Fingerprint Analysis Replication

Tianyi Lan, Jacqueline Liu

April 12, 2018

0. Abstract

Looking at the results of a survey regarding fingerprint evidence in criminial trials, we replicate conclusions drawn in *Certainty and Uncertainty in Reporting Fingerprint Evidence* (Kadane and Koehler), namely that the language used to present the evidence influences the strength with which respondents view it. Further analysis showed that an individual's inherent likelihood of convicting is correlated with their political views, any jury or law experience they have had, the frequency with which they watch CSI or similar shows, and their engagement with the survey. A comparison between complete and incomplete survey responses found no significant differences between the two, though the small sample size made it hard to make any strong claims.

1. Introduction

Fingerprint analysis is often used in criminal trials as evidence for identification of the culprit. However an issue with this is that laypeople, namely jurors and judges, see fingerprint evidence as conclusive when in fact it is not. Fingerprint analysis looks at certain characteristics such as ridges, forks, loops, and arches; a "match" is declared when two prints have the same features. Problematically, there is no measure of the number of people who would "match" with the fingerprint of interest; it could be that the defendant has a very unique fingerprint and is the only person who could have left such a print at the crime scene. It could also be that half the world's population share the same characteristics and are a match. Expressing this uncertainty during a court case is of utmost importance which makes presentation language vital in criminal trials.

To better understand how the language used to present fingerprint evidence influences jurors' confidence in the evidence, Kadane and Koehler conducted an online survey in which potential jurors (American citizens over 18 years old with no felony history) read a hypothetical legal case and answered questions about their perception of evidence strength of the fingerprint testimony. Their findings are discussed in *Certainty and Uncertainty in Reporting Fingerprint Evidence*.

2. Replication

2.1. Methods

First we will attempt to replicate the results of Kadane and Koehler's paper, using the same sample of survey respondents they did. Their original work studied "strength of evidence", i.e. the certainty in the fingerprint match instilled in the respondents by the evidence presented.[1]

2.2. Results

Kadane and Koehler had concluded that there was no significant effect of cross-examination on the respondent's perceived strength of evidence. They had six unique phrasings with which the expert witness would present fingerprint evidence. Each phrasing was shown with and without cross-examination testimony, leading to 12

[1] "Strength of evidence" was measured through a series of six questions which were worded along the lines of "How confident are you that the fingerprint on the cash register tray was left by the defendant?". The final score was an average of the six responses when each were binned on a 1-7 scale.

different total treatment types, hereafter referred to as "conditions" (see Figure 1[2]). This means that across pairs of conditions (1 and 7, 2 and 8, etc. . .), the same language is used to present the evidence, differing only in that the first in the pair includes cross-examination.

	Condition	
Expert Testimony	**Cross-exam**	**No Cross-exam**
Cannot exclude Mr. Johnson	1	7
The likelihood of observing this amount of correspondence when two impressions are made by different sources is considered extremely low	2	8
Mr. Johnson is the source	3	9
Mr. Johnson is the source to a reasonable degree of scientific certainty	4	10
I effected an individualization on that print to Mr. Johnson	5	11
I effected an individualization on that print to Mr. Johnson to the exclusion of all possible other sources in the world	6	12

Figure 1: Twelve Conditions

conditions	estimate	p-value	lower bound	upper bound
1 vs. 7	0.21	0.53	-0.85	0.44
2 vs. 8	0.55	0.12	-1.25	0.15
3 vs. 9	0.13	0.72	-0.83	0.58
4 vs. 10	0.20	0.59	-0.95	0.54
5 vs. 11	-0.10	0.80	-0.65	0.85
6 vs. 12	0.19	0.60	-0.88	0.51

Table 1: Studying the effects of cross-examination

Pairwise t-tests showed that perceived strength of evidence was not significantly influenced by the absence/presence of cross-examination. All p-values were much too large to be significant (see Table 1), so we collapsed the 12 conditions into 6, no longer differentiating between conditions with and without cross-examination. This decision mirrors Kadane's in his original paper. Along with the effect of cross-examination, we were able to replicate all conclusions and plots in his original paper except for the one which plots strength of evidence by combined 6 conditions. Figure 2[3] compares the plot presented in Kadane's paper with the one generated in our own replication. For visualization benefits, condition 6 was actually moved up to the position of condition 2 so that there seems to an increasing trend. However, a few of the boxes in our replication seemed to still have different ranges and medians. All other figures were reproduced successfully, so the discrepancy for this one in particular remains unknown.

