1-30-2021

Using mixture models to examine group difference among jurors: an illustration involving the perceived strength of forensic science evidence

Naomi Kaplan-Damary
University of California, Irvine

William C. Thompson
University of California, Irvine

Rebecca Hofstein Grady
University of California, Irvine

Hal S. Stern
University of California, Irvine

Using mixture models to examine group difference among jurors: an illustration involving the perceived strength of forensic science evidence

Abstract

The way in which jurors perceive reports of forensic evidence is of critical importance, especially in cases of forensic identification evidence that require examiners to compare items and assess whether they originate from a common source. The current study discusses methods for studying group differences among mock jurors and illustrates them using a reanalysis of data regarding lay perceptions of forensic science evidence. Conventional approaches that consider subpopulations defined *a priori* are compared with mixture models that infer group structure from the data, allowing detection of subgroups that cohere in unexpected ways. Mixture models allow researchers to determine whether a population comprises subpopulations that respond to evidence differently and then to consider how those subpopulations might be characterized. The reanalysis reported here shows that mixture models can enhance understanding of lay perceptions of an important type of forensic science evidence (DNA and fingerprint comparisons), providing insight into how the perceived strength of that evidence varies as a function of the language forensic experts use to describe their findings. This novel application of mixture models illustrates how such models can be used, more generally, to explore the importance of juror characteristics in jury decision making.

Keywords

forensic science, testimony, reporting, identification, probability, evidence

Disciplines

Forensic Science and Technology

Comments

The following article is published as Kaplan-Damary, Naomi, William C Thompson, Rebecca Hofstein Grady, and Hal S. Stern. "Using mixture models to examine group difference among jurors: an illustration involving the perceived strength of forensic science evidence." *Law, Probability and Risk* (2021). Posted with permission of CSAFE.

doi:10.1093/lpr/mgaa016

Using mixture models to examine group difference among jurors: an illustration involving the perceived strength of forensic science evidence

Naomi Kaplan-Damary, William C. Thompson, Rebecca Hofstein Grady and Hal S. Stern
University of California, Irvine, CA, USA

[Received on 11 May 2020; revised on 14 November 2020; accepted on 24 November 2020]

The way in which jurors perceive reports of forensic evidence is of critical importance, especially in cases of forensic identification evidence that require examiners to compare items and assess whether they originate from a common source. The current study discusses methods for studying group differences among mock jurors and illustrates them using a reanalysis of data regarding lay perceptions of forensic science evidence. Conventional approaches that consider subpopulations defined *a priori* are compared with mixture models that infer group structure from the data, allowing detection of subgroups that cohere in unexpected ways. Mixture models allow researchers to determine whether a population comprises subpopulations that respond to evidence differently and then to consider how those subpopulations might be characterized. The reanalysis reported here shows that mixture models can enhance understanding of lay perceptions of an important type of forensic science evidence (DNA and fingerprint comparisons), providing insight into how the perceived strength of that evidence varies as a function of the language forensic experts use to describe their findings. This novel application of mixture models illustrates how such models can be used, more generally, to explore the importance of juror characteristics in jury decision making.

Keywords: forensic science; testimony; reporting; identification; probability; evidence

1. Introduction

The use of lay jurors as triers-of-fact has been a central element of the Anglo-American system of justice and is becoming more common in Europe as well (Bradley, 1996). Lay jurors are said to bring the 'conscience of the community' to fact-findings, thus assuring that the outcome of trials reflects the coalescence of a diverse range of values and opinions. Concerns have been raised, however, about the ability of lay jurors to understand the evidence, particularly in trials that turn on scientific testimony, such as criminal trials where forensic science is used to identify the defendant as the perpetrator (Thompson, 2018). Three major areas of research in this context focus respectively on the role of juror characteristics, expert witness characteristics and the language used to describe the evidence in understanding jury perception of forensic testimony. In this introduction, we briefly review findings in these areas and describe how they motivate the present study which focuses on identifying group differences among jurors that may affect their evaluations of the language used by forensic examiners. The research described here is especially timely given ongoing discussions in

the forensic science community (e.g. in the Organization of Scientific Area Committees for Forensic Science[1]) about appropriate language for conclusions.

Considerable work has been done on the relationship between jury characteristics and jury decision-making. MacCoun (1989), in his survey of experimental research on jury decision-making, looked at the influence of jurors' observable pre-trial characteristics on their verdicts, and concluded that the effect was negligible. A more comprehensive survey by Devine *et al*. (2001) of 206 studies on factors associated with verdicts carried out between 1955 and 1999, focused on four primary categories of factors: (a) procedural characteristics, (b) participant characteristics, (c) case characteristics, and (d) deliberation characteristics. The participant characteristics considered included personality traits (such as authoritarian tendencies), demographic features (gender, socio-economic status, race, education, age), social attitudes, self-perception and self-assurance. Again it was found that few if any juror characteristics were useful predictors of juror verdict preferences. Similarly, studies that look specifically on lay reactions to the language used by forensic scientists (our focus in this article) have found little evidence of group differences. The few variables that predict outcomes in some studies (e.g. subjective numeracy, Scurich (2015); gender and confidence in math skills, Thompson *et al*. (2013)), have not consistently been predictive in other studies (e.g. Thompson and Newman, 2015; Martire *et al*. 2013; 2014).

However, all previous studies of juror characteristics have been 'hypothesis-driven'. In other words, they tested hypotheses generated in advance by the researchers about what characteristics might be important or what group differences might exist. This traditional research strategy may miss sub-group structure in the data that researchers fail to anticipate because the groups cohere in unexpected ways. The study reported here shows how this limitation can be addressed with a 'data-driven' approach to exploring group differences. Using statistical methods that are novel in this context, our approach infers the existence of groups from the data itself, allowing detection of coherent subgroups that might otherwise not have been detected. To our knowledge, this is the first article to use this approach in the study of jury decision-making. Our approach uses mixture models, a class of statistical models that test whether the data are more consistent with a model assuming subpopulations than a model that assumes homogeneity (no subpopulations) using a Monte Carlo approach (Lindsay and Lesperance, 1995; Morduch and Stern, 1997). To illustrate this approach, we reanalyse data from a recent study of lay evaluations of the language used by forensic scientists to characterize the strength of their findings (Thompson *et al*., 2018a), showing how our novel application of statistical methods can be used to provide fresh insights.

