumstantial.

internal logic, or a general belief that someone whose case has proceeded to trial is unlikely to be innocent, regardless of the evidence. When the examiner made an identification, participants who chose to convict assigned higher probabilities that Cole committed the crime than those who did not convict. The choice to not convict in the identification condition may be due to the "beyond a reasonable doubt" threshold: the only evidence that participants read matched the gun to the crime scene, but did not have evidence that tied Cole specifically to the crime. This distinction between the gun and Cole can be seen in the relationship between those who chose not to convict when the bullets matched in Figure 11: they overall assigned a higher probability that gun was used in the crime than they did that Cole was involved in the crime (mean values of 75.57 and 60.06 with standard deviations of 25.94 and 25.47, respectively).

3.7 Summary

A common feature in many of these charts is scale compression - most individuals limited their Likert scale selection to the two highest values in terms of credibility, reliability, and scientificity. This demonstrates that, across all experimental conditions, participants perceive the examiner as credible, and the evidence as reliable and scientific. In this study, we were unable to discern a difference in perception of reliability, credibility, or scientificity between the algorithm and non-algorithm conditions or between demonstrative evidence and standard testimony conditions, though there are suggestions that effects may be present but not detectable due to scale compression (e.g. Figure 9). Feelings regarding the strength of evidence, conviction

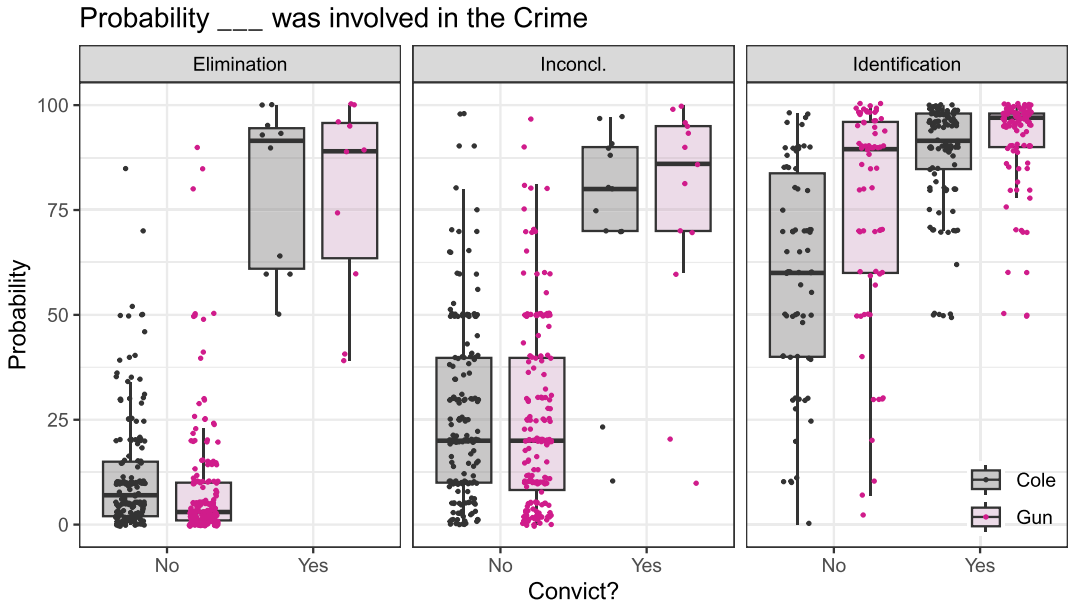

Figure 11: Probability that Cole committed the crime (black) and that the gun was used in the crime (pink) based on whether the participants chose to convict. Overall, participants responded to the testimony presented (identification, inconclusive, elimination) in a reasonable manner. The largest discrepancy between probabilities for Cole and probabilities for the gun was in identification; that is, many participants recognized that information about the gun's use (or not) in the crime may not provide the same level of information about cole's guilt or innocence.

decision, and probability Cole/the gun was involved in the crime varied by the conclusion of the firearms examiner, as expected.