3. Conviction proneness for complete responses

3.1. Methods

Amongst all the features of the survey participants, Kadane had concluded that there was a strong relationship between perceived strength of evidence and conviction proneness, and that those with stronger concern for convicting the guilty would assign more weight to fingerprint evidence. Thus, for this part of the analysis, we will look into the "conviction proneness" of each respondent, a measure of how likely they are to convict a defendant based on their level of agreement with the following statement:

[2]Taken from Kadane's paper.
[3]Left figure taken from Kadane's paper.

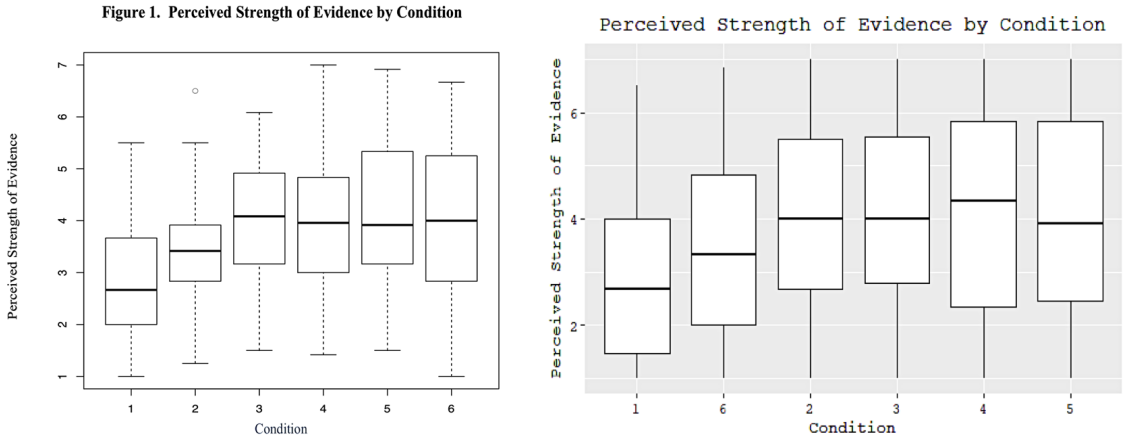

Figure 2: Different Plots from Replication

> Our criminal justice system should be less concerned about protecting the rights of the people charged with crimes and more concerned about convicting the guilty.[4]

This variable will now be treated as the response in our models. Specifically, as it has five levels ranging from Strongly Agree to Strongly Disagree, an ordinal logistic regression model will be used to identify predictive factors of conviction proneness for complete responses in the data set.

3.2. Results

We first looked at some EDA results. The boxplot shown in Figure 3 appeared to have decreasing trend: respondents who were more concerned about convicting the guilty assigned more weight to the fingerprint evidence. This was consistent with Kadane's conclusion.