Numerous studies have examined how the language used by an expert witness may impact on the jury. O'Barr (1982) found that witnesses who employed 'powerful speech' and thus exuded confidence were seen by mock jurors as more convincing, truthful, competent, intelligent and trustworthy compared to witnesses who used 'powerless speech' and projected low confidence. Furthermore, mock jurors rated expert witnesses who used a formal speech style (which included usage of lay terminology and people's names) as more convincing, competent, qualified and intelligent compared to a hypercorrect style which referred to people in impersonal ways (i.e. 'the client'), used technical terminology, and preferred a pedantic word choice. Bank and Poythress Jr (1982) concluded that in the field of mental health, the impact of the expert witness may depend to a large degree on elements of persuasion, for example the ability to weave expert observations into a scenario which

[1] See the Standard for Friction Ridge Examination Conclusions https://www.nist.gov/topics/organization-scientific-area-committees-forensic-science/friction-ridge-subcommittee)

incorporates important evidence, a free-flowing, narrative style, and components of 'powerful' versus 'powerless' speech. Champagne *et al.* (1990) found that jurors were favourably impressed by experts who had strong communications skills, appearance of knowledge, and impressive educational credentials. Ivković and Hans (2003) found that jurors preferred clear, well-paced, and concise testimony, given enthusiastically and supported by technical aids. Regarding the content of the testimony, jurors were influenced by its completeness, consistency and complexity, the understanding of which required a clear explanation. The Witness Credibility Scale (WCS), developed and validated by Brodsky *et al.* (2010), linked witness credibility to four primary factors: 'knowledge', 'likeability', 'trustworthiness', and 'confidence'. In their study of expert testimony during homicide cases, McCarthy-Wilcox and NicDaeid (2018) found that juror perceptions of the experts' credibility were based upon the latter's academic qualifications, confidence shown when answering questions, demeanour and status as government employees. Jurors described a deeper understanding as a result of narrative testimony and this was reported to be a key factor in the jurors' acceptance that the witness was credible. Overall, it appears that clear, concise, credible and confidant testimony by expert witnesses has the greatest effect on jurors, but questions remain as to the impact of different types of statement language and that is the issue addressed here.

Until recently, scientific forensic evidence was generally perceived as authoritative, clear cut and immune to bias. However, the picture regarding comprehension of forensic evidence is more complex. Motivated by a desire to increase the accuracy and logical cohesiveness of scientific testimony, and to ensure its presentation in a form that can be understood by jurors, researchers have been drawn to the way in which the evidence is delivered in a court of law. A considerable body of research has examined the question of how potential jurors understand conclusions presented in many different quantitative and qualitative formats. Eldridge (2019) reviewed a large portion of the available literature on juror comprehension of forensic science testimony and concluded that jurors 'often undervalue evidence, particularly if it is in a discipline that they may have previously considered to be less discriminating. They do not understand numerical testimony well, although they may prefer to hear it, and they vary widely in their interpretation of verbal expressions . . . '.

Concerns about the impact of conclusion language on the interpretation of evidence have played a key role in recent debates among forensic scientists about the best ways to present their findings in reports and testimony in criminal trials. Categorical statements such as 'the defendant's right index finger has been identified as the source of the fingerprint found on this knife' might be easy to understand but may also fail to acknowledge the uncertainty associated with the conclusion or the possibility of error, thus over-stating the conclusion (National Academy of Sciences, 2009; PCAST, 2016). In response to these criticisms, forensic scientists have, in recent years, been considering various alternative methods of presentation that allow more nuanced statements regarding strength of evidence, including numerical statements such as likelihood ratios and match probabilities, as well as qualitative statements about the strength of evidence (Thompson, 2018; Thompson et al., 2018b). Such statements are easier to justify scientifically, but may be more difficult for laypeople to understand, raising the possibility that injustice might arise due to misunderstanding or misinterpretation.

Social scientists have begun to explore this issue, although the research is in its infancy (Martire *et al.* 2013, 2014; Thompson and Newman, 2015; Mitchell and Garrett, 2019). The studies published thus far share a common limitation — they treat jurors as a single, homogeneous population, not allowing for the possibility that different subgroups in the diverse population of potential jurors may respond to scientific evidence in different ways. At best the studies have examined the effects

of a few demographic and background variables, finding little of note. This article uses a statistical method that is novel in this context in an effort to generate fresh insights about heterogeneity in the population when it comes to juror interpretation of forensic conclusions.

Section 2 reviews the setting in which juror interpretations of forensic evidence testimony is investigated here. Section 3 presents approaches for modelling subpopulations. In Section 3.1 an exploratory analysis of subpopulations defined by specific characteristics is conducted. An alternative approach using mixture models, new to this area of research, which does not require pre-defined subpopulations is presented in Section 3.2. Section 4 discusses the findings and possible implications for practice. A concluding summary is provided in Section 5.

2. Studies of juror perception of forensic evidence

To explore juror perception, we utilize the data from three studies used in Thompson *et al*. (2018a) in which jury-eligible adults recruited from Amazon's Mechanical Turk (mTurk) evaluated the relative strength of statements used to report a forensic scientist's conclusion following a fingerprint comparison (Study 1 and 2) or DNA comparison (Study 3).

2.1 *Types of forensic evidence statements*

One area in which juror comprehension of forensic evidence has been studied deals with statements which compare two pieces of evidence, one from a known source and the other of unknown or questioned origin, in order to determine whether the two share a common source. The questioned and the known items vary from one area of forensic evidence to another. For example, in fingerprint comparison, it is common that the crime scene print is the element with questioned source and the suspect's print is of known source, while in the case of glass comparison, the crime scene glass fragments often have a known source while the glass fragments from the suspect are of questioned origin. The logic of forensic examination involves the identification of similarities and differences between the evidence from a crime scene and the evidence associated with a suspect. In order to reach a conclusion, the examiner has to consider the likelihood of finding the observed similarities and differences under two hypotheses regarding the source of the items; (1) that the items have the same source; and (2) that the items have a different source. In doing this the examiner takes into account the quantity and quality of characteristics that agree as well as how unusual the matching characteristics are. Forensic scientists report the strength of their source conclusions to the court in various ways. Here we summarize one classification of statements based on the Thompson *et al*. (2018a) study that serves as a starting point for our investigation. Thompson *et al*. (2018a) consider statements in several categories:

(1) Statements regarding the relative probability of the observed results if the items have the same source or different source, which can be:
   a. Likelihood Ratios (LRs)
   b. Verbal Strength of Support Statements (SOS)
(2) Statements regarding the probability of the observed results if the items have a different source, which can be:
   a. Random match probabilities (RMP)
   b. Verbal statements about the Likelihood of Observed Similarity (LOS)

(3) Qualitative Source Probability Statements (QSP), sometimes known as statements of posterior probability.
(4) Categorical Conclusions (CC).

LRs represent the relative probability of the observed similarities and discrepancies in the evidence under two alternative hypotheses about the source of the items (same-source or different-source). One example of a statement of this type is 'The genetic features observed in the evidentiary sample are 100,000 times more likely if the sample contains DNA from defendant than if the sample contains DNA from a random unknown Caucasian.' SOS statements are non-numerical statements about the degree to which the results of a forensic comparison support the proposition that the items have the same source (or a different source). An example of such statement is 'there is moderately strong support for the theory that the suspect is the source'. RMPs are sometimes used when a comparison reveals matching features in two items. The examiner estimates and reports the frequency of the matching features in a reference population. For example, 'the probability that a random Caucasian-American would match this DNA profile is 0.0000001 or 1 in 10 million'. QSPs are used by forensic examiners to express opinions on the probability that two items have a common source in a qualitative manner. For example, the forensic scientist might say it is 'moderately probable' or 'highly probable' or 'practically certain' that two items have a common source. CC statements are used by forensic scientists in some disciplines to simply state the conclusion about whether two items have a common source. An example for such statement is 'the crime scene fingerprint matches the suspect's' or 'I identified the suspect as the source of the print'. For a more detailed explanation of the different types of statements see Thompson *et al.* (2018a).

