Science & Justice 63 (2023) 38–53

Contents lists available at ScienceDirect

Science & Justice

journal homepage: www.elsevier.com/locate/scijus

Calibrating the perceived strength of evidence of forensic testimony statements

Thomas Busey *, Morgan Klutzke

Indiana University, United States

ARTICLE INFO

Keywords:
Forensic testimony
Categorical conclusions
Articulation language

ABSTRACT

Pattern comparison disciplines use categorical statements to express conclusions. We measured the strength of evidence for six different scales as perceived by members of the general public and fingerprint examiners. The statements came from different types of scales, and included categorical conclusions, likelihoods, strength of support statements, and random match probabilities. We used an online interface that required participants to first correctly sort the statements in a given conclusion scale, and then place each statement on a single evidence axis that ranged from most support imaginable for same source to most support imaginable for different sources. We analyzed the data using both the raw values and a Thurstone–Mosteller model based on ordinal values. We found systematic differences between examiners and members of the general public, such that examiners distinguished between Identification and Extremely Strong Support for Common Source, while members of the general public did not. Statements that included numerical values tended to be placed lower than categorical conclusions, and members of the general public tended to place the highest categorical conclusion in each scale at the very top of the evidence axis. The results suggest that laypersons can distinguish between statements meant to represent moderate vs strong evidence, but tend to place categorical conclusions above statements that involve numerical values.

1. Introduction

In pattern comparison forensic disciplines such as fingerprints, firearms, toolmarks, and footwear, conclusions made by forensic examiners are often expressed as *categorical conclusions*. These are *categorical* in the sense that there are a limited number of possible statements in the scale, unlike a likelihood ratio that can, in theory, take on an infinite number of values. They are *conclusions* in the sense that they are making a statement about the origin of a questioned impression, such as "I identified this latent print to the suspect." These types of statements could be interpreted as a posterior, in that they are phrased as a statement about the likelihood of a proposition, rather than the likelihood of observing evidence given a proposition. Statements such as these have been criticized as being overinterpreted by laypersons [1] and perhaps too strong given the error rates observed in error rate (black box) studies [2].

In response to criticism that categorical conclusions are interpreted as absolutist in nature, the Friction Ridge Subcommittee of OSAC has begun to consider language that is more similar to a strength-of-evidence statement [3]. For example, 'Extremely strong support for common source' might be a replacement for 'Identification' in the fingerprint discipline. This revised statement is still a statement about a proposition and therefore is different than a likelihood ratio, which is a statement about evidence *given* a proposition. This revised statement has the potential to move the language in the direction of more nuanced articulation language and may avoid the incorrect assumption of perfect accuracy by jury members. However, this new language has not been tested to determine whether it is interpreted differently from traditional articulation statements.

Articulation language serves as a proxy or summary for the evidence that has accumulated in the mind of the examiner, and for this language to be properly calibrated it must be understood by both the forensic practitioner and the layperson. Should there be differences between how each statement is understood, this represents a mis-calibration of the evidence that might result in a jury member, defendant, or prosecutor interpreting the strength of the forensic evidence in different ways. While an examiner may qualify some conclusions on the stand during testimony [4], the vast majority of cases do not go to trial. Instead, these qualifications or hedges may be ignored or misunderstood by a

* Corresponding author at: Department of Psychological and Brain Sciences, 1101 E 10th St., Indiana University, Bloomington, IN 47408, United States.
E-mail address: busey@indiana.edu (T. Busey).

https://doi.org/10.1016/j.scijus.2022.10.003
Received 22 November 2021; Received in revised form 3 September 2022; Accepted 23 October 2022
Available online 28 October 2022

prosecutor or defense attorney, who may encourage a suspect to take a plea deal when the evidence may not support one and could result in the conviction of an innocent person.

