JOURNAL OF DATA SCIENCE xx (xx), 1–18 DOI: 0000
July 2022

Supplement to 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, 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 Ordinal Logistic Regression (Likert Scales)

Because there are not enough observations in all categories, only categories with enough observations are considered. Most analyses are also limited to main effects. Due to the scale compression mentioned throughout the article, this analytical approach is not recommended, as it ignores key aspects of the data collection process (such as the inclusion of the complete scale). If there are only two categories for consideration, the response is considered as binomial. If there are more than two categories, ordered logistic regression is first implemented using the 'VGAM' package (the 'polr' package implementation failed to find starting values in several cases), and assumptions of proportional odds are tested by comparing the likelihood to the model without the parallel odds assumption. Unless otherwise noted, the parallel odds assumption holds. In a few cases, there were not enough observations for the model to be computed without the parallel odds assumption.

1.1 Credibility

1.1.1 How credible did you find the testimony of Terry Smith (the firearm examiner)?

Figure 1 indicates that only the top two categories of the Likert scale have enough data for a formal analysis. Thus, only the top two categories were considered using a binomial generalized

---

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

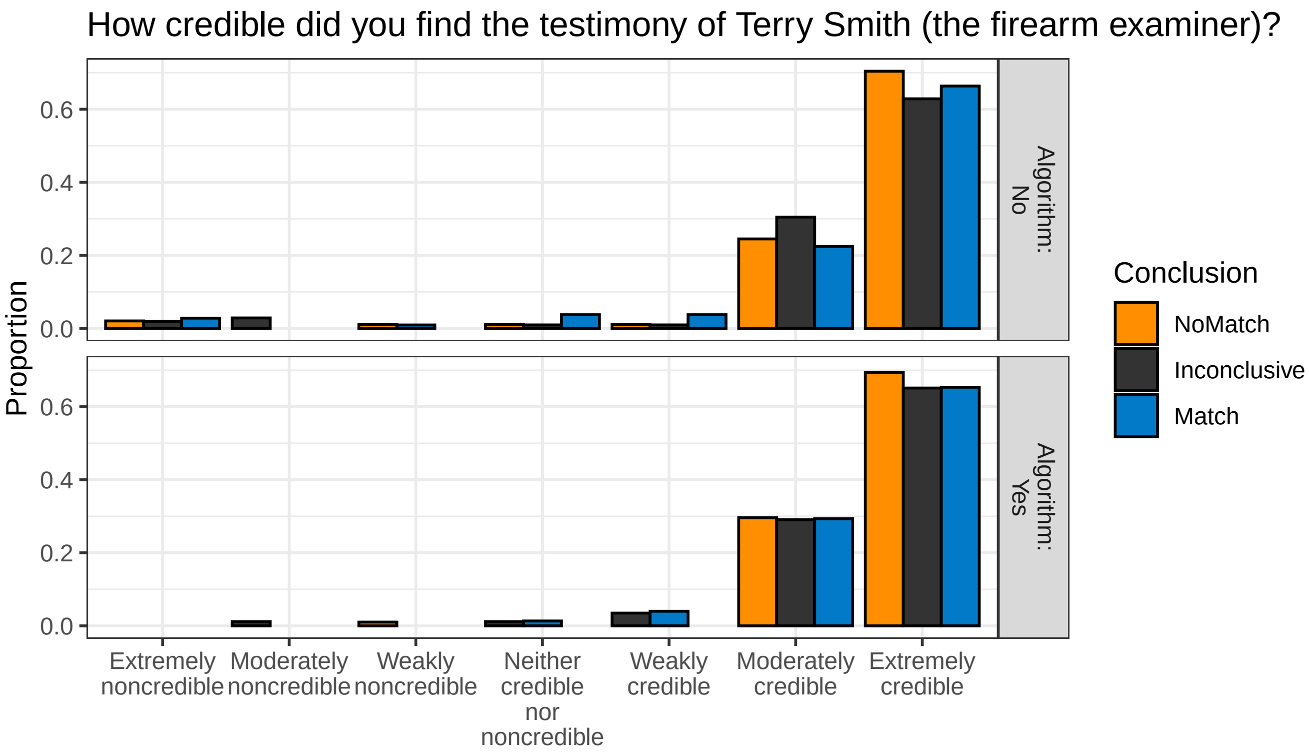

Figure 1: Histogram of Firearms Examiner Credibility

linear model. This model does not fully consider the responses or response options given to participants, and is not recommended. There were not significant differences between conditions.

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: firetestcred
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                         534     645.82
## Algorithm   1  0.41865       533     645.40   0.5176
## Conclusion  2  0.95523       531     644.44   0.6203
## Picture     1  0.19970       530     644.24   0.6550
```

1.1.2 How credible did you find the testimony of Adrian Jones (the algorithm expert)?

Similar to Figure 1, Figure 2 shows that most individuals only selected the top two categories of the Likert scale. Thus, as before, only the top two categories will be considered in statistical

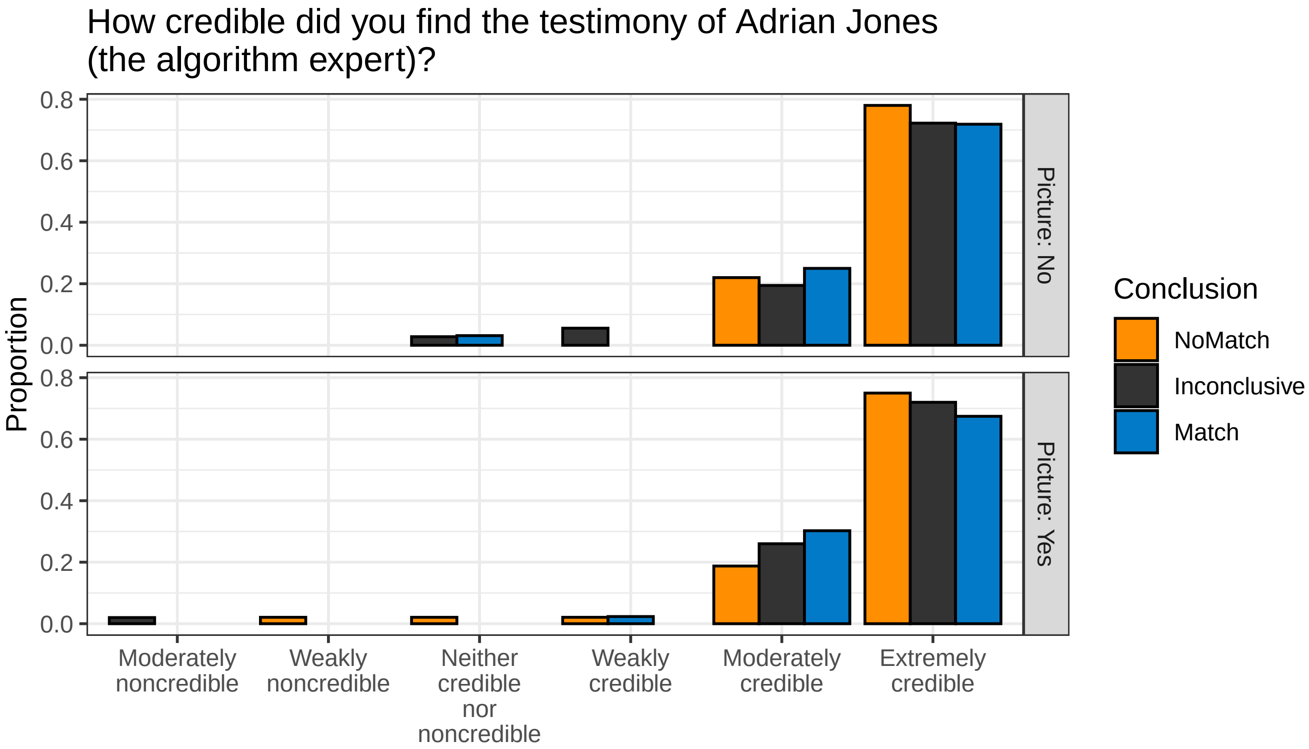

Figure 2: Histogram of Algorithm Expert Credibility

analysis (although this does not reflect how the data was collected).

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: algtestcred
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                         249     277.82
## Conclusion  2  1.32358       247     276.50   0.5159
## Picture     1  0.19895       246     276.30   0.6556
```

1.2 Reliability

1.2.1 In general, how reliable do you think firearm evidence is?

Figure 3 has observations from each condition combination in the top three categories of the Likert scale (weakly reliable, moderately reliable, and extremely reliable), so an ordered logistic regression using the three top categories is used.