2.2 *Study design and analysis*

The analysis is based on 3 studies. In each, jury-eligible adults evaluated pairs of statements, indicating which of the two was perceived as being stronger evidence that the items originate from the same source. Figure 1, repeated from Thompson *et al.* (2018a) provides an example of a pair of statements given to participants to compare as part of Study 2.

A paired comparison study design was used for assessing the perceived strengths of the statements because pilot studies indicated that people can provide more meaningful responses when asked to evaluate the relative strength of two statements than when asked to evaluate a large number simultaneously. The data from the studies were analysed using a paired comparison model. The model assigns a strength parameter to each statement, $\lambda_i, i = 1, \ldots, N$, where N is the number of

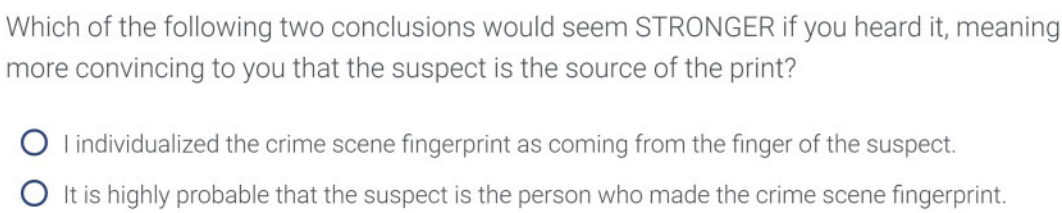

FIG. 1. Example of how pairs of statements were presented to participants

(Thompson, W. C., Grady, R. H., Lai, E., and Stern, H. S. (2018a). Perceived strength of forensic scientists' reporting statements about source conclusions. *Law, Probability and Risk*, 17(2):133–155.).

statements. The probability, π_{ij}, that statement i is found stronger than statement j in terms of these parameters, is assumed to be $\pi_{ij} = F(\lambda_i - \lambda_j)$, where F is a cumulative distribution function (CDF). When F is the Normal CDF, the paired comparison model is known as the unstructured Thurstone–Mosteller model (Thurstone, 1927) and when F is Logistic, the model is known as the unstructured Bradley–Terry model (Bradley and Terry, 1952). Let $Y_{kij} = 1$ indicate that statement i is preferred to j by subject k and 0 otherwise. In the case of the Bradley–Terry model the likelihood of the observed data is

$$L(\lambda \mid y) = \prod_{k \in K} \prod_{ij \in I_k} \left(\frac{e^{\lambda_i - \lambda_j}}{1 + e^{\lambda_i - \lambda_j}} \right)^{y_{kij}} \cdot \left(\frac{1}{1 + e^{\lambda_i - \lambda_j}} \right)^{1 - y_{kij}}$$

where K is the number of subjects and I_k are the statements subject k compared. The λ parameters are estimated using maximum likelihood and the resulting estimates provide an indication of the relative strength of the statements in the studied population.

Note that the probabilities remain the same under an additive constant change of the parameters (the parameters are not identifiable) and thus one of the parameters should be set to a constant. The statement RMP3 ('one person in 100,000') was used as the reference category since it was present in all studies. The estimated strength parameter for RMP3 is reported as zero in each case. The estimate presented for each other statement indicates the strength of that item relative to RMP3, with positive numbers indicating the statement is perceived as stronger than RMP3, and negative numbers indicating it is perceived to be weaker.

In each study nine statements were paired randomly, creating a pool of 36 possible comparisons. The number of pairs evaluated by each participant was limited to a selection of 16. Some statements were used in all three studies; some in just one or two of the studies. Table 1, reproduced from Thompson *et al.* (2018a), presents a complete list of statements used throughout the studies. Study 1 includes various numerical statements that allow us to focus on differences in the population with respect to quasi-quantitative and quantitative reports while studies 2 and 3 make it possible to compare qualitative categorical statements to newer quantitative approaches. In addition to the variation in the statements that were included, the studies used different forensic evidence types. Studies 1 and 2 were based on fingerprint comparisons and Study 3 on DNA comparison. Study 1, 2 and 3 included 120, 121 and 138 participants, respectively.[2]

Five characteristics of each participant were measured in all three studies.

(1) Gender: male, female
(2) Age
(3) SNS: subjective numeracy on a scale from 1 to 6, averaged across 8 questions. This is a self-report measure of perceived ability to perform various mathematical tasks and preferences for the use of numerical versus prose information (Fagerlin *et al.*, 2007).
(4) Education level, going from 1 (some high school) to 9 (received a doctoral degree)
(5) Forensic knowledge: self-rated knowledge of forensic science. This is a single 7-point question that asked 'How knowledgeable are you about forensic science?' going from 1 (Not Knowledgeable) to 7 (Extremely Knowledgeable).

[2] These sample sizes differ slightly from those reported in Thompson et al. (2018a) because unusual covariate values were identified and eliminated.

TABLE 1 *The statements used in the studies, Thompson, W. C., Grady, R. H., Lai, E., and Stern, H. S. (2018a). Perceived strength of forensic scientists' reporting statements about source conclusions.* Law, Probability and Risk, *17(2):133–155.*

Study	Statement	
		Match frequency (RMP)
2,3	RMP4:	'one person in 10 million'
1, 2, 3	RMP3:	'one person in 100, 000'
1, 2	RMP2:	'one person in 1, 000'
1	RMP1:	'one person in 10'
		Likelihood ratios (LR)
3	LR4:	'10, 000, 000 times more likely' if suspect rather than random person is source
3	LR3:	'100, 000 times more likely' if suspect rather than random person is source
		Categorical conclusions (CC)
3	CC5:	'suspect was the source' person is source
2	CC4:	'individualized. . . as coming from the finger of the suspect'
2	CC3:	'identified. . . to the finger of the suspect'
2	CC2:	'matches the fingerprint of the suspect'
3	CC1:	'suspect could have been the source'
		Likelihood of observed similarity (LOS)
2	LOS1:	'likelihood of observing this amount of corresponding ridge detail when two fingerprints are made by different people is considered extremely low'
		Source probability statements (SP)
1	SP3:	'a practical certainty that suspect was the source'
1, 2, 3	SP2:	'highly probable'
1	SP1:	'moderately probable'
		Strength of support (SOS)
1, 2, 3	SOS4:	'extremely strong support'
3	SOS3:	'very strong support'
1	SOS2:	'moderate support'
1	SOS1:	'weak support'

3. Models for subpopulations

In order to assess the validity of the assumption made by Thompson *et al*. (2018a) that all people are governed by the same preferences, this article explores two approaches to identifying subpopulations. The first approach is an exploratory analysis, which considers subpopulations defined *a priori* by specific characteristics (e.g. age). A second approach that does not require us to *a priori* identify which characteristics might be relevant uses a mixture model to identify subpopulations. We first use a mixture model that assumes that subpopulations are not described by the covariates and then try to explore whether any of the covariates help to explain the observed subpopulations.