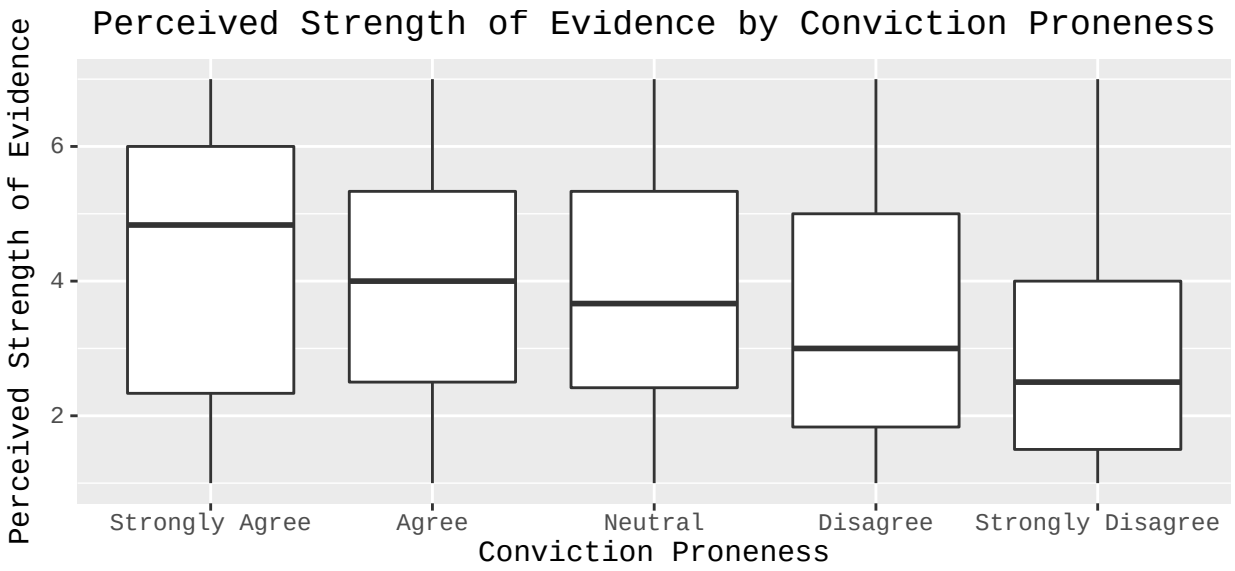

Figure 3: Perceived Strength of Evidence vs. Conviction proneness

[4]The response to this question is one of the five: strongly agree, agree, neutral, disagree, or strongly disagree. Participants marking Strongly Agree or Agree to the conviction proneness question showed a strong concern for convicting the guilty, whereas participants with Disagree or Strongly Disagree response corresponded to a strong concern for protecting defendant's rights (and therefore were less likely to convict).

Variable	Chi-Square Value	p-value
Political view	28.50	0.00
CSI frequency	44.20	0.00
Jury experience	10.40	0.04

Table 2: Predictors which are correlated with conviction proneness

Conviction proneness was the only predictor that was strongly correlated with their perceived strength of evidence. Interested in what types of people are more likely to convict, we conducted chi-square tests on all participant features against conviction proneness as our new response variable. Results showed that political view, CSI watching frequency, and jury experience are the three significant predictors of conviction proneness (Table 2)

Using conviction proneness as the new response variable with forward selection by AIC, the best model of conviction proneness included five predictors with no interactions: political leaning, CSI (or other criminal TV shows) watching frequency, jury experience, experience with law enforcement and engagement. The user engagement variable was devised from the comment section of the survey responses where leaving longer comments would correspond to more engaged participants. This model yielded an AIC of 1783 and a BIC of 1849.

Coefficient	Values
politicalsomewhat conservative	-0.53
politicalsomewhat liberal	-0.62
politicalliberal	-0.93
csiOnce every few months	0.07
csiOnce a month	0.44
csiOnce a week	0.69
csiSeveral times / week	0.93
juryYes	-0.31
lawExpYes	-0.28
engagedfFalse	0.14
engagedfTrue	0.40

Table 3: Coefficients of Ordinal Regression

Figure 4(a) plots the cumulative probabilities of falling into each category of conviction proneness for an participants the baseline traits. Specifically, for a conservative person with no law enforcement or jury experience who rarely or never watches CSI or other similar shows and left no comments, the probability he or she answers “strongly agree” or “agree” (inclined to convict) is 0.41 with only a 0.26 probability of answering “strongly disagree” or “disagree”. Figure 4(b) shows that someone with the same traits but identifies as liberal has now only 22% chance of answering “strongly agree” or “agree”, with a 47% chance of not convicting. On the other hand, Figure 4(c) shows watching CSI increases the likelihood an individual is conviction prone, going from 0.41 to 0.64 probability that they agreed with the conviction proneness statement. Looking at Table 3, we can see all the coefficient values. Like being liberal, having jury and/or law experience decreases one’s conviction proneness, though to a lesser extent. User engagement increases it, though not as dramatically as watching CSI shows one a month or more.