Traditional categorical conclusions in the friction ridge discipline have included Identification, Inconclusive, and Exclusion [5,6]. Various organizations have criticized categorical conclusions as either prone to overinterpretation or implying absolute certainty [1,7,8]. Alternatives to categorical conclusions include likelihood ratios, random match probabilities, and strength of support statements. Likelihood ratios are numerical values that reflect the ratio of two probabilities: the probability of the observations given a same source proposition and the probability of the observations given a different sources proposition. Likelihood ratios are widely used in forensic DNA applications where probabilistic genotyping software provides a numerical result [9] and the propositions could include sub-source, activity and offence level propositions. Morrison [10] has argued that the likelihood ratio need not be quantitative but could be based on the expert's subjective evaluation. This approach is widely used in Europe [11,12] but has not seen widespread adoption in the US.

Strength of Support statements can express either the degree to which a set of observations supports a particular conclusion or the probability of the observations *given* one or more propositions. As such, these statements are similar to a likelihood ratio. Random Match Probabilities (RMPs) are the compliment of likelihood ratios if the observations have probability 1.0 under the same-source proposition. However, RMPs are potentially confusing because it may not be clear to a layperson whether 1 in 10 or 1 in a million is better [13] (and we see evidence for this in our data as well). RMPs also suffer from the fallacy of the transposed conditional, because a layperson may assume that a low random match probability implies common source when in fact only the other facts of the case allow for a complete characterization of the probability of the proposition *given* the evidence [14].

Work on juror understanding of evidence has focused on whether categorical scales or numerical likelihood ratios are better understood by members of the jury [15] and calls for a unified scale across disciplines [16]. Thompson and Newman [17] found that prior beliefs about a discipline affect evidence interpretation by mock jurors, suggesting that no one-size-fits-all approach is possible across all disciplines. A similar result was reported by Garrett, Crozier and Grady [18]. The choice of wording will also matter; Howes, Kirkbride, Kelty, Julian and Kemp [19] found that reports from forensic glass analysis would be difficult for a lay audience to comprehend. Martire, Kemp and Newell [15] reviewed the comprehension of various numerical and verbal statements and argued that not only must statements accurately reflect the strength of the evidence, but they must be phrased such that they are interpreted appropriately because they identified systematic biases in the interpretation of conclusion statements. Spellman [20] argued that probabilistic statements such as likelihood ratios and RMPs are very difficult for laypersons to understand even after extensive training and McQuiston-Surrett and Saks [21] found that qualitative statements were more damaging to the defense than quantitative statements. However, Thompson, Kaasa and Peterson [22] identified circumstances where laypersons made judgments that were in line with Bayesian expectations under certain conditions. In the end, it may be that a focus on the reliability of the evidence is more important than the exact phrase used to describe the conclusion [23]. The perceived reputation of the examiner and the sophistication of the methods may actually play a greater role than the testimony itself [24].

Within the fingerprint discipline, Garrett, Mitchell and Scurich [25] compared categorical statements against probabilistic statements and found that members of the general public viewed categorical and strong probabilistic statements similarly, but distinguished between strong and weak probabilistic statements. This suggests that there is a probabilistic statement that is viewed as equivalent to a categorical statement, but low probabilistic values imply less support for a common source proposition. However, members of the general public generally were not calibrated in absolute terms when interpreting probabilistic statements.

The goal of the present work is to establish how different articulation statements are understood by both fingerprint examiners and members of the general public. We will measure these strengths on both relative and absolute scales, with endpoints that are defined by hypothetical strengths to provide measurements relative to these endpoints, but also consider relative measurements to compare different statements to guide the development of new conclusion scales.

Thompson, Grady, Lai and Stern [26] addressed this question with a very straightforward design. They presented pairs of statements to members of the public (Amazon Mechanical Turk workers) and asked the participants: "Which of the following two conclusions would seem STRONGER if you heard it, meaning more convincing to you that the suspect is the source of the print?" [26]. This process requires that possible pairs must be compared, and in three different studies they compared a variety of different statements using both fingerprint and DNA scenarios. They modeled the choice data using a Thurstone–Mosteller model that produces strength parameters for each conclusion statement. They found that participants could distinguish between statements meant to imply higher strength of support from those meant to imply lower strength of support. They caution against the term 'match', and noted the potential misinterpretation of RMPs. The study found that categorical conclusions tended to be interpreted as providing strong support, which the authors found concerning. Overall the study provides direct comparison across different statements based only on relative judgements of strength of evidence.