3.1 *An exploratory analysis of defined subpopulations*

As an initial step, the data collected by Thompson *et al*. (2018a) was used to investigate the ranking patterns of different subpopulations identified by characteristics of the participant (age, SNS, gender, level of education and forensic knowledge). A series of analyses are carried out, each

TABLE 2 *A comparison of the different unstructured Bradley–Terry (BT) log likelihoods of the three studies. Models found to be a significant improvement over the BT model, assuming a single set of preferences for the entire population are marked with an asterisk*

	Model	Log Likelihoods		
		Study 1	Study 2	Study 3
1	BT for the complete data	–912.43	–951.5	–1137.14
2	BT for high and low levels of forensic knowledge	–909.75	–943.77	–1128.78
3	BT for female and male	–909.15	–946.4	–1130.64
4	BT for old and young age	–908.20	–940.17	* –1117.81
5	BT for high and low levels of education	*–895.12	–945.66	–1130.75
6	BT for high and low SNS levels	*–875.96	*–935.5	*–1121.56
7	A mixture of two subpopulation models	*–735.90	*–813.18	*–1044.721

considering two subpopulations defined by a single characteristic. The original study sample was split into two groups according to each characteristic as defined here:

(1) Gender (Male, Female)
(2) Age (greater than or equal to 32, lower than 32)
(3) SNS total (greater than or equal to 5, lower than 5)
(4) Level of Education (greater than or equal to 6, lower than 6)
(5) Forensic knowledge (greater than or equal to 4, lower than 4)

The cut-points for defining the groups based on quantitative characteristics are the median values. For each subpopulation (e.g. people of age greater than or equal to 32) an unstructured paired comparison Bradley–Terry model (Bradley and Terry, 1952) was fitted and the log likelihood of the model was calculated. The log likelihood of the complementary subpopulation (e.g. age lower than 32) was then calculated. The sum of these two log likelihoods is the log likelihood of a two subpopulation Bradley–Terry model.

Table 2 presents the value of the maximized log likelihood for a set of models. The first row reports the values for the data analysed by Thompson *et al.* (2018a). The Bradley–Terry model was used here instead of the Thurstone–Mosteller model applied there. Rows 2–6 are the values for the models that define subpopulation via median split on the specified variable. The final row presents results for a mixture model approach described in Section 3.2. The models in row 2–6 can be compared to the model in row 1 (separately for each column) via a traditional significance test. It should be noted that the Bonferroni correction (Wasserman, 2004, pp. 165–166) was used to account for multiple significant tests, meaning that only models with a p-value less than $\frac{\alpha}{m}$ (where $m = 18$ is the cumulative number of suggested models in all 3 studies and $\alpha = 0.05$ is the significance level), were found to be an improvement over the Bradley–Terry model fitted to the entire population.

The primary finding after adjusting for multiple comparisons is that there appear to be differences based on subjective numeracy levels (SNS). The Bradley–Terry model fitted separately to the two SNS groups was found to be a significant improvement over the Bradley–Terry model fitted to the entire population in all three studies. In addition, it produced the highest or second highest likelihood for each study (among rows 2–6).

TABLE 3 *SNS Analysis. In each study, the estimated strength parameters of the low and high SNS groups are presented*

Study 1				Study 2				Study 3			
low		high		low		high		low		high	
SP3	0.41	SP3	0.79	CC2	0.99	RMP4	1.10	LR4	0.96	LR4	1.31
SOS4	0.05	SOS4	0.47	RMP4	0.93	CC2	1.02	RMP4	0.93	RMP4	1.07
RMP3	0.00	RMP3	0.00	CC3	0.30	CC3	0.25	SOS4	0.30	CC5	0.42
SP2	–0.27	SP2	–0.43	CC4	0.08	RMP3	0.00	CC5	0.27	LR3	0.38
RMP2	–0.76	RMP2	–1.57	RMP3	0.00	CC4	–0.27	LR3	0.21	RMP3	0.00
SP1	–1.09	SOS2	–2.29	SOS4	–0.40	SOS4	–1.07	RMP3	0.00	SOS4	–0.23
SOS2	–1.20	SP1	–2.46	LOS1	–0.70	LOS1	–1.60	SOS3	–0.20	SOS3	–1.07
RMP1	–1.81	RMP1	–3.77	SP2	–0.81	RMP2	–1.60	SP2	–1.02	SP2	–1.35
SOS1	–2.14	SOS1	–4.65	RMP2	–1.41	SP2	–2.29	CC1	–2.25	CC1	–3.46

Table 3 presents the estimated strength parameters of the low and high SNS groups for each of the three studies. The standard errors are approximately 0.2. It is generally true that the ordering of the statements is the same in the high and the low SNS groups. Occasionally there is a reversal in the order. For example, in Study 1 the position of SOS2 and SP1 differs.

When splitting the data based on SNS, statements within categories maintained their order. For example, in Study 1 among SOS statements, 'extremely strong support' (SOS4) was viewed as stronger than 'moderate support' (SOS2), which was in turn seen as stronger than 'weak support' (SOS1). In addition, in almost no category of statements were all statements perceived as stronger or weaker than the statements in any other category except for the high SNS group of Study 3 in which SOS statements were perceived as weaker than RMP and LR statements. Furthermore, the relative ranking of statements across studies was generally consistent. For example, RMP3 ('1 in 100,000') was consistently ranked higher than SP2 ('highly probable'), and SOS4 ('extremely strong support') was also ranked consistently as stronger than SP2 although the difference between these statements was not always significant.

An interesting finding is that the scale used by subjects in the high SNS group is more spread-out than that used by the low SNS group, suggesting that the high SNS subjects are more able to distinguish between the statements. This is especially true for the high SNS group of Study 1. In this group, the estimated strength parameters range from –4.65 to 0.79 which means that according to the Bradley–Terry model, the estimated probability that the strongest statement is preferred to the weakest statement is 0.996. In the low SNS group, the estimated strength parameters range from –2.14 to 0.41 which means that the estimated probability that the strongest statement is preferred to the weakest in 0.928.

It should be noted that the values of the estimated strength parameters in the different studies should not be compared since the values of these estimators are calculated from comparison with the specific statements included in the study. The relative order of the estimators may be compared across studies. The values of the estimated strength parameters for different subpopulations within each study may be compared as well.

To conclude, the exploratory approach suggests that low and high SNS subjects react differently but that the differences in the perceived strength are relatively small. The ordering of the statements

is similar in both groups but high SNS subjects tend to distinguish more between stronger and weaker statements. A limitation of this approach is that it focuses only on a single characteristic of the individual at a time. A more general approach which allows the data to more flexibly define subpopulations that most differ in terms of their ranking is considered in the next section.

3.2 *A mixture model approach for identifying subpopulations*

The mixture (or mixture of experts) model provides an alternative approach that does not require pre-defined subpopulations. Formally, the mixture model associates an unobserved or latent variable with each individual that indicates the subpopulation to which the subject belongs. The subpopulations are characterized by different values of the strength parameters (as in the previous section).

A second component of the mixture model relates the latent variables to covariates. The covariates determine the probability of belonging to a certain subpopulation. For example, when there are only two subpopulations, the latent variable has two options and therefore one might assume a logistic model (Morduch and Stern, 1997).