3.3. Conclusion & implications

Political views and CSI watching frequency seem to be the strongest predictors of conviction proneness, changing the likelihood of an individual being conviction prone by more than 15 percentage points. In addition, political views, jury experience and law enforcement experience decrease the probability of convicting whereas user engagement and CSI watching frequency increase the probability an individual is inclined to convict. As some of the predictive factors are not commonly expected and noticed, by noticing the significant effects of

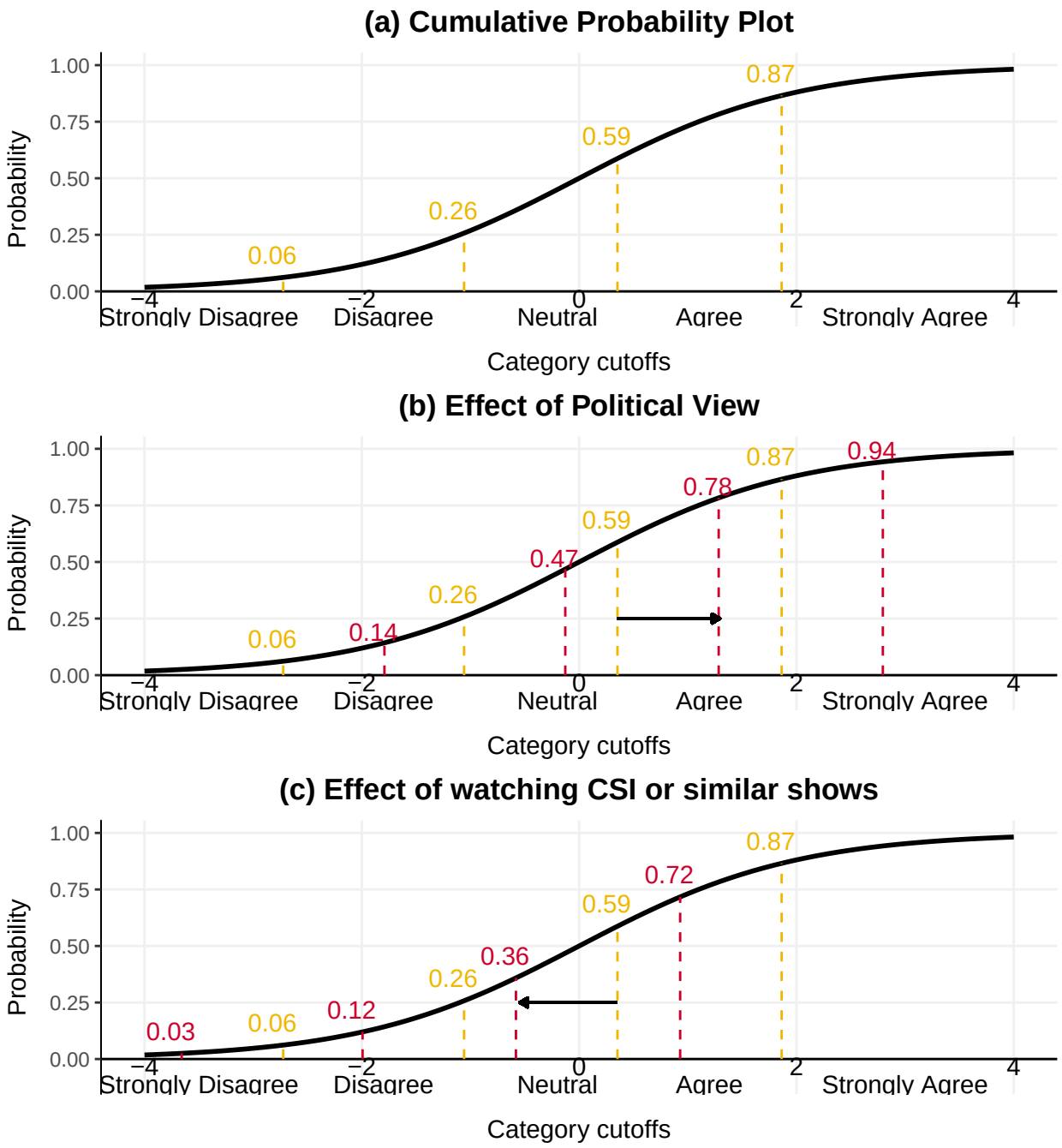

Figure 4: Effects of Predictors on Conviction Proneness

those predictors as suggested by the results, one may design more strictly controlled experiments/surveys for analyzing perceived strength of evidence and these results could therefore shed some light on how to select fairer jurors for future criminal cases.

4. Incomplete data

4.1. Methods

We will also identify survey respondents who qualified to be in the analysis [5] except for the fact that they did not finish the survey. We will study if and how this population differs from the one used in Kadane and Koehler's analysis. We will also study if the engagement of the participants is correlated with either the original outcome of interest, strength of evidence, or the new response, conviction proneness. We'll measure user engagement by the answers in the comments sections.

4.2. Results

We studied the data and data dictionary to filter through the samples to understand how the 600 survey participants were identified. We then looked at the breakdown of all participants (see Figure 5) and found our sample of interest (those who were eligible to be in the study but did not complete the survey) included only 3 participants. This extremely small sample size limits our options for future analysis. Furthermore, there is one participant who met all study requirements and yet was given a label of "screened" rather than being included in the study. Because the data was labeled by an outside company, we are unable to identify the reason for this discrepancy.