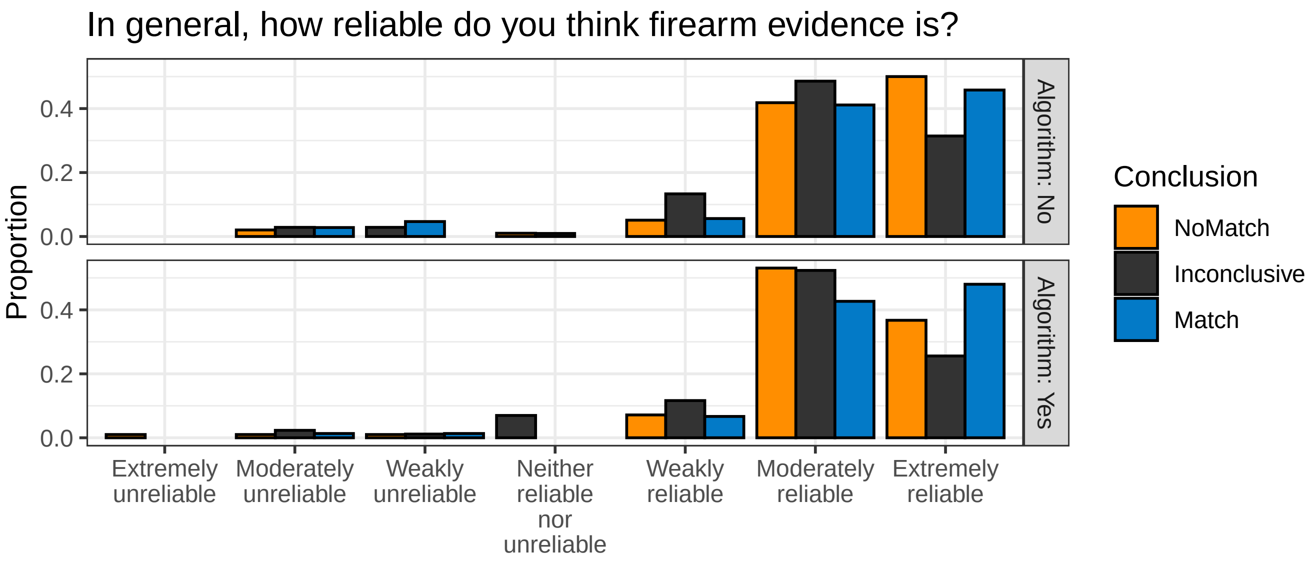

Figure 3: Histogram of perceived firearm reliability as a field

```
##                             Estimate Std. Error    z value     Pr(>|z|)
## (Intercept):1             2.69064642  0.2301020 11.6932783 1.379566e-31
## (Intercept):2            -0.04658948  0.1823326 -0.2555192 7.983221e-01
## ConclusionInconclusive -0.63775184  0.2057890 -3.0990566 1.941379e-03
## ConclusionMatch           0.14463400  0.2051155  0.7051343 4.807267e-01
## PictureYes               -0.04996001  0.1683530 -0.2967574 7.666517e-01
## AlgorithmYes             -0.23425313  0.1693210 -1.3834855 1.665161e-01
```

1.2.2 How reliable do you think the firearm evidence in this case is?

Based on Figure 4, the top three categories contain results and are used in analysis (using ordered logistic regression).

```
##                             Estimate Std. Error    z value     Pr(>|z|)
## (Intercept):1             2.75992073  0.3442074  8.0181921 1.073129e-15
## (Intercept):2             0.33859995  0.2606274  1.2991725 1.938847e-01
## ConclusionInconclusive -1.02844224  0.3351858 -3.0682750 2.152984e-03
## ConclusionMatch          -0.05880294  0.3246099 -0.1811495 8.562502e-01
## PictureYes                0.06098488  0.2724282  0.2238567 8.228688e-01
```

1.2.3 How reliable do you think the firearms examiner's subjective opinion of the bullet comparison is, in this case?

In this case, there may not be enough observations in the "weakly reliable" category for analysis - so only the highest two categories of the Likert scale are analyzed.

```
##                             Estimate Std. Error    z value     Pr(>|z|)
## (Intercept):1             3.48591557  0.2692876 12.9449556 2.509082e-38
```