The description here allows for U subpopulations. Let x_k be the covariate vector of subject k, and u_k indicate the group or subpopulation membership of subject k, $u_k = 1, \ldots, U$ where u_k is not observable. A logistic model for subpopulation membership is specified by

$$Pr(U_k = u_k) = \frac{e^{\beta_{u_k} x_k}}{\sum_{u=1}^{U} e^{\beta_u x_k}},$$

where $\beta_U = 0$. Combining this model with a separate Bradley–Terry model (as in Section 2.2) for each subpopulation yields the following likelihood, $L(\lambda, \beta | x, y)$ as

$$\prod_{k \in K} \sum_{u_k=1}^{U} \prod_{ij \in I_k} \left(\frac{e^{\lambda_i^{(u_k)} - \lambda_j^{(u_k)}}}{1 + e^{\lambda_i^{(u_k)} - \lambda_j^{(u_k)}}} \right)^{y_{kij}} \cdot \left(\frac{1}{1 + e^{\lambda_i^{(u_k)} - \lambda_j^{(u_k)}}} \right)^{1 - y_{kij}} \cdot \left(\frac{e^{\beta_{u_k} x_k}}{\sum_{u=1}^{U} e^{\beta_u x_k}} \right). \tag{3.1}$$

The maximum likelihood estimators for the parameters can be calculated by directly maximizing the likelihood or by using the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977). The EM algorithm is an iterative method that can be used to find maximum likelihood estimates of parameters where a model depends on unobserved latent variables as is the case here. The EM iterates between computing expected values of the sufficient statistics for the latent variables (in this case the probabilities for each value of U_k) and maximizing the complete data likelihood (including the latent variables) given the expected sufficient statistic. More details are provided in Appendix A. The EM used here was given a relatively good staring point using a procedure developed by the authors to improve the EM convergence, this is also described in Appendix B.

The bottom row of Table 2 presents the maximized log likelihood of the two component ($U = 2$) mixture model with no covariates ($x_k =$ "intercept only") in the three studies. The two-component mixture model achieves the highest log likelihood in all of the three studies (as seen in this table).

Tables 4–6 and Fig. 2 present the results of applying the two-component mixture model to the data for all three studies. The strength parameter estimates and the standard errors are presented in Tables 4–6 and a graphic visualization of the estimated strength parameters is presented in Fig. 2. A

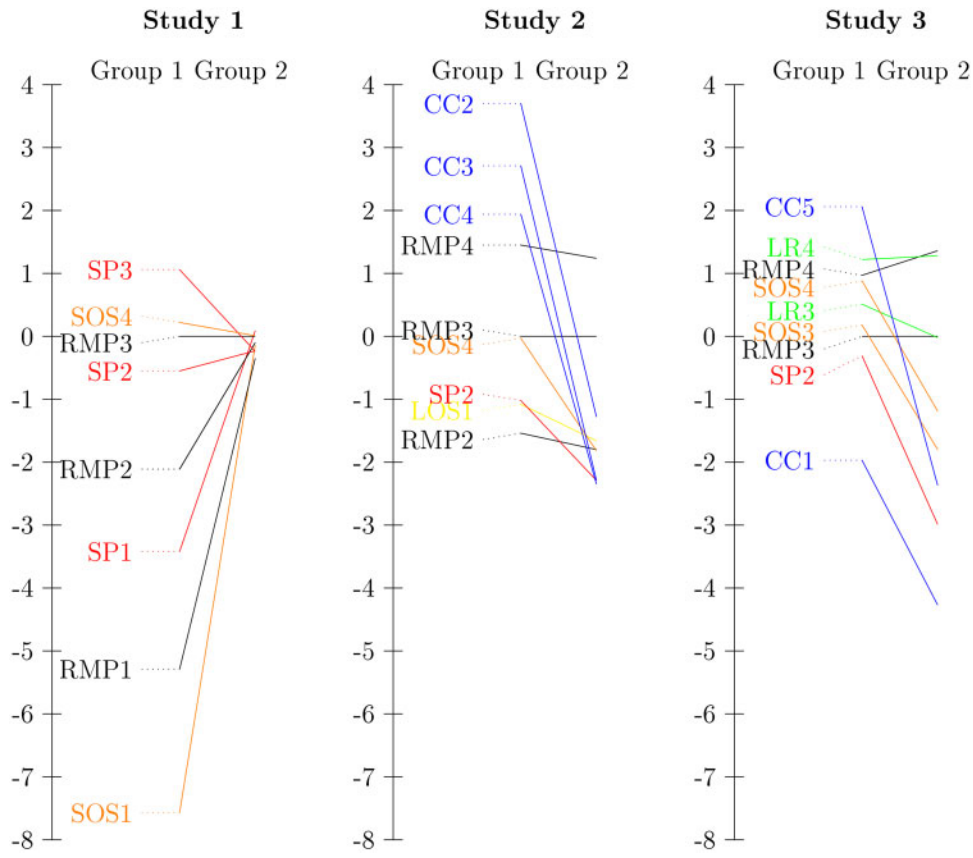

FIG. 2. Graphical display of the estimated statement strength parameters from the mixture model analysis of the data from studies 1–3. The colours of the statement abbreviations indicate the classes of statements as presented in Table 1. For each study, the abbreviations on the left appear at the location on the vertical axis indicating the estimated strength parameter for the statement in Group 1. The line connects to the value of the estimated strength parameter for that statement in Group 2. For example, In Study 1, the location of SP3 on the vertical axis is 1.06 which is the estimated strength parameter for this statement in Group 1. The line connects to –0.24 which is the value of the estimated strength parameter in Group 2 (See Table 4).

high estimated strength parameter corresponds to a statement which was perceived as a particularly strong statement. As in Thompson *et al*. (2018a) the strength parameter for RMP3 is set equal to zero to identify the model parameters.

Using the data of Study 1, the mixture model identifies two subpopulations. One subpopulation (Group 1), estimated to comprise 75% of the population, ranks the statements in the order identified by Thompson *et al*. (2018a). The second subpopulation appears not to be able to distinguish among the statements. This means that the way in which the expert words the conclusion has no influence on them whatsoever. It is also possible that 25% of the population just choose randomly and that is the reason it appears that they are not able to distinguish among the statements. Subjects of Group 1 ranked the statements in their natural order within a category of statements (e.g. SOS). They perceived SP3 as the strongest statement and SOS1 as the weakest. These subjects did not choose all

TABLE 4 *The statement ranking results of Study 1 using a mixture model*

Study 1					
Group 1 (75%)	Estimate	SE	Group 2 (25%)	Estimate	SE
SP3	1.06	0.27	SP1	0.09	0.27
SOS4	0.22	0.25	SOS4	0.01	0.28
RMP3	0	–	RMP3	0	–
SP2	–0.55	0.25	RMP2	–0.1	0.23
RMP2	–2.11	0.25	SOS1	–0.14	0.26
SOS2	–3.28	0.32	SP2	–0.23	0.26
SP1	–3.42	0.32	SP3	–0.24	0.28
RMP1	–5.29	0.4	SOS2	–0.29	0.26
SOS1	–7.57	0.62	RMP1	–0.35	0.23

TABLE 5 *The statement ranking results of Study 2 using a mixture model*

Study 2					
Group 1 (59%)	Estimate	SE	Group 2 (41%)	Estimate	SE
CC2	3.7	0.41	RMP4	1.24	0.28
CC3	2.71	0.33	RMP3	0	–
CC4	1.94	0.31	CC2	–1.28	0.29
RMP4	1.45	0.22	LOS1	–1.66	0.28
RMP3	0	–	RMP2	–1.8	0.25
SOS4	–0.03	0.24	SOS4	–1.82	0.28
SP2	–1.02	0.26	CC3	–2.29	0.32
LOS1	–1.08	0.27	SP2	–2.29	0.29
RMP2	–1.54	0.21	CC4	–2.35	0.3

statements of one category as being stronger or weaker than the statements in any other category and they perceived SPs as being stronger than their corresponding RMPs.