The examiner's conclusion had the largest effect, both in the expected areas of strength of evidence and probability that Cole/the gun was involved in the crime, as well as in areas of reliability and scientificity. There was also some difference between perceptions of the algorithm and the traditional bullet analysis method. In particular, the explanation of the algorithm received lower scores of understanding than the explanation of the firearms examiner's bullet comparison. This may be due to math aversion - terms used in the algorithm description sound difficult to understand, which may reduce the willingness of participants to try to parse the explanation, even if the explanation itself is not technical.

4 Discussion

In this study, questions using Likert scales were frequently subject to scale compression, complicating our attempts to statistically model responses by condition. There are several solutions to this problem, but one of the most promising is to alter the scenario to introduce more doubt about the process of firearms examination, shifting participant answers towards the middle of the Likert scale. Introducing discussion of erroneous convictions due to firearms evidence, adding more information about error rates in firearms examination, and including instructions to the jury from the judge before the examiner's testimony are all options which have occurred in real court cases and which would be expected to reduce scale compression and improve the statistical ability to

discern effects of algorithms and demonstrative evidence on interpretation of firearms testimony.

Another solution is to move away from Likert scales to responses in different formats: probabilities, numerical chance, betting, and opinion of guilt. We are in the process of executing a short follow-up study which examines different ways to ask these questions, and our goal is to explore not only how participant responses change, but to also demonstrate ways to model these different response types effectively.

Very few studies are executed perfectly; this study is no exception. There were two minor mistakes in the transcript which were present for approximately the first half of participants. These typos included referring to the firearms expert as Alex Smith as opposed to Terry Smith in all scenario questions, and for the cross examination in the elimination testimony. In the case of non-algorithm inconclusive testimonies, the question: "Can you describe the process of obtaining these test fired bullets?" was missing, but the response: "The test-fired bullets came from a test fire of the gun recovered from the traffic stop." remained unchanged. There was no indication that participants were confused by these typos, but because Prolific recruits participants for separate demographic categories, these typos are confounded with the demographic variables, because blocks for demographics with higher participation on Prolific (younger ages, whites) fill up more quickly than blocks with lower participation on the site.

In addition to the data-driven modifications above, we also noticed during the execution of this study that participants may not have fully understood the difference between initial testimony and cross examination, or which witnesses were testifying for the prosecution vs. the defense. The transcript format provided in this study followed the same format as the court transcripts - speakers were indicated by "Q:" and "A:", but the identity of the speaker and their alignment within the courtroom could be easily confused. To address this, we plan to implement a more visual representation of a courtroom transcript, using graphics to show each individual who is speaking and subtle cues to indicate which side they are testifying for.

We expect that these combined modifications will produce a study with more nuanced participant responses and will alleviate the scale compression seen in this experiment.

Supplementary Material

The Supplementary Material includes: (1) Statistical models and additional graphs for study questions; (2) Code for the creation of the survey app; (3) Survey data and testimony outline; (4) Source files for paper.

Acknowledgement

Thank you to the study participants, without whom this work would not be possible.

Computation and visualizations were made possible thanks to the following R packages: 'tidyverse'(Wickham et al., 2019), 'RColorBrewer' (Neuwirth, 2022), 'patchwork' (Pedersen, 2022), 'gt' (Iannone et al., 2023), 'MASS' (Venables and Ripley, 2002), 'emmeans' (Lenth, 2023), 'shiny' (Chang et al., 2023), and 'mgcv' (Wood, 2017).

Funding

This work was funded (or partially funded) by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreements 70NANB15H176 and

70NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College and University of Nebraska, Lincoln.

References

Abramson J (2018). Jury selection in the weeds: Whither the democratic shore? *University of Michigan Journal of Law Reform*, 52(1): 1. https://doi.org/10.36646/mjlr.52.1.jury

baku13 (2005). L7 105mm tank gun Cut model. https://commons.wikimedia.org/wiki/File:105mm_tank_gun_Rifling.jpg.

Bornstein BH, Greene E (2011). Jury decision making: Implications for and from psychology. *Current Directions in Psychological Science*, 20(1): 63–67. https://doi.org/10.1177/0963721410397282

Cardwell BA, Henkel LA, Garry M, Newman EJ, Foster JL (2016). Nonprobative photos rapidly lead people to believe claims about their own (and other people's) pasts. *Memory & Cognition*, 44(6): 883–896. https://doi.org/10.3758/s13421-016-0603-1

Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, et al. (2023). *shiny: Web application framework for R*. R package version 1.7.4.1.