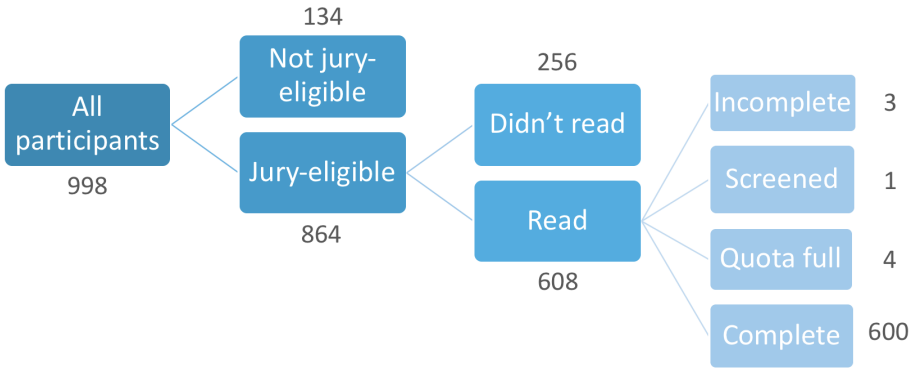

Figure 5: Missing Data Analysis

Question 22 asked "Do you have any comments for the researchers regarding this survey?". 395 of the 600 participants wrote a response, varying from "No" to in-depth discussions of the legal system. Using regular expressions, we divided the participants into three categories: no response, unengaged, and engaged, where unengaged comments were defined as those under 15 characters with leading with "no", "nothing", or "n/a". This left us with 205 non-responses, 198 not engaged, and 197 engaged. We also studied user engagement as a continuous variable - the number of characters in the comment. The distribution is very right skewed, with 0 being the most common length by far and 2 being the median.

Looking into the distribution of our missing data in Table 4, we see that all three participants answered the questions regarding the strength of evidence (questions 3 through 8). Figure 6 shows the distribution of the strength of evidence scores from the 600 complete responses, with the vertical lines denoting the scores of the incomplete responses. Because the complete distribution is so widely spread, we cannot conclude that the incomplete scores are extreme in any way. Plotting the distribution by the condition assigned (2 for "Alice", and 3 for "Bob" and "Eve") again shows little deviance from the expected distribution. Given the incomplete

[5] Respondents were filtered out through a series of requirements. They needed to complete the survey, be eligible for jury duty (namely 18+ years of age, no felony record, and U.S. citizens) and have had actually read the material and answer questions in a meaningful manner. The latter most requirement was held by asking two reading comprehension questions; only respondents who answered correctly could be included in the analysis.