Table 7 presents the explanatory variables' coefficient estimates and the standard errors for the logistic part of the mixture model. The SNS, age and education estimates are found to be significant which means that these variables help distinguish between the groups. Group 1 is characterized by people with higher SNS, age and education. The coefficient in Table 7 describe the impact of a difference in that variable on the likelihood of being in subpopulation 1. The meaning of the SNS coefficient for example is that for every increase of 1 point in SNS, the odds of belonging to Group 1 compared to Group 2 increases by a factor of 2.64 ($= e^{0.97}$).

The results of fitting the mixture model to Studies 2 and 3 are quite different from the Study 1 results. The key finding in these studies is that there are two subpopulations which differ in the relative strength of different statement categories. In Study 2, the two groups seem to agree on the strength of the RMP statements but perceive the CC statements quite differently. Group 1 (59% of

TABLE 6 *The statement ranking results of Study 3 using a mixture model*

Study 3					
Group 1 (56%)	Estimate	SE	Group 2 (44%)	Estimate	SE
CC5	2.06	0.29	RMP4	1.36	0.32
LR4	1.22	0.21	LR4	1.28	0.36
RMP4	0.97	0.19	RMP3	0	0
SOS4	0.88	0.21	LR3	–0.02	0.3
LR3	0.51	0.2	SOS4	–1.19	0.3
SOS3	0.18	0.21	SOS3	–1.8	0.32
RMP3	0	0	CC5	–2.37	0.48
SP2	–0.31	0.23	SP2	–2.99	0.35
CC1	–1.97	0.29	CC1	–4.27	0.42

TABLE 7 *The explanatory variables' coefficient estimates using data from Study 1*

	Estimate	Std. Error
(Intercept)	–6.22	1.73
SNS total	0.97	0.29
Age	0.06	–0.03
Gender	0.32	0.53
Education	0.33	0.16
Forensic knowledge	–0.28	0.18

the population) tends to find CC statements stronger than RMP statements as opposed to Group 2 (41% of the population) which favours the numerical information presented by RMP statements.

Study 3 resembles Study 2 in the sense that the two groups tend to agree on the RMP statements and find the CC statements relatively different. Members of Group 1 (56% of the population) tend to choose CC5, 'suspect was the source', as stronger than RMP statements while subjects of Group 1 (44% of the population) are not impressed by this statement and rank it quite low. In both studies the two groups rank SOS4 and SP2 statements below RMP4 but in each study the groups differ in the relative distance between these statements.

Table 8 presents a comparison of π_{ij}, the probability that statement i is preferred to statement j, for statements included in Studies 2 and 3, i.e. RMP3, RMP4, SOS4 and SP2. In both studies, Group 2 is the one that favours the numerical statements. The consistency of the two studies is also shown in the columns that represent these groups (marked in grey) as the probabilities are close.

Interestingly, the explanatory variables are not found to be helpful in distinguishing between the groups in Studies 2 and 3 (the estimates are not significant) and thus an EM with no explanatory variables (x_k = "intercept only") is used there.

Here the division of the population into two groups which best describe the observed data was presented. A crucial question is whether the two-subpopulation model does in fact provide a substantially better fit than a single model fitted to the entire population. One approach to determine

TABLE 8 *Comparison of π_{ij} for statements included in Studies 2 and 3. Columns marked in grey represent groups that favour the numerical statements*

	Study 2		Study 3	
Statements	Group 1	Group 2	Group 1	Group 2
RMP3 vs RMP4	0.19	0.22	0.27	0.2
RMP3 vs SOS4	0.51	0.86	0.29	0.77
RMP3 vs SP2	0.73	0.91	0.58	0.95
RMP4 vs SOS4	0.81	0.96	0.52	0.93
RMP4 vs SP2	0.92	0.97	0.78	0.99
SOS4 vs SP2	0.73	0.62	0.77	0.86

the number of subpopulations is to conduct a Monte Carlo LR test (Everitt, 1981). In this setting, the Monte Carlo test supports the presence of two subpopulations rather than one. It also suggests that the sample sizes are not large enough to determine whether there are more than two subpopulations.

4. Discussion

The 'data-driven' study reported here allows detection of coherent subgroups from the data itself in contrast to the traditional research strategy which is 'hypothesis driven'. The advantage of this approach is the ability to detect subgroup structure in the data that might be missed otherwise. The study examines whether the data are more consistent with a model assuming subpopulations than a model that assumes homogeneity (no subpopulations). Using data from the three studies of Thompson *et al.* (2018a) it was found that there is evidence that different subpopulations tend to perceive the statements presented by the examiner differently. This in itself is an important finding since there is a need to acknowledge that even when solid scientific evidence is presented, it may be interpreted at least by a certain percent of the population in a way that is different from the way the examiner intended.

The exploratory analysis suggests that there appear to be differences based on the subjective numeracy levels (SNS) and that high SNS subjects tend to better distinguish between stronger and weaker statements which supports the findings of Thompson *et al.* (2013) and Scurich (2015). The mixture model approach suggests that there is a subpopulation that finds quantitative summaries of the strength of the evidence more compelling. One might expect that SNS would be associated with membership in this subpopulation, however we do not find this relationship in the data. This latter finding is more consistent with the results of Thompson and Newman (2015); Martire *et al.* (2013; 2014).

As noted in Section 2.2, Study 1 by its design allowed us to focus on differences in the population with respect to quasi-quantitative and quantitative scales while studies 2 and 3 allowed us to compare qualitative categorical statements to newer quantitative approaches.

The mixture model which does not require pre-defined subpopulations identified two subpopulations in Study 1 that differ completely in the way that they perceive the statements: the subjects of Group 2 were unable to distinguish between statements while those of Group 1 ranked the statements in the same order found in Thompson *et al.* (2018a). One possible explanation is that

participants in Group 1 paid more attention to the details of what the expert was saying than those in Group 2. Alternatively, Group 1 participants may have been more capable of processing and appreciating the significance of the expert's statements about the strength of the evidence than those in Group 2. Group 1 consisted of older, better-educated participants with higher subjective numeracy than Group 2, which makes it plausible that they were more conscientious or more discerning in their evaluations of the evidence. These possible explorations could be tested and clarified through further research.