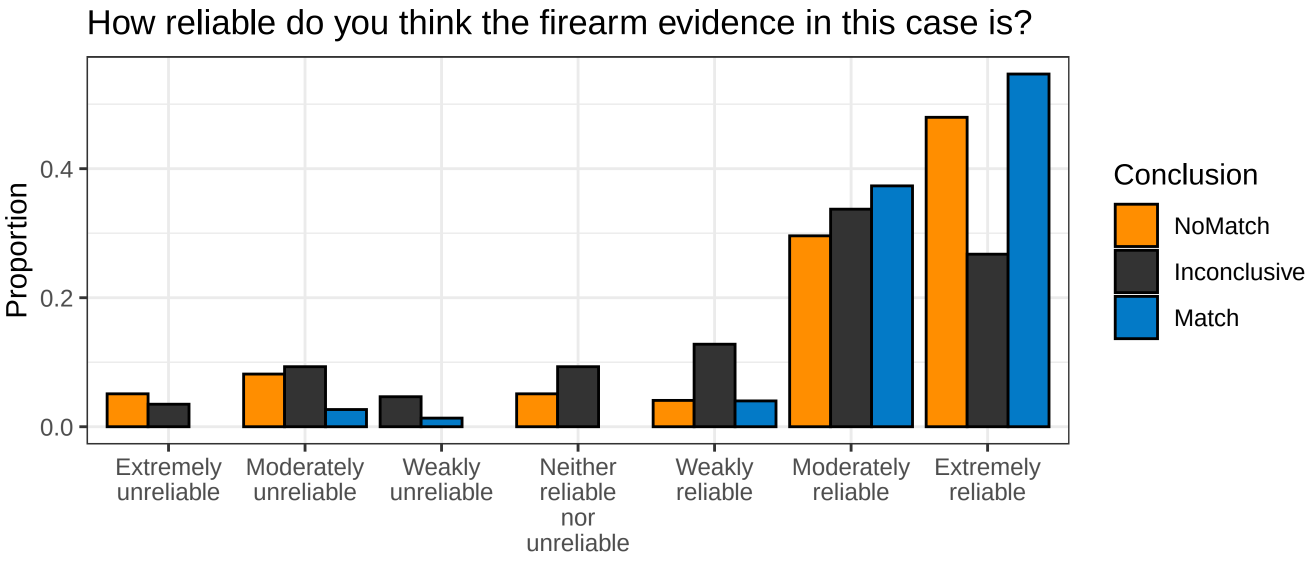

Figure 4: Histogram of overall case reliability

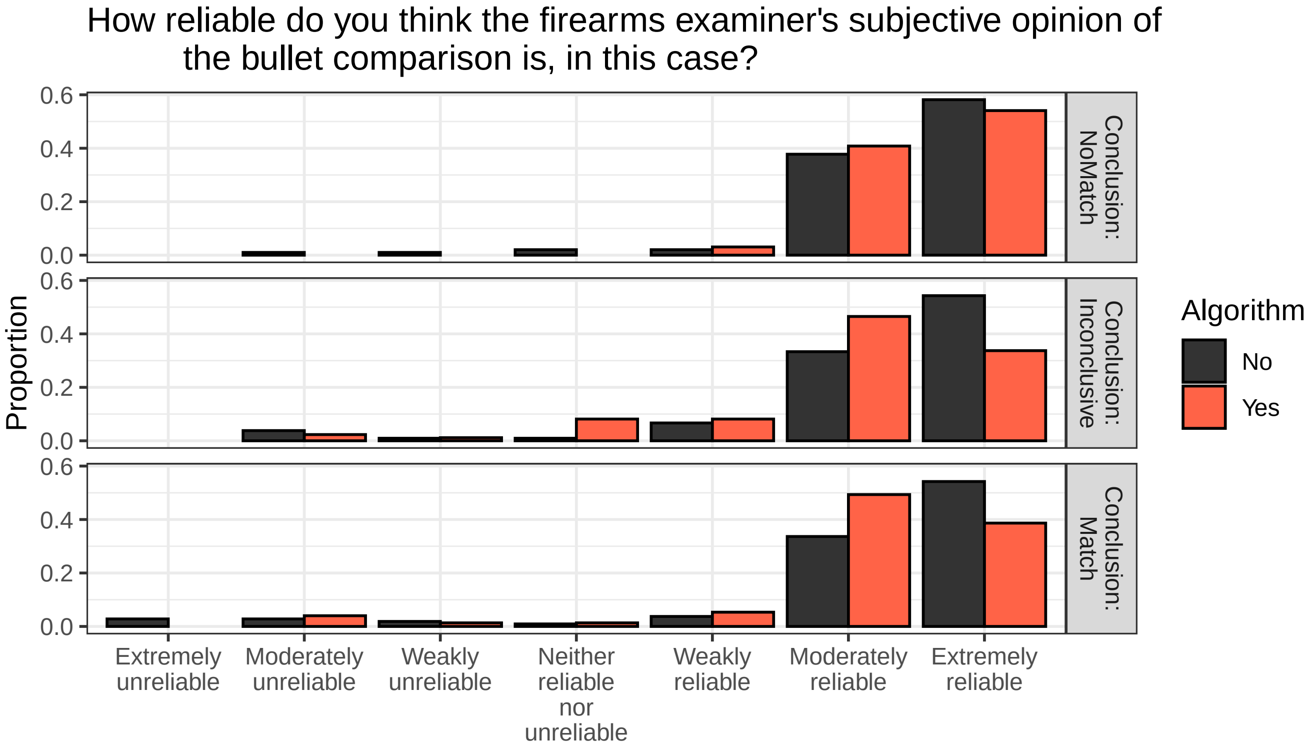

Figure 5: Histogram of perceived firearm exam reliability

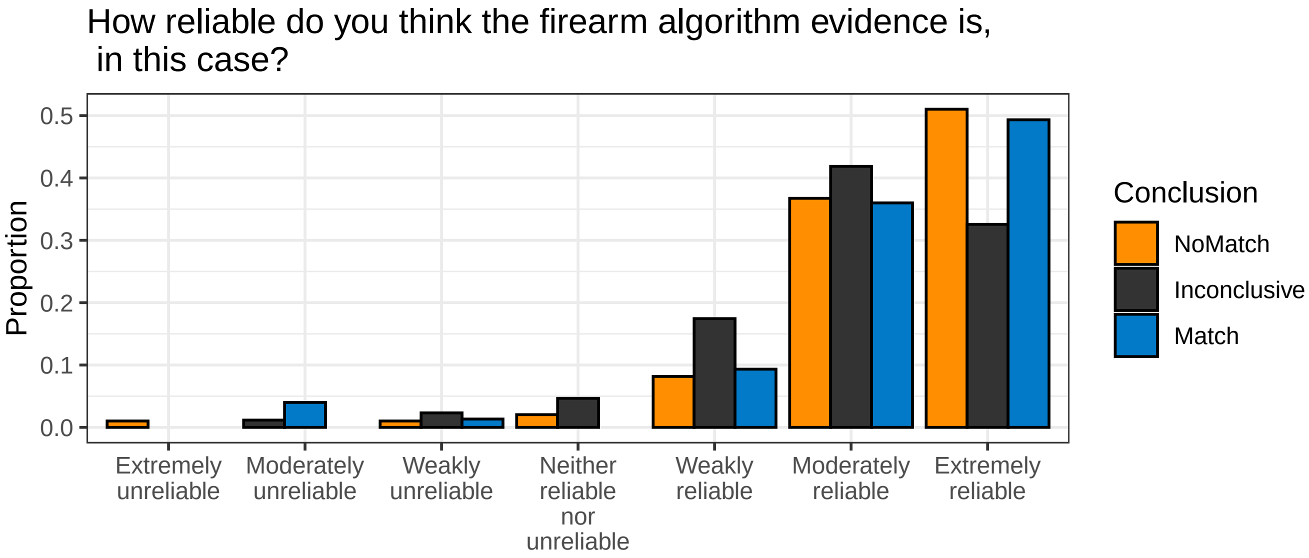

Figure 6: Histogram of perceived algorithm reliability

```
## (Intercept):2           0.62481123  0.1899804  3.2888202 1.006083e-03
## ConclusionInconclusive -0.43256548  0.2083372 -2.0762755 3.786847e-02
## ConclusionMatch        -0.28182315  0.2112147 -1.3342969 1.821066e-01
## PictureYes             -0.08513639  0.1718398 -0.4954406 6.202891e-01
## AlgorithmYes           -0.51673224  0.1729951 -2.9869753 2.817525e-03
```

1.2.4 How reliable do you think the firearm algorithm evidence is, in this case?

In Figure 6, there are enough observations in the three highest categories for ordered logistic regression.

```
##                           Estimate Std. Error    z value     Pr(>|z|)
## (Intercept):1           2.53041886  0.3294040  7.6818090 1.568576e-14
## (Intercept):2           0.20609156  0.2570606  0.8017236 4.227129e-01
## ConclusionInconclusive -0.60818214  0.3317086 -1.8334830 6.673077e-02
## ConclusionMatch        -0.04176306  0.3195061 -0.1307113 8.960037e-01
## PictureYes             -0.10797218  0.2700328 -0.3998483 6.892682e-01
```

1.3 Scientificity

1.3.1 In general, how scientific do you think firearm evidence is?

The top three categories are used for analysis (Figure 7).

```
##                           Estimate Std. Error    z value     Pr(>|z|)
## (Intercept):1           2.63271547  0.2326023 11.3185274 1.062415e-29
## (Intercept):2          -0.09899810  0.1834667 -0.5395972 5.894748e-01
## ConclusionInconclusive -0.60334328  0.2027738 -2.9754504 2.925587e-03
## ConclusionMatch        -0.03547145  0.2063989 -0.1718587 8.635486e-01
```