	Alice	Bob	Eve
condition	2.00	3.00	3.00
q3	4.00	3.00	7.00
q4	4.00	3.00	7.00
q5	4.00	3.00	7.00
q6	50.00	50.00	100.00
q7	4.00	3.00	7.00
q8	50.00	50.00	99.00
q9	4.00		7.00
q10	4.00		6.00
q11	4.00		1.00
q12			1.00
q14			1.00
q15			2.00
q16			4.00
q17			1.00
q18			

Table 4: Incomplete Responses

responses came from different conditions and ended at different questions in the survey, there does not seem to be a systemic reason for the dropout.

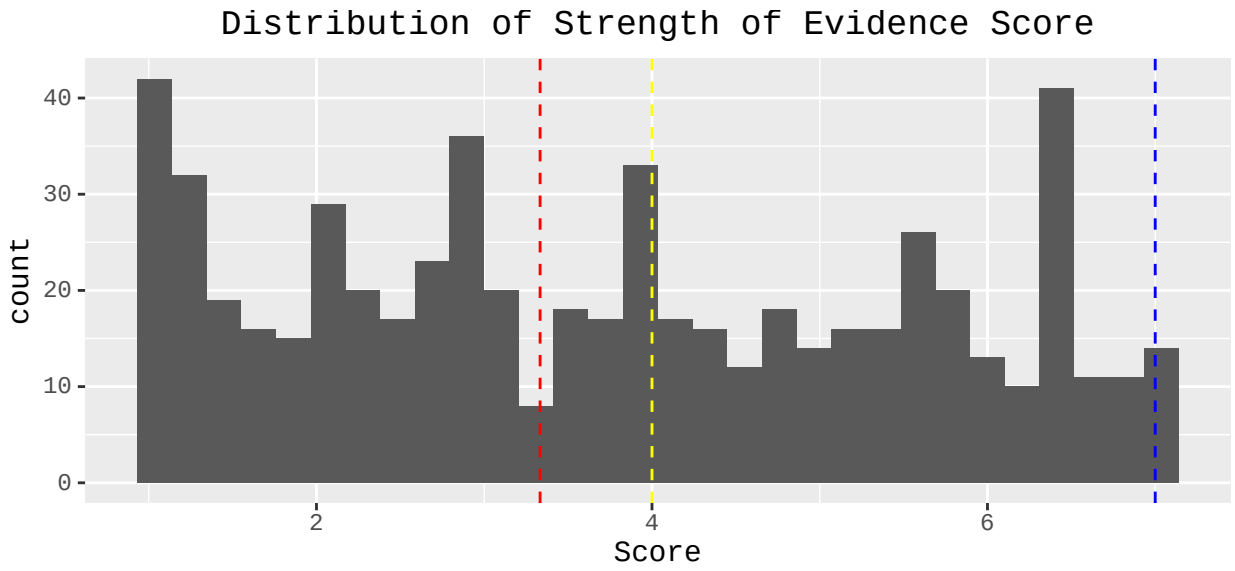

Figure 6: Distribution of Strength of Evidence Score

A closer look at Table 4 shows that Alice, Bob and Eve were very consistent in their strength of evidence scores, giving the same rating on each of the six questions. This is consistent with the complete responses where 42 consistently gave the lowest score possible, 17 gave the middle rating like Alice and Eve, and 14 gave the highest score possible like Bob. Given this makes up over 10% of our participants, again we do not think Alice, Bob or Eve were outliers in their response patterns. Bayesian hypothesis testing with a null hypothesis that the mean strength of evidence scores of the incomplete responses is equal to that of the complete responses against the alternative that they are not equal gives weak evidence in favor of the null:

$$BF = \frac{L(X_{incomplete}|H_a)}{L(X_{incomplete}|H_0)} = 0.629$$

Notice that rather than using the alternative hypothesis with the maximum likelihood, Bayesian hypothesis

testing integrates across all possible alternatives, weighted by some prior distribution. Because the Bayes factor, which represents the shift between the prior and posterior likelihoods, is so close to 1 that we cannot draw any conclusions about the data.