A more interesting picture is presented in Studies 2 and 3, both of which show that the groups seem to agree on the strength of the RMP statements though one subpopulation prefers categorical statements. This finding may reflect a general difference among individuals in their willingness or ability to make use of numbers, and their preferences for numerical versus more qualitative statements. Previous research has suggested that people who express high confidence in their ability to draw correct conclusions from numerical data respond more strongly and correctly to statistical data about the value of forensic evidence than do people who express less confidence (Kaasa *et al.* 2007). This research may have implications for the process of juror selection, prompting attorneys to consider the preferences of potential jurors in accordance with the expected nature and presentation of evidence in court.

A limitation in using these studies stems from the fact that since not all statements are included in all studies, it is difficult to reach a more general conclusion. In addition, as noted in Section 3, a model which allows for a larger number of subpopulations might provide a better description of the population but since the data sets are relatively small and the number of parameters increases in such models, they can be hard to fit. For these reasons it is recommended that additional studies which include a greater number of participants be carried out. Moreover, alternative approaches for dealing with heterogeneity in the data are currently under investigation. An additional issue is the limited covariates information. To better characterize the subpopulations found in this study, such as the group that prefers categorical versus RMP statements, additional personal characteristics should be taken into account in future studies.

The discovery that the population of potential jurors consists of subgroups which interpret expert testimony in different ways, may prompt courts to review existing policies. Though the findings of this study do not necessarily support the demand for highly numerate 'Blue ribbon' juries (Halle, 2014; Hans and Helm, 2019), the group differences in preferred types of statements suggest that forensic scientists do need to present their findings in multiple ways. Perhaps there is no single approach that works best for all potential jurors, but by offering several complementary alternative statements, forensic experts may be able to assure that their testimony is more broadly understood. Similarly, training of attorneys and judges as well as the pre-trial instruction of jurors could well facilitate a clearer and more uniform understanding of the terms used to describe the strength of forensic evidence presented in court (Evans *et al.*, 2019). It is our hope that research like that reported here will cast light on how this might be done.

5. Summary

Data from three studies investigating the perceptions of the lay public regarding presentations of forensic evidence point to the existence of subpopulations which perceive statements reported by expert witnesses differently. In Study 1, which focused on differences with respect to quasi-quantitative and quantitative statements, the participants' numeracy, age, education and possibly the

degree of attention paid to the proceedings (or capacity to understand them) affected their ability to distinguish between statements. In studies 2 and 3, which included both numerical and categorical statements, differences were observed in the relative strength assigned to the two categories of statements. These findings could influence the selection of juries in cases where forensic evidence is considered, and might support the argument for training important actors in the judicial process (prosecutors, juries and judges) in the appropriate interpretation of forensic evidence. The justice system might well be advised to consider multiple ways of presenting evidence in court in order to provide for its better understanding by diverse subpopulations.

The mixture models illustrated here have a broad range of potential applications in research on jury decision-making. They may prove particularly helpful in identifying differences among subpopulations that researchers fail to anticipate because subpopulations cohere in unexpected ways. They are an important new tool for mock jury research.

Appendix

Appendix A. The EM algorithm

The EM algorithm is an iterative method for finding maximum likelihood estimates of parameters in situations with missing data or with unobserved latent variables (Dempster *et al*. 1977). EM is used to calculate maximum likelihood estimates for the mixture model described in Section 3.2. The algorithm is described in this appendix with slightly different notation. Z_k denotes the subpopulation indicator here. To develop the algorithm the complete data likelihood is first defined for the observed paired comparison data Y and the latent variables (subpopulation indicators) Z_k,

$$L(\theta|Y,Z) = \prod_{k\in K}\prod_{z_k=1}^{U}\prod_{ij\in I_k}\left[\left(\frac{e^{\lambda_i^{(Z_k)}-\lambda_j^{(Z_k)}}}{1+e^{\lambda_i^{(Z_k)}-\lambda_j^{(Z_k)}}}\right)^{y_{kij}}\cdot\left(\frac{1}{1+e^{\lambda_i^{(Z_k)}-\lambda_j^{(Z_k)}}}\right)^{1-y_{kij}}\cdot\left(\frac{e^{\beta_{Z_k}x_k}}{\sum_{u=1}^{U}e^{\beta_u x_k}}\right)\right]^{I_{\{Z_k=z_k\}}}.$$

where $\theta = (\lambda, \beta)$ are the parameters of the model and dependence on the covariates X is suppressed in the notation. Each iteration of the algorithm is comprised of two steps. The E-step or Expectation step computes the expectation of the complete data log likelihood conditional on current estimates of the parameter θ. The M-step or Maximization step then updates the parameter estimates. In our case, assuming we have completed the tth iteration, the next step involves defining $Q(\theta|\theta^{(t)})$, the expected complete data log likelihood given $\theta^{(t)}$ where the expectation is over the latent variables $Z_k, k \in K$. Evaluating the function Q requires computation of $E(I_{\{Z_k=z_k\}}|\theta^{(t)}, Y) = \Pr(Z_k = z_k|\theta^{(t)}, Y)$ for $z_k = 1, \ldots, U$. These conditional probabilities are computed as

$$W_{z_k}^{(t)} = Pr(Z_k = z_k|\theta^{(t)}, Y) \propto \prod_{ij\in I_k}\left[\left(\frac{e^{\lambda_i^{(Z_k)}-\lambda_j^{(Z_k)}}}{1+e^{\lambda_i^{(Z_k)}-\lambda_j^{(Z_k)}}}\right)^{y_{kij}}\cdot\left(\frac{1}{1+e^{\lambda_i^{(Z_k)}-\lambda_j^{(Z_k)}}}\right)^{1-y_{kij}}\cdot\left(\frac{e^{\beta_{Z_k}x_k}}{\sum_{u=1}^{U}e^{\beta_u x_k}}\right)\right].$$

Plugging these expectations into the expression for Q yields

$$Q(\theta|\theta^{(t)}) = \sum_{k\in K}\sum_{z_k=1}^{U}\sum_{ij\in I_k}\log\left[\left(\frac{e^{\lambda_i^{(z_k)}-\lambda_j^{(z_k)}}}{1+e^{\lambda_i^{(z_k)}-\lambda_j^{(z_k)}}}\right)^{y_{kij}}\cdot\left(\frac{1}{1+e^{\lambda_i^{(z_k)}-\lambda_j^{(z_k)}}}\right)^{1-y_{kij}}\right]^{I_{\{Z_k=z_k\}}}\cdot W_{z_k}^{(t)}$$
$$+\sum_{k\in K}\sum_{z_k=1}^{U}\sum_{ij\in I_k}\log\left[\left(\frac{e^{\beta_{Z_k}x_k}}{\sum_{u=1}^{U}e^{\beta_u x_k}}\right)\right]^{I_{\{Z_k=z_k\}}}\cdot W_{z_k}^{(t)}.$$

The M-step maximizes $Q(\theta|\theta^{(t)})$ over the parameters θ to give the next iterate $\theta^{(t+1)}$. Examination of the form of Q indicates that $\lambda^{(t+1)}$ and $\beta^{(t+1)}$ can be obtained independently of one another by maximizing the relevant Bradley–Terry and multinomial logistic terms. The E- and M-steps are iterated until the parameter estimates do not change.