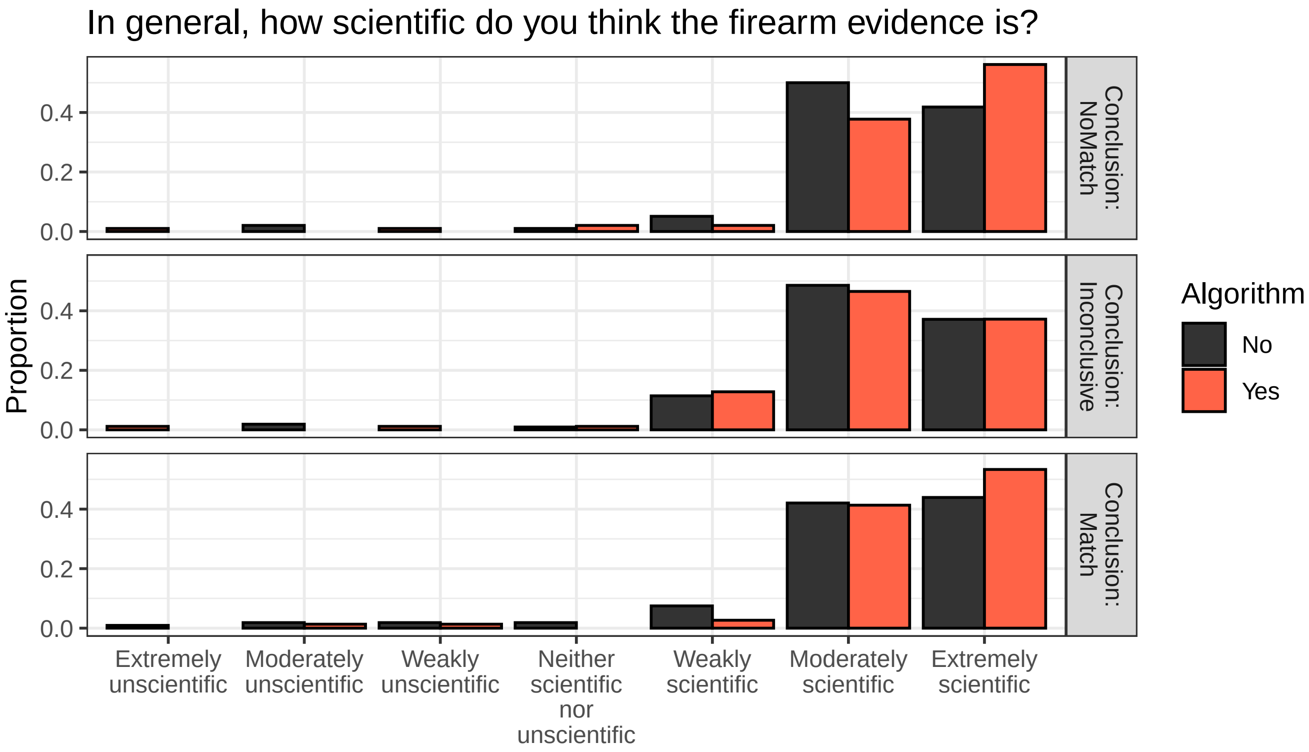

Figure 7: Histogram of perceived firearm scientificity as a field

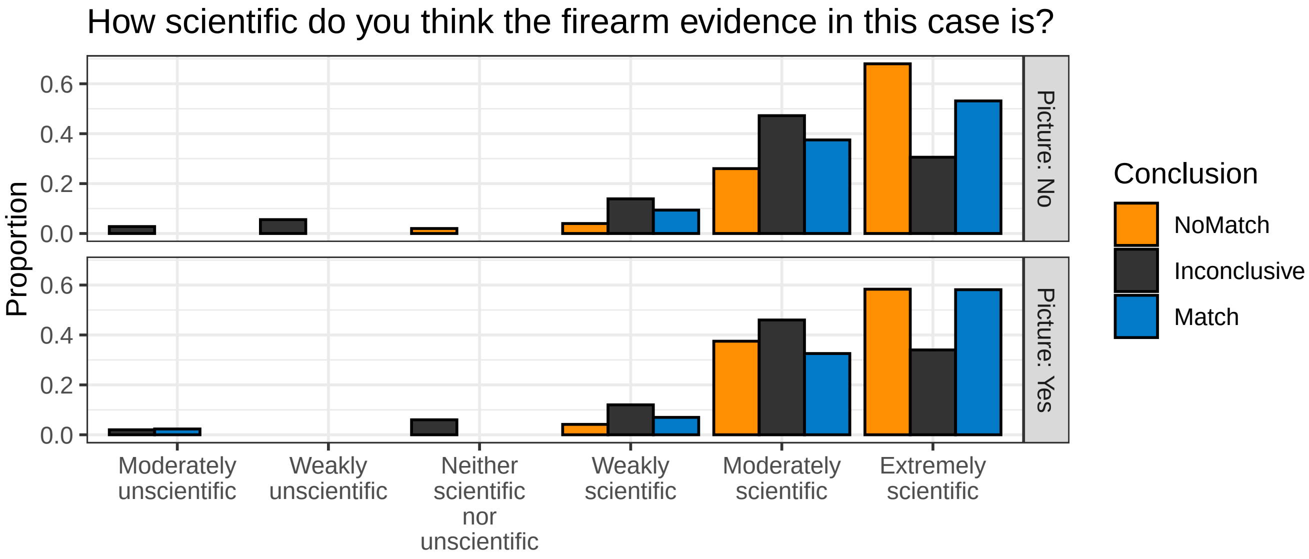

Figure 8: Histogram of perceived overall scientificity in this case

```
## PictureYes              0.02853508  0.1672198  0.1706442 8.645036e-01
## AlgorithmYes            0.32016403  0.1683871  1.9013569 5.725528e-02
```

1.3.2 How scientific do you think the firearm evidence in this case is?

Shown in Figure 8, the top three categories are considered for analysis.

```
##                            Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)               2.6563989  0.5543173  4.7921995 1.649627e-06
## ConclusionInconclusive -1.8349916  0.6316622 -2.9050206 3.672289e-03
## ConclusionMatch        -0.8206576  0.6791730 -1.2083188 2.269247e-01
## PictureYes              0.1948171  0.4913303  0.3965096 6.917292e-01
```

1.3.3 How scientific do you think the firearms examiner's subjective opinion of the bullet comparison is, in this case?

Figure 9 shows this graph of scientificity. There are observations in all three of the top categories, so an ordered logistic regression was used in this case as well.

```
##                             Estimate Std. Error     z value     Pr(>|z|)
## (Intercept)              2.405621778  0.4221794  5.69810255 1.211482e-08
## ConclusionInconclusive -0.961413670  0.4367253 -2.20141513 2.770665e-02
## ConclusionMatch        -0.772645299  0.4410679 -1.75176046 7.981501e-02
## PictureYes              0.009540393  0.3271550  0.02916169 9.767356e-01
## AlgorithmYes           -0.101246442  0.3330589 -0.30398963 7.611358e-01
```

1.3.4 How scientific do you think the firearm algorithm evidence is, in this case?

Figure 10 indicates that the top three categories may be used.

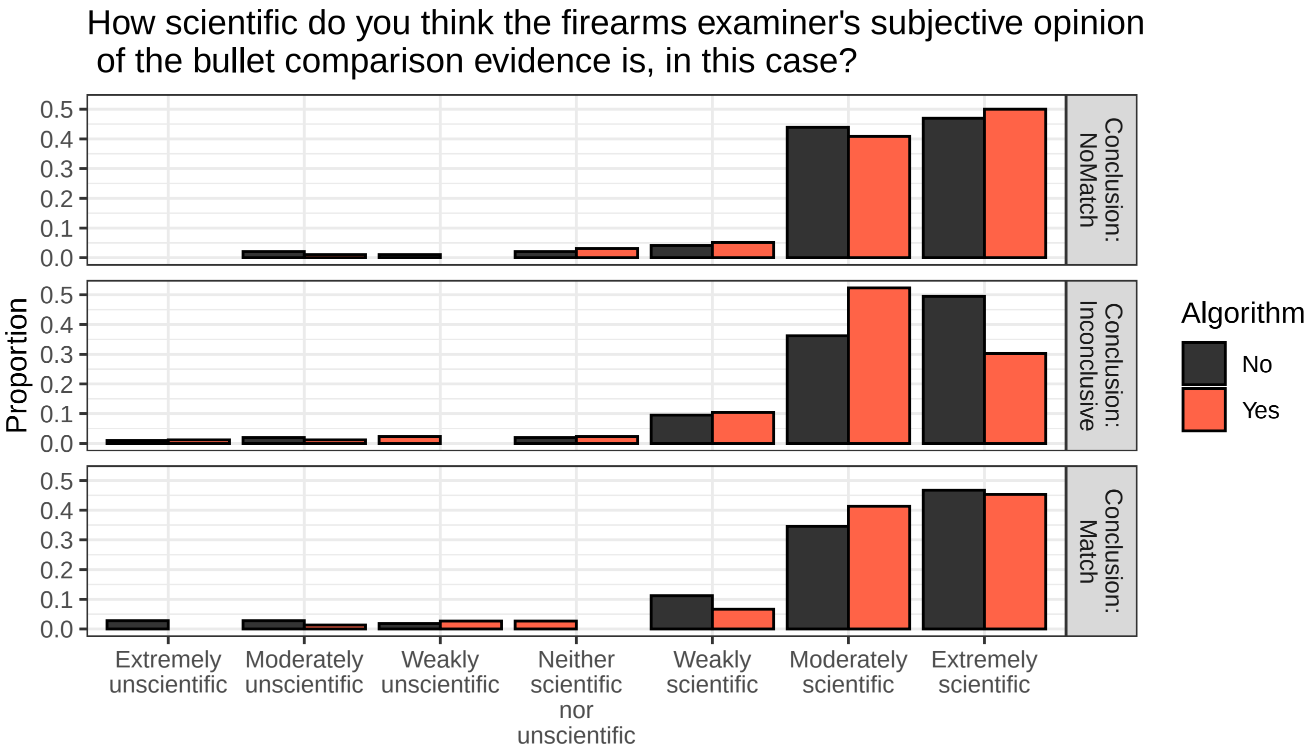

Figure 9: Histogram of perceived scientificity of the bullet comparison of the firearm examiner