Dunning D, Heath C, Suls JM (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, 5(3): 69–106. Publisher: SAGE Publications Inc. https://doi.org/10.1111/j.1529-1006.2004.00018.x

Garrett B, Mitchell G, Scurich N (2018). Comparing categorical and probabilistic fingerprint evidence. *Journal of Forensic Sciences*, 63(6): 1712–1717. https://doi.org/10.1111/1556-4029.13797

Garrett BL, Scurich N, Crozier WE (2020). Mock jurors' evaluation of firearm examiner testimony. *Law and Human Behavior*, 44(5): 412–423. https://doi.org/10.1037/lhb0000423

Gremi-ch (2009). English: A 5.66x45mm (.223 rem.) boat tailed FMJ spitzer bullet laying on a ruler with a scale in centimeter. https://commons.wikimedia.org/wiki/File:GP90-bullet.JPG?uselang=fr.

Hare E, Hofmann H, Carriquiry A (2017). Automatic matching of bullet land impressions. *Annals of Applied Statistics*, 11(4): 2332–2356. https://doi.org/10.1214/17-AOAS1080 MR3743299

Hofmann H, Carriquiry A, Vanderplas S (2021). Treatment of inconclusives in the AFTE range of conclusions. *Law, Probability and Risk*, 19(3): 317–364. Tex.eprint. https://academic.oup.com/lpr/article-pdf/19/3-4/317/38817993/mgab002.pdf.

Iannone R, Cheng J, Schloerke B, Hughes E, Lauer A, Seo J (2023). *gt: Easily create presentation-ready display tables*. R package version 0.9.0.

Kellermann K (2013). Trial advocacy: Truthiness, falsiness, and nothingness. *Jury Expert*, 25: 38. 00001.

Koehler JJ (2001). When are people persuaded by DNA match statistics? *Law and Human Behavior*, 25(5): 493–513. https://doi.org/10.1023/A:1012892815916

Lenth RV (2023). *emmeans: Estimated marginal means, aka least-squares means*. R package version 1.8.7.

MacCoun RJ, Kerr NL (1988). Asymmetric influence in mock jury deliberation: Jurors' bias for leniency. *Journal of Personality and Social Psychology*, 54(1): 21–33. Publisher: American Psychological Association. https://doi.org/10.1037//0022-3514.54.1.21

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

Neuwirth E (2022). *RColorBrewer: ColorBrewer palettes*. R package version 1.1-3.

Pedersen TL (2022). *patchwork: The composer of plots*. R package version 1.1.2.

President's Council of Advisors on Science and Technology, (2016). Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature Comparison Methods, *Technical report*.

Schweitzer NJ, Saks MJ, Murphy ER, Roskies AL, Sinnott-Armstrong W, Gaudet LM (2011). Neuroimages as evidence in a mens rea defense: No impact. *Psychology, Public Policy, and Law*, 17(3): 357–393. https://doi.org/10.1037/a0023581

Song J, Chu W, Vorburger TV, Thompson R, Renegar TB, Zheng A, et al. (2012). Development of ballistics identification- from image comparison to topography measurement in surface metrology. *Measurement Science & Technology*, 23: 054010, 1–6. https://doi.org/10.1088/0957-0233/23/5/054010

Swofford H (2017). Information paper. *Technical report*, Defense Forensic Science Center. https://osf.io/8kajs.

Swofford H, Champod C (2022). Probabilistic reporting and algorithms in forensic science: Stakeholder perspectives within the American criminal justice system. *Forensic Science International: Synergy*, 4: 100220, 1–25. ISSN 2589-871X. https://doi.org/10.1016/j.fsisyn.2022.100220

Vanderplas S, Nally M, Klep T, Cadevall C, Hofmann H (2020). Comparison of three similarity scores for bullet LEA matching. *Forensic Science International*, 308. https://doi.org/10/gn6487

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. Springer, New York, 4th edition. ISBN 0-387-95457-0. MR1337030

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. (2019). Welcome to the Tidyverse. *The Journal of Open Source Software*, 4(43): 1686. https://doi.org/10.21105/joss.01686

Wood S (2017). Generalized additive models: An introduction with r. MR3726911