Appendix B. Calculation of starting points for the EM

The EM algorithm used here utilizes the following algorithm which produces relatively good starting points in order to reduce the time to convergence.

(1) The n individuals are randomly divided into C clusters (the number of clusters is chosen in advance). For the work presented here C is 2.
(2) In each cluster an unstructured Bradley–Terry model is fit to the data using the *BradleyTerry2* package (Turner and Firth, 2012) for the R software package. The strength parameters are denoted by $\lambda^{(c)}$.
(3) The likelihood of individual k using cluster c parameters is calculated as follows.

Let $Y_{ij} = \begin{cases} 1 & \text{if } i \text{ is prefered over } j \\ 0 & \text{o.w} \end{cases}$

and $\Pi_{ij} = P(Y_{ij}=1) = \frac{e^{\lambda_i^{(c)}-\lambda_j^{(c)}}}{1+e^{\lambda_i^{(c)}-\lambda_j^{(c)}}}$, then the likelihood of individual k is:

$$\prod_{(ij)\in I_k}(\Pi_{ij})^{y_{ijk}}(1-\Pi_{ij})^{1-y_{ijk}},$$

where I_k represents the questions presented to individual k.

(4) The likelihoods of the various clusters are compared for each individual. The individual is assigned to the cluster with the highest likelihood.
(5) The process (steps 2–4) is repeated until the cluster assignment remains constant. The final values of the strength parameters are used as the initial values for the EM algorithm described in Appendix A.
(6) The initial values of beta for the EM algorithm described in Appendix A are taken to be the coefficients from a logistic regression of the final cluster assignment on the covariates.

Funding

This work was funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement 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

Bank, S. C. and Poythress Jr, N. G. (1982). The elements of persuasion in expert testimony. *The Journal of Psychiatry & Law*, **10**(2):173–204.

Bradley, C. M. (1996). The convergence of the continental and the common law model of criminal procedure. *Criminal Law Forum*, **7**: 471–484.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, **39**(3/4):324–345.

Brodsky, S. L., Griffin, M. P., and Cramer, R. J. (2010). The witness credibility scale: An outcome measure for expert witness research. *Behavioral Sciences & the Law*, **28**(6):892–907.

Champagne, A., Shuman, D., and Whitaker, E. (1990). An empirical examination of the use of expert witnesses in American courts. *Jurimetrics Journal*, **31**(4):375–392.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1):1–22.

Devine, D. J., Clayton, L. D., Dunford, B. B., Seying, R., and Pryce, J. (2001). Jury decision making: 45 years of empirical research on deliberating groups. *Psychology, Public Policy, and Law*, **7**(3):622.

Eldridge, H. (2019). Juror comprehension of forensic expert testimony: A literature review and gap analysis. *Forensic Science International: Synergy*.

Evans, A., Kim, C., Laraia, N., Lutz, Z., Osborne, A., Parchuke, E., Williams, L., Zook, A., and Quarino, L. (2019). Toward a more effective use and understanding of forensic evidence in courts of law. *Widener Law Review*, **25**:1.

Everitt, B. S. (1981). A monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioral Research*, **16**(2):171–180.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., and Smith, D. M. (2007). Measuring numeracy without a math test: Development of the subjective numeracy scale. *Medical Decision Making*, **27**(5):672–680.

Halle, J. M. (2014). Avoiding those wearing propeller hats: The use of blue ribbon juries in complex patent litigation. *University of Baltimore Law Review*, **43**:435.

Hans, V. and Helm, R. (2019). Professional judges, lay judges, and lay jurors. In K., Darryl, J. I. T.Brown, and B., Weisser, editors, *The Oxford Handbook of Criminal Process*, chapter **10**, page 214. Oxford University Press, Oxford.

Ivković, S. K. and Hans, V. P. (2003). Jurors' evaluations of expert testimony: Judging the messenger and the message. *Law & Social Inquiry*, **28**(2):441–482.

Kaasa, S. O., Peterson, T., Morris, E. K., and Thompson, W. C. (2007). Statistical inference and forensic evidence: Evaluating a bullet lead match. *Law and Human Behavior*, **31**(5):433–447.

Lindsay, B. G. and Lesperance, M. L. (1995). A review of semiparametric mixture models. *Journal of Statistical Planning and Inference*, **47**(1–2):29–39.

MacCoun, R. J. (1989). Experimental research on jury decision-making. *Science*, **244**(4908):1046–1050.

Martire, K. A., Kemp, R. I., Sayle, M., and Newell, B. (2014). On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect. *Forensic Science International*, **240**:61–68.

Martire, K. A., Kemp, R. I., Watkins, I., Sayle, M. A., and Newell, B. R. (2013). The expression and interpretation of uncertain forensic science evidence: Verbal equivalence, evidence strength, and the weak evidence effect. *Law and Human Behavior*, **37**(3):197.

McCarthy Wilcox, A. and NicDaeid, N. (2018). Jurors' perceptions of forensic science expert witnesses: experience, qualifications, testimony style and credibility. *Forensic Science International*, **291**:100–108.
Mitchell, G. and Garrett, B. L. (2019). The impact of proficiency testing information and error aversions on the weight given to fingerprint evidence. *Behavioral Sciences & the Law*, **37**(2):195–210.
Morduch, J. J. and Stern, H. S. (1997). Using mixture models to detect sex bias in health outcomes in Bangladesh. *Journal of Econometrics*, **77**(1):259–276.
National Academy of Sciences. (2009). *Strengthening forensic science in the United States: a path forward*. Retrieved from http://www.nap.edu/catalog.php? record_id=12589
O'Barr, W. M. (1982). Linguistic evidence: Language. *Power, and Strategy in the Courtroom*, **3**.
PCAST (President's Council of Advisors on Science and Technology). (2016). *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Report to the President. Retrieved from https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports/
Scurich, N. (2015). The differential effect of numeracy and anecdotes on the perceived fallibility of forensic science. *Psychiatry, Psychology and Law*, **22**(4):616–623.
Thompson, W. C. (2018). How should forensic scientists present source conclusions. *Seton Hall Law Review*, **48**:773.
Thompson, W. C., Grady, R. H., Lai, E., and Stern, H. S. (2018a). Perceived strength of forensic scientists' reporting statements about source conclusions. *Law, Probability and Risk*, **17**(2):133–155.
Thompson, W. C., Kaasa, S. O., and Peterson, T. (2013). Do jurors give appropriate weight to forensic identification evidence? *Journal of Empirical Legal Studies*, **10**(2):359–397.
Thompson, W. C. and Newman, E. J. (2015). Lay understanding of forensic statistics: Evaluation of random match probabilities, likelihood ratios, and verbal equivalents. *Law and Human Behavior*, **39**(4):332.
Thompson, W. C., Vuille, J., Taroni, F., and Bidermann, A. (2018b). After uniqueness: the evolution of forensic science opinions. *Judicature*, **102**:18.
Thurstone, L. L. (1927). The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology*, **21**(4):384.
Turner, H. and Firth, D. (2012). Bradley-Terry models in R: the BradleyTerry2 package. *Journal of Statistical Software*, **48**(9).
Wasserman, L. (2004). *All of Statistics*. Springer, NY.