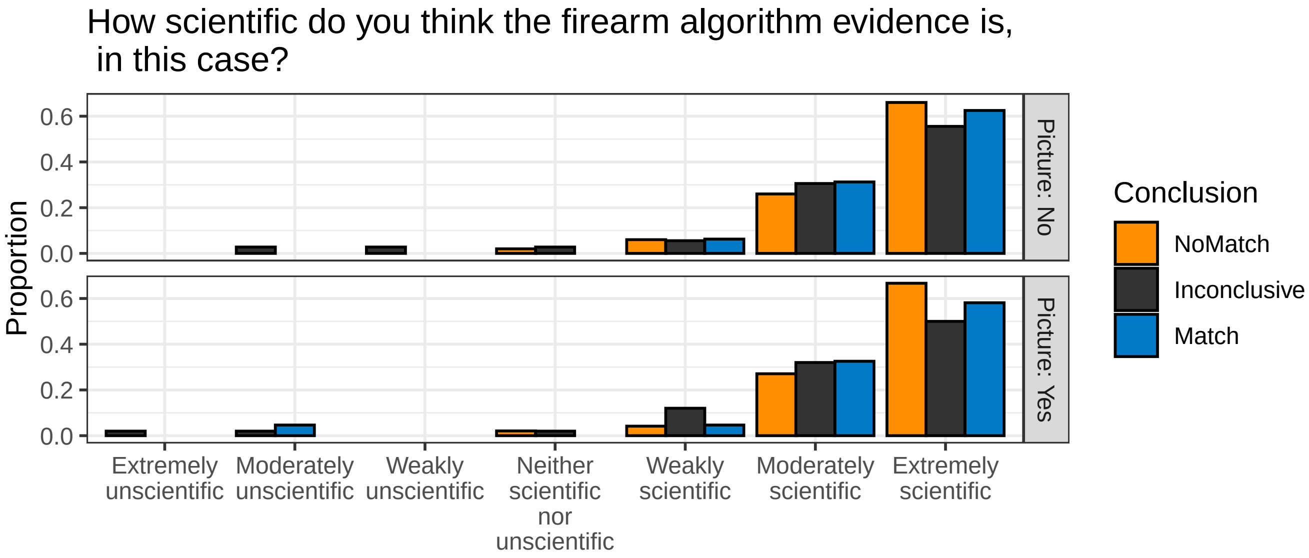

Figure 10: Histogram of perceived algorithm scientificity in this case

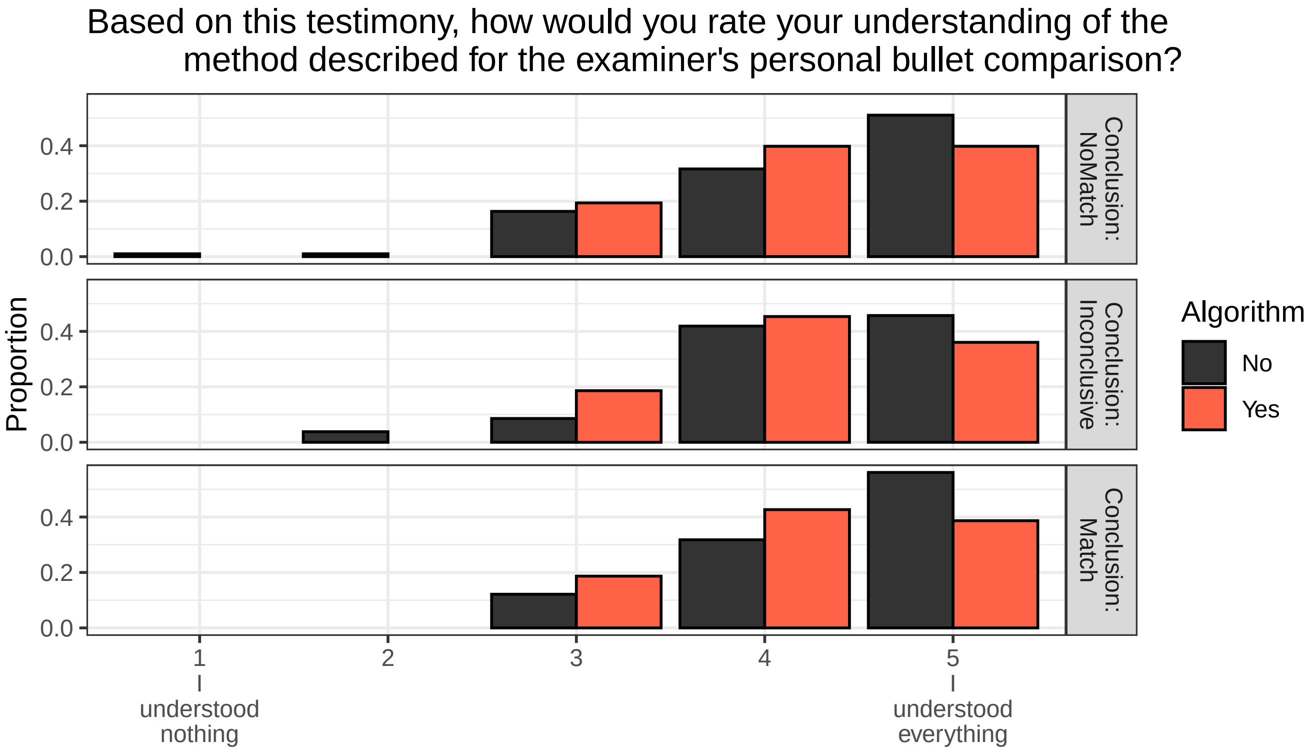

Figure 11: Histogram of understanding for the explanation of the firearms examiner

```
##                              Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)                 2.6593151  0.5396457  4.9278912 8.312189e-07
## ConclusionInconclusive -0.8200085  0.6042670 -1.3570299 1.747717e-01
## ConclusionMatch         -0.1325378  0.6992955 -0.1895304 8.496771e-01
## PictureYes              -0.1866046  0.5234277 -0.3565050 7.214624e-01
```

1.4 Understanding

1.4.1 Based on this testimony, how would you rate your understanding of the method described for the examiner's personal bullet comparison?

The top three categories are used, based on Figure 11.

```
##                              Estimate Std. Error    z value     Pr(>|z|)
## (Intercept):1             2.00624872  0.1997633 10.0431280 9.850013e-24
## (Intercept):2             0.10125841  0.1770278  0.5719916 5.673277e-01
## ConclusionInconclusive -0.03983138  0.1939499 -0.2053695 8.372835e-01
## ConclusionMatch           0.11820234  0.1967775  0.6006902 5.480463e-01
## PictureYes              -0.12340507  0.1603832 -0.7694390 4.416328e-01
## AlgorithmYes            -0.52192165  0.1615615 -3.2304826 1.235814e-03
```