A strength of this approach is that it relies only on ordinal relations, and by modeling these ordinal relations with a variant of a general linear model, they bootstrapped their way into a ratio scale of the various terms. This is a clever way to compute the relative perceived strengths of the evidence for the articulation statements that they could include in each experiment.

A downside to this approach is that it presents each statement in isolation, rather than as part of a complete scale. It may be, for example, that the perceived strength of a given statement is determined by the other statements in that scale. Our group previously observed this in the behavior of examiners using simulated casework comparisons [27]. We measured the use of the Identification conclusion in a scale that included only Inconclusive and Exclusion. We then compared this use of Identification to that in an expanded scale that included Support for Common Source' and Support for Different Sources. We found that when presented with a scale with additional categories, participants *redefined the meaning of Identification*, using it less often than when they had only three statements to choose from. Thus, the meaning of a statement may depend in part on what the other possibilities are in the conclusion scale, a conclusion that was recently replicated [28]. It is also possible that in any categorical scale, the top category is essentially interpreted in absolutist terms, but more quantitative or numerical scales may not be interpreted this way.

To compliment the Thompson, Grady, Lai and Stern [26] study and to extend it to new proposed language, in the current study we adopted a different approach. We designed an online interface to allow participants to directly manipulate different statements as shown in Fig. 1. Our approach extends existing methods designed to compare the relative strength of different forensic conclusion statements, but brings in the psychophysical and psychometric approaches described by Cohen, Ferrell and Johnson [29]. They grounded the judgments made by participants in a visual display, which improves the interpretation of small frequencies or proportions. They demonstrated that while typical s-shaped functions between estimates and ground truth proportions were observed (i.e. observers typically over-estimated small proportions), the biases in judgments of proportions were systematic across observers, and validates this approach for measuring values at even the extreme endpoints of a scale. Martire, Kemp and Newell [15] also took advantage of both numerical and visual displays to provide accurate estimation of proportions by their participants.