Looking at the relationship between strength of evidence score and user engagement, we see in Figure 7 that the distribution of the scores do differ. Those who were not engaged seemed to have found the evidence less convincing than those who were engaged, while participants who left no comment clustered more in the middle. The Kolmogorov-Smirnov test, however, concludes that the distributions are not significantly different from one another. An ANOVA test finds the difference in means are also not statistically significant. The length of comments also seems to have no correlation with strength of evidence score or conviction proneness.

4.3. Conclusion

We found no evidence suggesting that engagement is correlated with strength of evidence score, treating comments as categorical or continuous variable. There is weak evidence that the incomplete responses may have the same mean as the complete ones, suggesting no difference in their distributions.

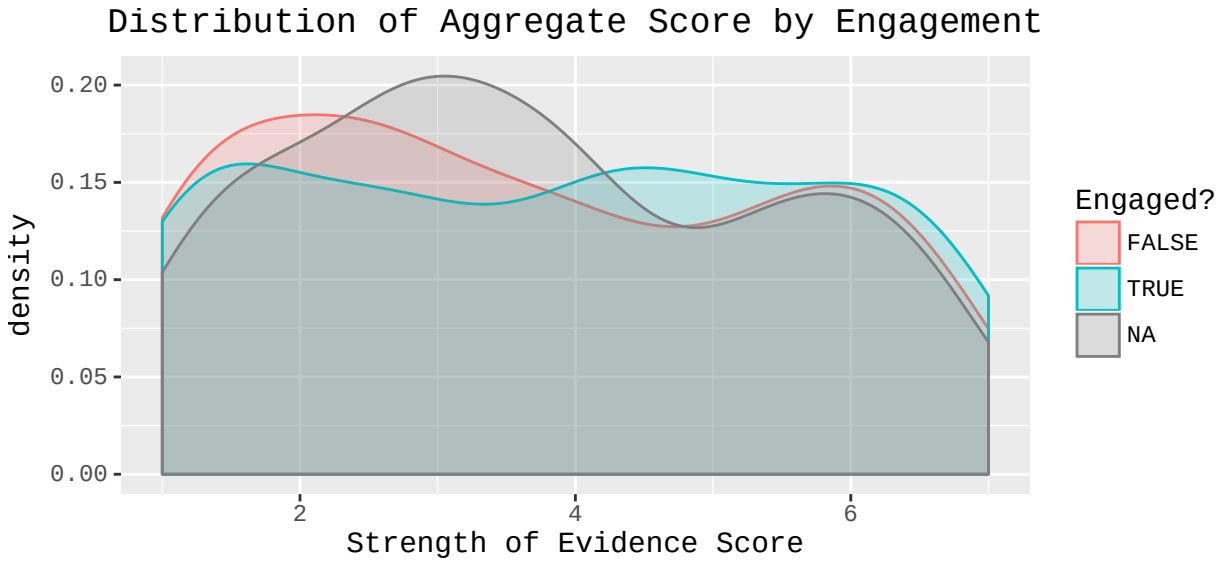

Figure 7: Distribution of Aggregate Score by Engagement

5. Future steps

To analyze whether the distributions of perceived strength of evidence differ between the complete responses (600 participants) and the incomplete group (3 participants), Bayesian hypothesis testing was used with the default Cauchy prior for the alternative hypothesis centered at 3.75 with a scale of $\sqrt{2}/2 \approx 0.707$. However, a prior that is more spread out and with values truncated between 1 and 7 should be more appropriate for our data. Thus a new prior could yield different results and possibly stronger evidence for whether the distributions of perceived strength of evidence differ between the complete vs. partial responses. In addition, in this analysis, participants that answered incorrectly on the reading comprehension questions were filtered out. However, there are no such criterion in real-life jury selections, so such exclusion might lead to biased results. In future analysis, with possibly a slightly different design of the survey, more demographic information would be available for partial responses and those who failed the reading comprehension questions could be potential valid respondents for the analysis.