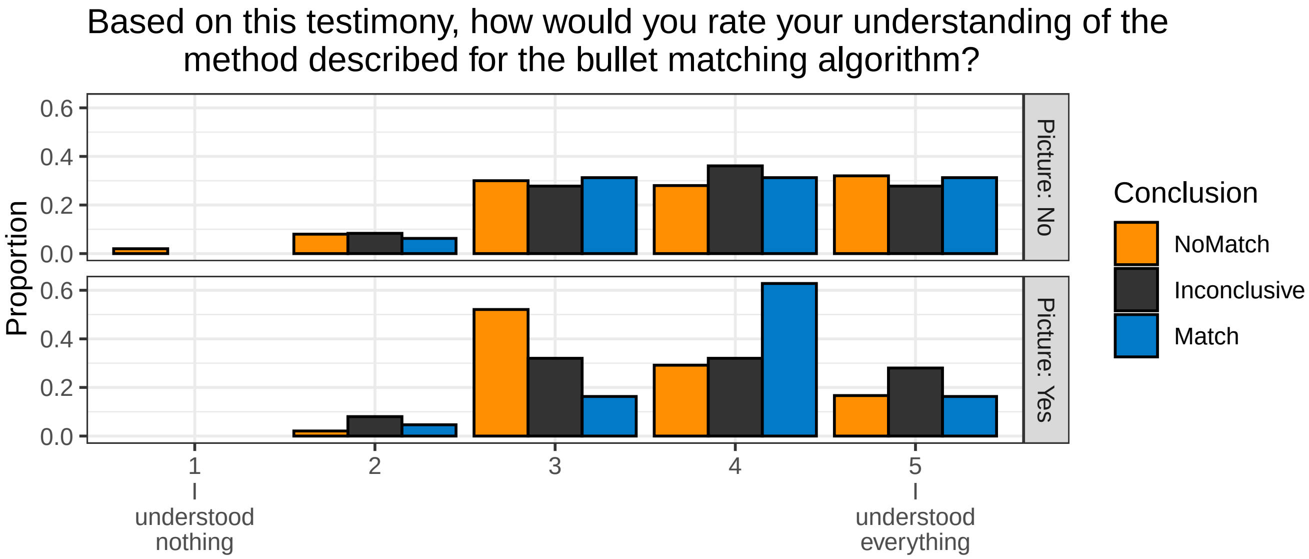

Figure 12: Histogram of understanding for the algorithm explanation

1.4.2 Based on this testimony, how would you rate your understanding of the method described for the bullet matching algorithm?

Here, all but the lowest category of understanding have observations, shown in Figure 12.

```
##                          Estimate Std. Error    z value     Pr(>|z|)
## (Intercept):1            2.6771725  0.3169749  8.4460091 3.014302e-17
## (Intercept):2            0.4191797  0.2248766  1.8640433 6.231563e-02
## (Intercept):3           -1.1599859  0.2357655 -4.9200835 8.650731e-07
## ConclusionInconclusive   0.2285620  0.2714496  0.8420055 3.997849e-01
## ConclusionMatch          0.3907155  0.2822361  1.3843568 1.662492e-01
## PictureYes              -0.2358587  0.2294883 -1.0277590 3.040632e-01
```

1.5 Strength

1.5.1 Strength of Evidence against Cole

Figure 13 shows the participants' percieved strength of evidence against the defendant. All categories are considered in this case.

```
##                          Estimate Std. Error     z value     Pr(>|z|)
## (Intercept):1            0.1736004  0.1720523   1.0089980 3.129756e-01
## (Intercept):2           -0.6988997  0.1756604  -3.9786986 6.929351e-05
## (Intercept):3           -1.4768982  0.1872769  -7.8861717 3.115979e-15
## (Intercept):4           -2.1048885  0.2016071 -10.4405452 1.618784e-25
## (Intercept):5           -2.9139033  0.2231067 -13.0605816 5.530366e-39
## (Intercept):6           -3.2467080  0.2314644 -14.0268143 1.068443e-44
## (Intercept):7           -3.9233876  0.2480721 -15.8155133 2.432194e-56
## (Intercept):8           -5.1296546  0.2882367 -17.7966758 7.499004e-71
```