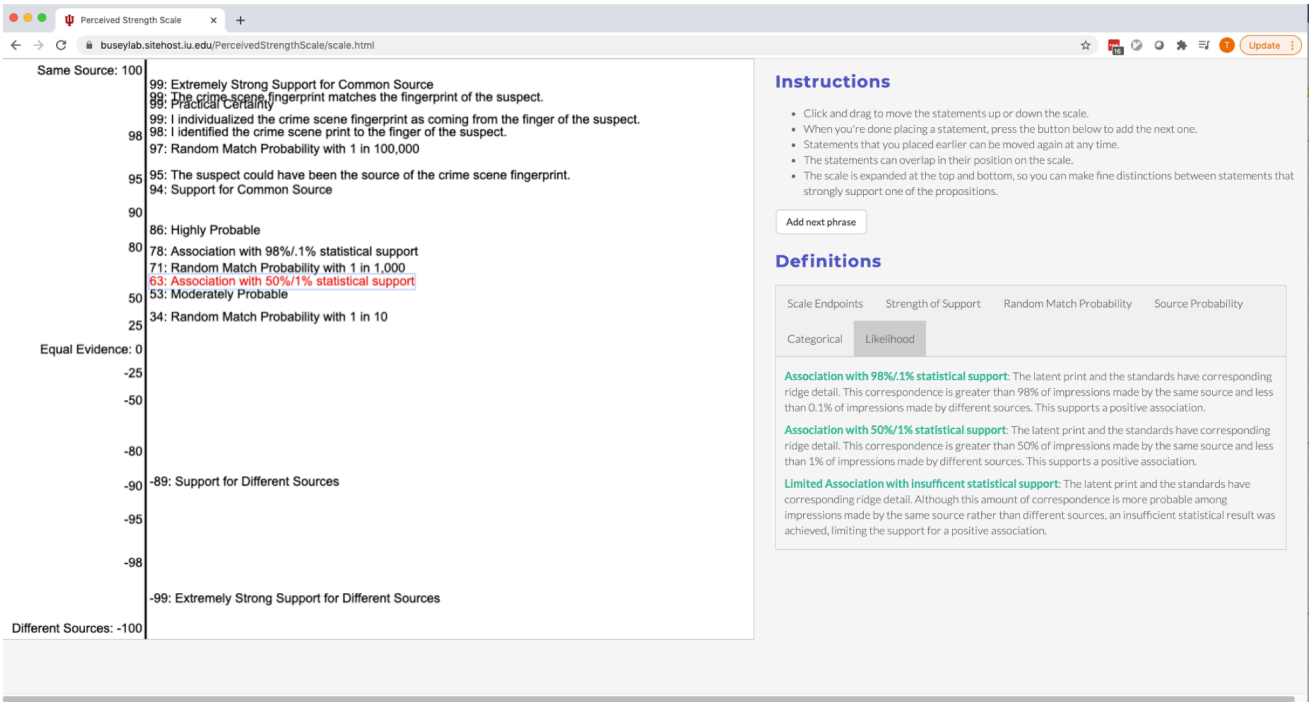

Fig. 1. Interface to measure the perceived strength of support for various articulation statements. Statement positions are hypothetical for purposes of illustration. Note that not all statements have yet been placed in this example, and the interface allows adjustments of all statements, not just the currently-added line (red text). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We designed the visual interface shown in Fig. 1 to help participants visualize both the relative and absolute evidence provided by different conclusion statements. The vertical axis is an evidence scale that ranges from −100 (most support imaginable for the different sources proposition) to 100 (most support imaginable for the same source proposition). To reflect the s-shape function noted by Cohen, Ferrell and Johnson [29], we expanded the scale at the endpoints. A similar scale was used by Martire, Kemp, Sayle and Newell [30], with the exception that their scale ranged from −10,000 to 10,000. The interface in Fig. 1 allows the participant to not only place each statement within the context of the other statements in that scale, but allows for comparisons across a broad set of statements and scales. The complete experiment available here and the reader is encouraged to visit the site to interact with the interface: https://buseylab.sitehost.iu.edu/PerceivedStrengthScale/.

To see just the interface part of the study, visit: https://buseylab.sitehost.iu.edu/PerceivedStrengthScale/scale.html which skips the consent form and the instructional video.

This interface was used to measure the perceived strength of evidence from three populations: fingerprint examiners (N = 126), members of the Indiana University and Bloomington Indiana community (N = 45) and jury-eligible adults from Amazon's Mechanical Turk (N = 143).

Table 1 illustrates the six different scales, each of which had various articulation statements, along with the shorthand statements that are used in tables and graphs below. The scales were taken from different styles of conclusion reporting within the forensic disciplines, and included recent language provided by the Defense Forensic Science Center (DFSC) of the US Army Crime Lab (USACIL) [31]. Note that this language has recently changed from the original formulation [32] and expresses two cumulative probabilities. Although this is not a true likelihood ratio (which is the ratio of two conditional probabilities given two propositions) we still refer to this language as the Likelihood scale in the experiment and analyses because it is a quantitative measure of the strength of the evidence.

The conclusion statements associated with each scale were placed sequentially after the conclusion statements for that scale were sorted, and the complete list of scales, articulation statements, and definitions are found Fig. 11. Further details of the methods are found below.

2. Method

The study was conducted using a web-based interface written in Javascript, with data stored remotely in a MySQL server. All data was collected according to the Human Subject protocol approved by Indiana University.

2.1. Participants

Fingerprint examiners were recruited from contacts gathered from forensic conferences, as well as placement on the CLPEX forum and snowball recruitment from those who had participated who were encouraged to recruit colleagues. We have no guarantee that all participants who indicated that they were fingerprint examiners were in fact members of the discipline, but we used a unique code on the web links to indicate that the link was obtained from the site that specifically recruited examiners or who was recruited by us. This allowed us to verify the provenance of the weblink, and we are reasonably confident that participants who indicated they were fingerprint examiner and use the discipline-specific link were members of the discipline. These participants were uncompensated. The only other inclusion criteria was that they were at least 18 years old and qualified to testify on fingerprint evidence in the United States. Of the fingerprint examiners, 3 reported they were a trainee, 11 reported less than 2 years of experience, 7 reported 2–4 years of experience, 13 reported 5–7 years of experience, 20 reported 8–12 years of experience, 37 reported 13–20 years of experience, and 31 reported more than 20 years of experience.