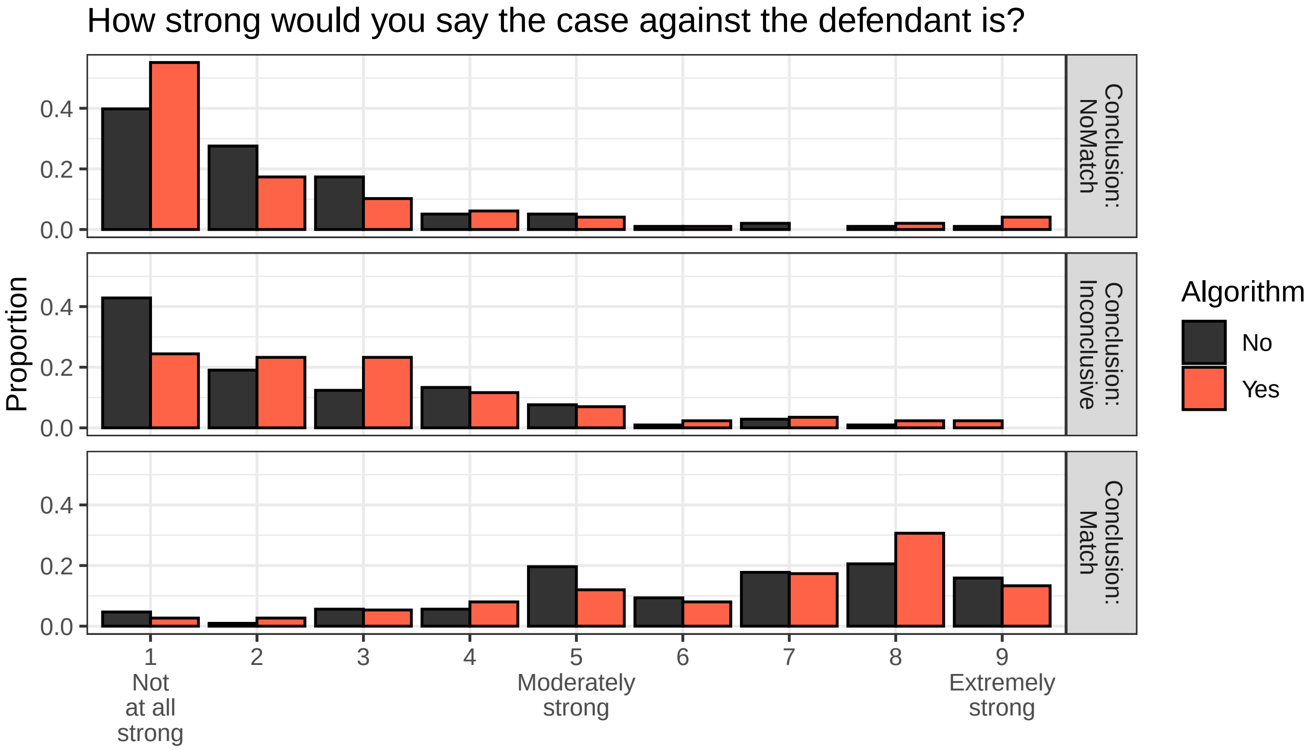

Figure 13: Histogram of perceived strength of evidence against the defendant

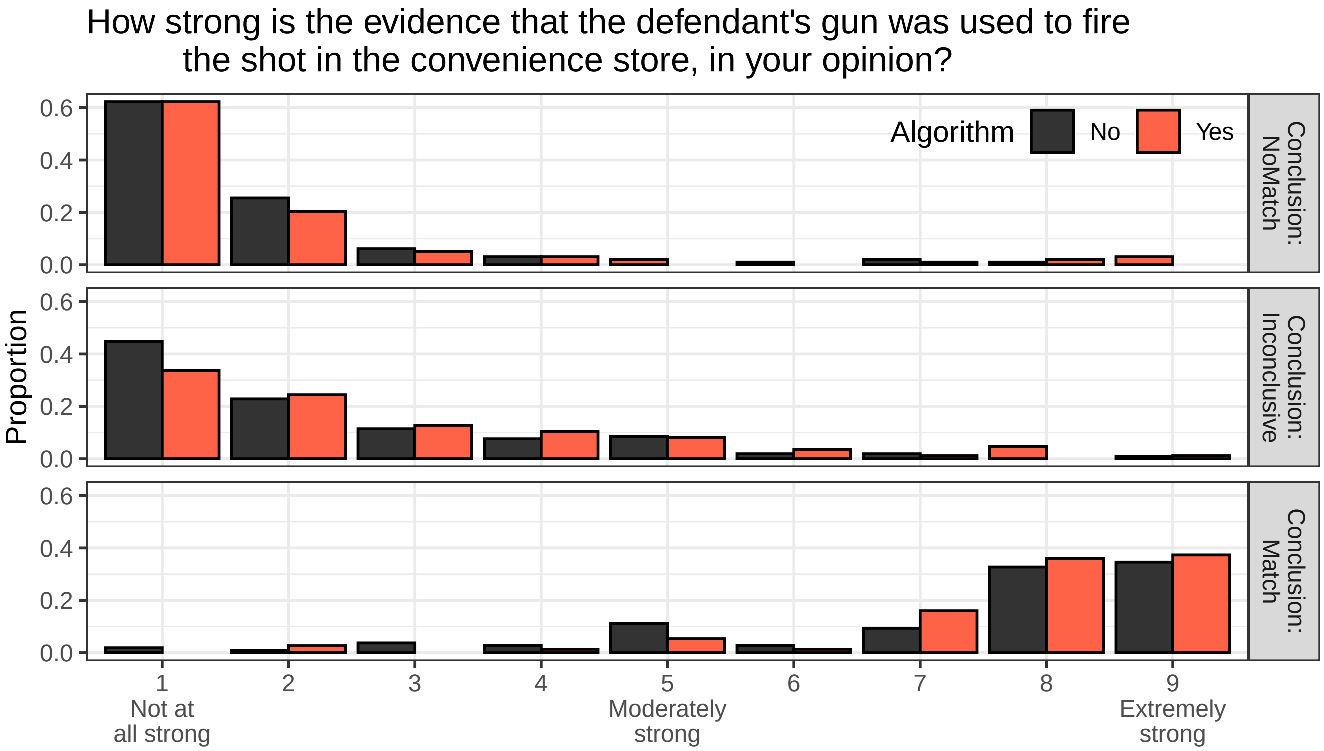

Figure 14: Histogram of perceived strength of evidence against the gun

```
## ConclusionInconclusive  0.5794179  0.1864103   3.1082934 1.881712e-03
## ConclusionMatch         3.6254826  0.2369178  15.3027046 7.334425e-53
## PictureYes             -0.3410194  0.1515947  -2.2495466 2.447774e-02
## AlgorithmYes            0.1443232  0.1521668   0.9484538 3.428985e-01
```

1.5.2 Strength of Evidence Against the Gun

In this case, there are not enough observations for the algorithm condition in each level for non-parallel odds (Figure 14). Therefore, only parallel odds are computed. All levels are included in the model.

```
##                          Estimate Std. Error    z value     Pr(>|z|)
## (Intercept):1          -0.6243571  0.1819157  -3.432123 5.988753e-04
## (Intercept):2          -1.7057502  0.1988435  -8.578354 9.624035e-18
## (Intercept):3          -2.2727735  0.2128996 -10.675328 1.327877e-26
## (Intercept):4          -2.7732129  0.2287528 -12.123186 7.960358e-34
## (Intercept):5          -3.5055672  0.2562690 -13.679247 1.350838e-42
## (Intercept):6          -3.7292055  0.2642989 -14.109801 3.305072e-45
## (Intercept):7          -4.3357854  0.2837070 -15.282619 9.984606e-53
## (Intercept):8          -5.7081607  0.3157797 -18.076403 4.889575e-73
## ConclusionInconclusive  1.0040458  0.1972809   5.089423 3.591552e-07
## ConclusionMatch         5.0772852  0.2918746  17.395433 8.934839e-68
```

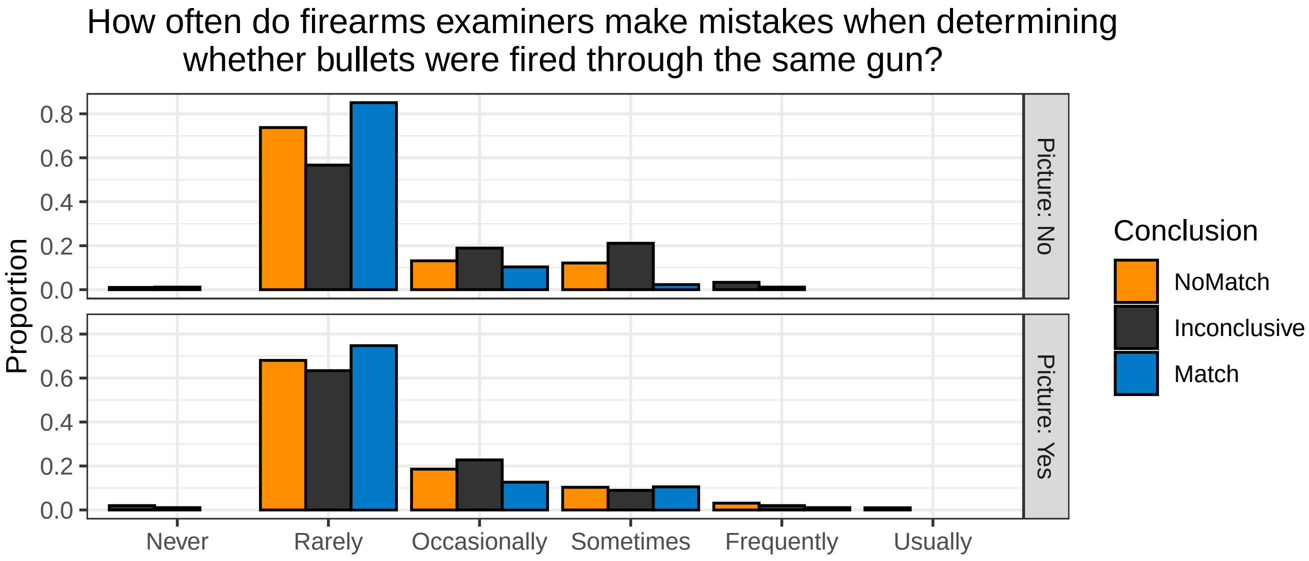

Figure 15: Histogram of perceived frequency of mistakes made by firearms examiners

```
## PictureYes            -0.1014063  0.1574368  -0.644108 5.195054e-01
## AlgorithmYes           0.2987159  0.1587286   1.881929 5.984566e-02
```

1.6 Mistakes

Figure 15 shows the perceived frequency that firearms examiners make mistakes. As can be seen, most individuals selected "Rarely", with little variations between conditions. Very few people selected the extreme values of the scale - "Never" and "Usually". Only the values of "Rarely", "Occasionally", and "Sometimes" are considered in the analysis.

```
##                           Estimate Std. Error      z value     Pr(>|z|)
## (Intercept):1          -1.30815117  0.2141733  -6.10790967 1.009445e-09
## (Intercept):2          -2.46653570  0.2386186 -10.33673008 4.806592e-25
## ConclusionInconclusive  0.47767483  0.2209376   2.16203507 3.061547e-02
## ConclusionMatch        -0.46577844  0.2516942  -1.85057254 6.423107e-02
## PictureYes              0.01634266  0.1914690   0.08535408 9.319799e-01
## AlgorithmYes            0.62735090  0.1924081   3.26052189 1.112074e-03
```

2 Beta Distributions

2.1 Probability

The graphs of probability are shown in Figure 16. In both cases, the only significant effect is that of conclusion. This analysis uses the 'gam' function from the 'mgcv' package.