We had two other participant groups recruited from the general public. The first was a group of members of the general public from the Bloomington, Indiana community. These were personally recruited by the first author and consisted of family and friends, church and community members, former students, and close associates. The goal was to obtain data from participants who would take the task seriously, were

Table 1
Six scales along with the articulation statements and shorthand terms. The shorthand terms are used only in the figures and tables in the current manuscript, and were not used during data collection. Note that the DFSC language is labeled as the Likelihood scale although the language actually consists of two probabilities and is not a true likelihood ratio.

Scale	Term	Manuscript Shorthand
Traditional	Identification	Identification
	Inconclusive	Inconclusive
	Exclusion	Exclusion
Categorical	I individualized the crime scene fingerprint as coming from the finger of the suspect.	I individualized…
	I identified the crime scene print to the finger of the suspect.	I identified…
	The crime scene fingerprint matches the fingerprint of the suspect.	The crime scene fingerprint matches…
	The suspect could have been the source of the crime scene fingerprint.	The suspect could have been the source…
Random Match Probability	Random Match Probability with 1 in 100,000	RMP 1 in 100,000
	Random Match Probability with 1 in 1000	RMP 1 in 1000
	Random Match Probability with 1 in 10	RMP 1 in 10
Likelihood (DFSC/ USACIL)	Association with 98 %/0.1 % statistical support	Association with 98 %
	Association with 50 %/1 % statistical support	Association with 50 %
	Limited Association with insufficient statistical support	Limited Association
Source Probability	Practical Certainty	Practical Certainty
	Highly Probable	Highly Probable
	Moderately Probable	Moderately Probable
Strength of Support	Extremely Strong Support for Common Source	Extremely Strong Support for CS
	Support for Common Source	Support for Common Source
	Support for Different Sources	Support for Different Sources
	Extremely Strong Support for Different Sources	Extremely Strong Support for DS

motivated to make fine distinctions between different statements, and would not rush through the experiment. These participants were uncompensated. The only inclusion criteria was that they were at least 18 years old and jury-eligible in the United States.

The second group that were members of the general public were recruited from Amazon's Mechanical Turk. We used similar recruitment strategies for Mechanical Turk as in Thompson, Grady, Lai and Stern [26]. The inclusion criteria was that they were at least 18 years old and jury-eligible in the United States. We also required a HIT approval rate of greater than 97, and Number of HITs approved above 5000, and location in the United States. These participants were compensated $2 for their participation.

There were 126 fingerprint examiners, 45 Bloomington community members, and 143 Mechanical Turk participants. Table 2 has details on age distributions for the three groups, and Table 3 has details on the education distribution for the three groups.

2.2. Instructions

To address the information gap between fingerprint examiners and members of the general public, we produced an 8 min video explaining the nature of fingerprint comparisons, how the results are communicated, and how to use our interface. The video may be viewed at https://iu.mediaspace.kaltura.com/media/t/1_d7zcg4bg and a transcript can be found in Table 5. We also included a sorting task that demonstrated to participants the nature of each scale as described below.

2.3. Procedure

All participants (including fingerprint examiners) first completed an informed consent form and then viewed the video instructions. This video explained the general procedures of fingerprint comparisons as well as how the results of the comparison are communicated. The second part of the video demonstrated how to interact with the interface.

The scale endpoints are somewhat problematic because in theory there is no upper or lower bound on the scale, and this can be difficult for subjects to understand [29]. However, we still need to define these for the user interface, and therefore defined the endpoints of the scale as follows:

This evidence scale describes a range of support that different conclusions might imply. The top of the scale is the same source proposition, which is the most support imaginable for the proposition that the two impressions came from the same finger. The bottom of the scale is a different sources proposition, which is the most support imaginable for the proposition that the two impressions came from different fingers. In the middle is equal evidence, which is the point on this scale where the evidence for the two propositions is equal.