2.1.1 Probability Cole Committed the Crime

```
##   model term                   df1 df2 F.ratio p.value
```

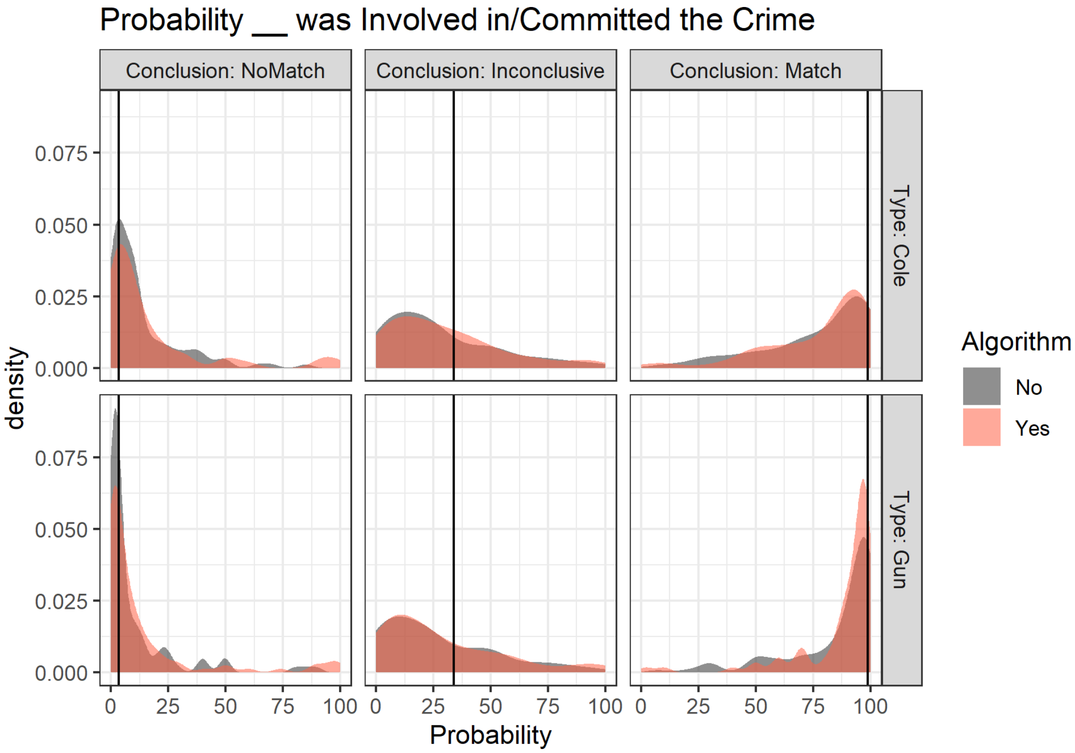

Figure 16: Probability the gun was used in the crime, or that Cole committed the crime. Black lines indicate bullet match scores for the algorithm.

```
##  Conclusion                     2 557 263.793  <.0001
##  Picture                        1 557   0.209  0.6481
##  Algorithm                      1 557   0.007  0.9327
##  Conclusion:Picture             2 557   1.518  0.2201
##  Conclusion:Algorithm           2 557   0.422  0.6557
##  Picture:Algorithm              1 557   0.585  0.4446
##  Conclusion:Picture:Algorithm   2 557   0.149  0.8620
```

2.1.2 Probability the Gun was Involved in the Crime

```
##  model term                   df1 df2 F.ratio p.value
##  Conclusion                     2 557 410.083  <.0001
##  Picture                        1 557   1.619  0.2038
##  Algorithm                      1 557   0.100  0.7514
##  Conclusion:Picture             2 557   0.660  0.5175
##  Conclusion:Algorithm           2 557   1.150  0.3175
##  Picture:Algorithm              1 557   0.692  0.4059
##  Conclusion:Picture:Algorithm   2 557   0.538  0.5845
```

3 Binomial Responses

3.1 Do you think guns leave unique markings on discharged bullets/casings?

Responses were recorded as in a yes/no format. 16 individuals indicated that they did not think think firearms left unique markings, out of 569 total responses.

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: unique_num
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                         568     145.83
## Conclusion  2  2.12000       566     143.71   0.3465
## Picture     1  2.88875       565     140.82   0.0892 .
## Algorithm   1  0.00591       564     140.81   0.9387
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

3.2 Conviction

Individuals were given the following question: “The State has the burden of proving beyond a reasonable doubt that the defendant is the person who committed the alleged crime. If you are not convinced beyond a reasonable doubt that the defendant is the person who committed the

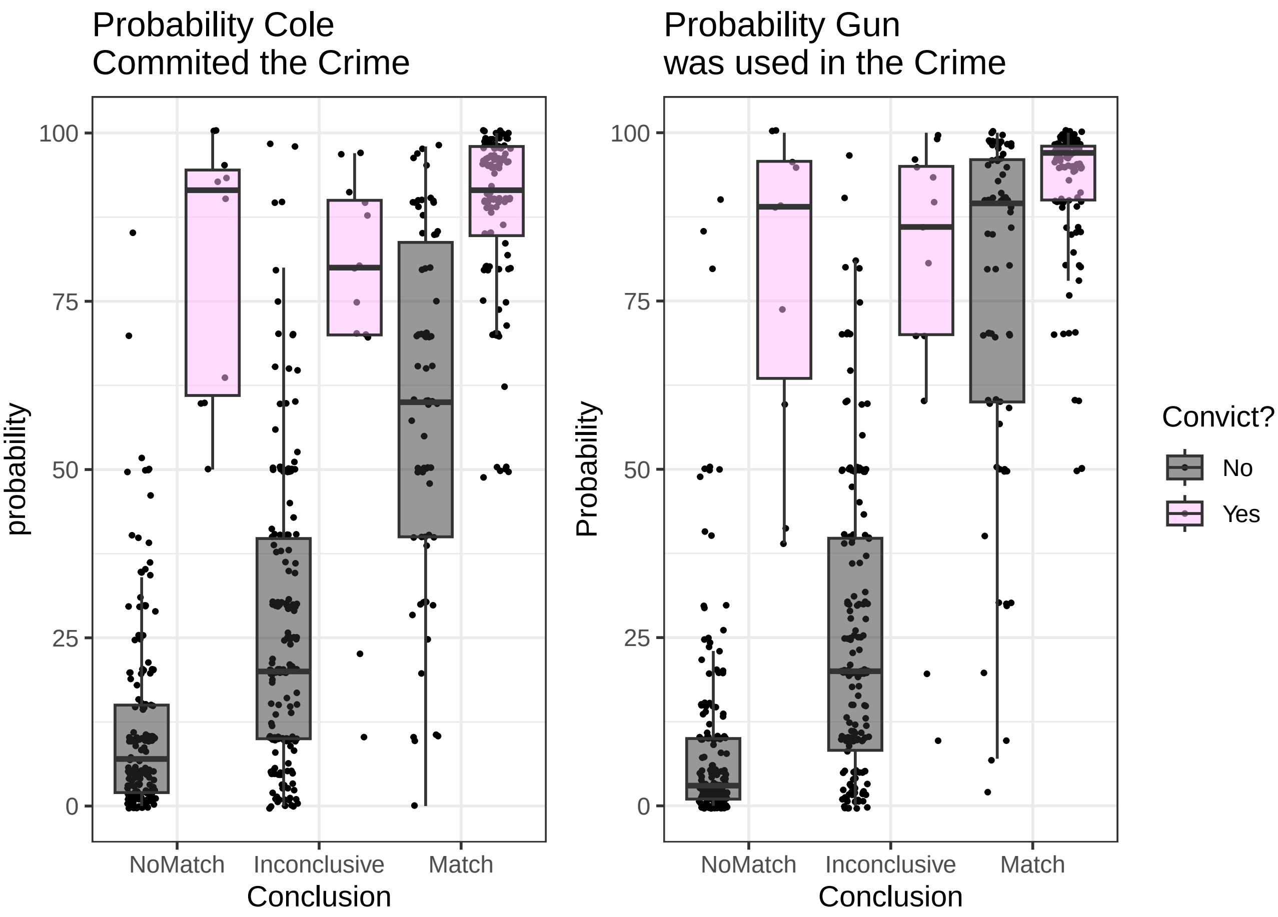

Figure 17: Probabilities based on whether the participants thought the defendant was guilty

alleged crime, you must find the defendant not guilty. Would you convict this defendant, based on the evidence that you have heard?". Results are shown in Figure 17.

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: guilt_num
##
## Terms added sequentially (first to last)
##
##
##                              Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                                          568     623.51
## Conclusion                    2  207.027       566     416.48  < 2e-16 ***
## Picture                       1    3.301       565     413.18  0.06924 .
## Algorithm                     1    3.656       564     409.53  0.05588 .
## Conclusion:Picture            2    0.125       562     409.40  0.93927
## Conclusion:Algorithm          2    5.088       560     404.31  0.07855 .
## Picture:Algorithm             1    0.494       559     403.82  0.48225
## Conclusion:Picture:Algorithm  2    4.215       557     399.60  0.12156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```