To familiarize participants with the task as well as how to interpret the endpoints of the scale, we gave participants a practice scale prior to introducing the remaining scales. We presented the dialog window shown in Fig. 2 in front of the main interface and ask participants to drag the statements to sort them. Although this practice task is trivial, some scales such as random match probability require some thought, thus necessitating this step in the experiment and this practice scale also familiarized participants with the sorting procedure.

Once the statements are in the correct order, a press of the Check button dismissed the dialog window and the participant viewed the main interface as shown in Fig. 1. The first statement of the current scale appeared in red in a random location in the scale and the participant was instructed to drag the statement to the location that corresponds to their estimate of the strength of support implied by that statement. Note that because the endpoints of this scale are ill-defined, we expanded the endpoints of the scale to allow for finer distinctions between phrases that provide a great deal of support for either proposition. For the practice task, we expected participants to drag the Same Source statement to the top of the scale, the Equal Evidence to the middle of the scale, and the Different Sources to the bottom of the scale. We did not use failure to drag these statements to these locations as exclusion criteria, but we had an extensive set of conditions that we did use to exclude participants for non-compliance with instructions as described below in the Participant Exclusion section.

After the participant finished placing each statement they clicked on the Add Next Phrase button, at which point the current statement changed color from red to black and a new statement appeared in a random location and in red text. The Add Next Phrase button was

Table 2
Age Demographics for the three groups.

Group	18–24	25–34	35–44	45–54	55–64	65–74	75+	Decline
Bloomington Community	12	3	5	7	8	1	2	0
Fingerprint Examiners	1	28	49	29	11	3	0	1
Mechanical Turk	2	27	29	19	9	6	0	0

Table 3
Education Demographics for the three groups. Highest degree obtained.

Group	Decline	Bachelor's	College Student	High School	Masters	PhD	Professional	Some College
Bloomington Community	0	10	9	1	9	6	1	2
Fingerprint Examiners	1	64	0	2	47	0	0	8
Mechanical Turk	1	41	2	13	9	1	2	23

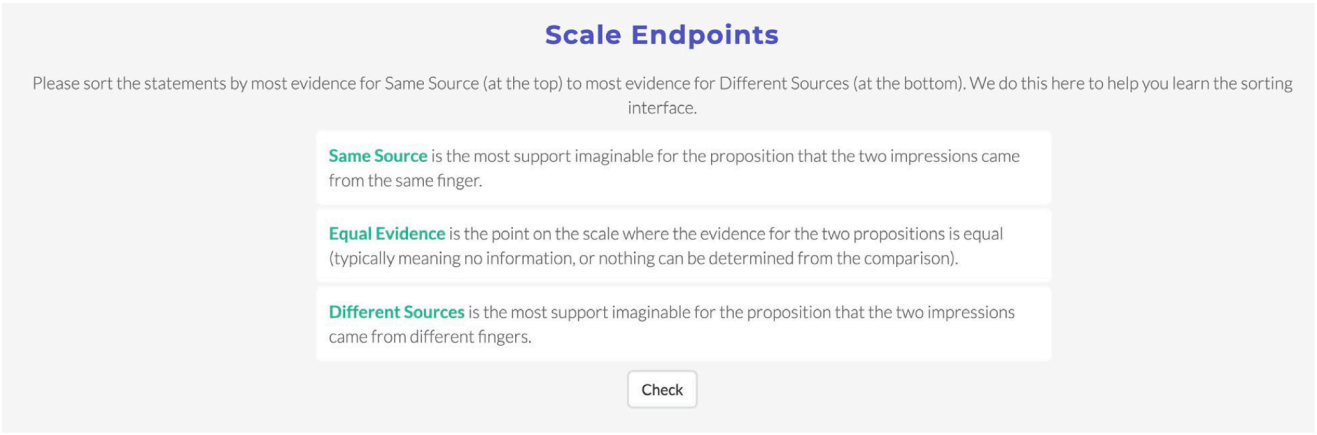

Fig. 2. Practice scale given to participants at the start of the experiment. The statements were presented in unsorted order and the participant was instructed to drag the statements such that the most evidence for same source is at the top, and the most evidence for different sources is at the bottom. The above figure is shown in the final correct sort order.

dimmed until the statement was moved to a location that was different from the starting location. Once all statements for the current scale were placed, the Add Next Phrase button changed to an Add Next Scale button. The practice statements were removed from the scale at the start of the first real conclusion scale. For the remaining scales, all statements remained on the screen until the completion of the experiment.

To verify that the participant had read and understood each statement in a conclusion scale, a dialog window containing all statements in random (unsorted) order was presented similar to that shown in Fig. 3. The order of the 6 scales was randomized across participants, so that the traditional scale shown in Fig. 3 appeared first for approximately 1/6 of the participants. The sorting task is important for each scale because it required participants to read each definition and compare each statement to the other statements in that conclusion scale. The Check button dismissed the dialog window only after the statements were in correct order. The only exception was the Categorical scale, where the ordering

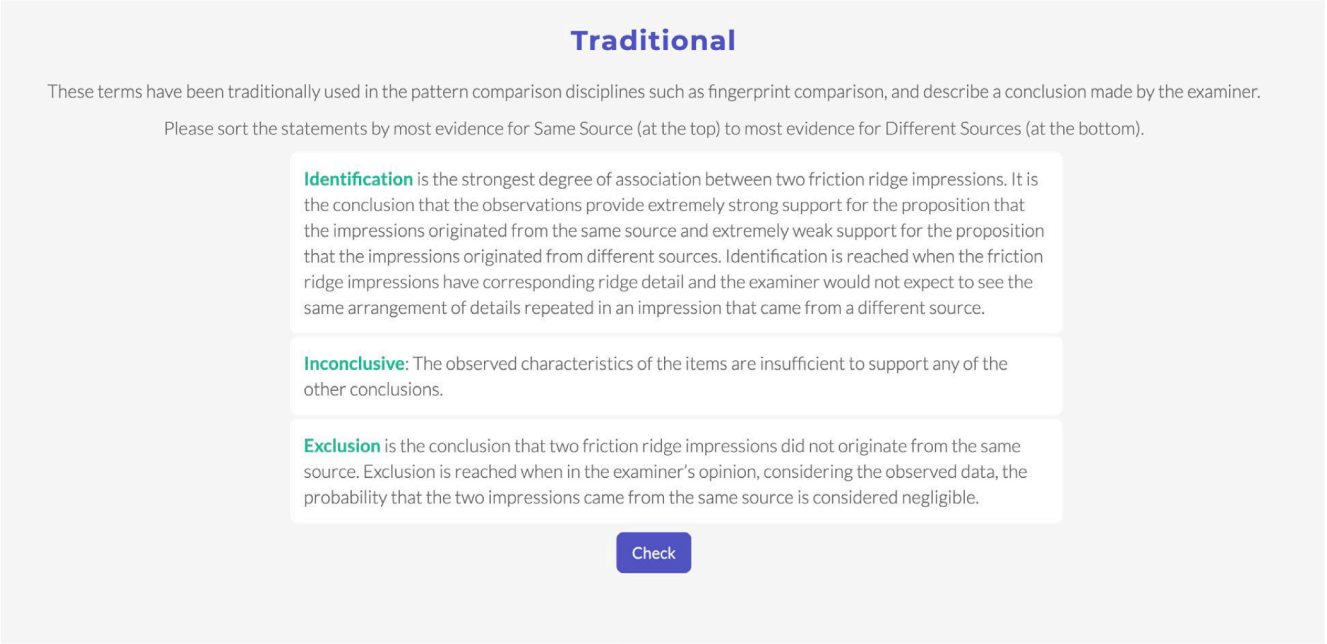

Fig. 3. Knowledge-check sorting task used for each scale (the Traditional scale is shown as an example). When each new scale is introduced, all of the statements associated with that scale are listed in random (unsorted) order. The participant must read each statement and then drag the statements in order such that the statement corresponding to the most evidence for same source is on the top, and the most evidence for different sources is on the bottom. The interface will only continue if the statements are sorted correctly.

between the "I individualized" and "I identified" statements is unclear and we did not want to bias participants by enforcing a particular order. Fig. 11 shows all scales in correct sort order.

After the statements for all six scales were positioned by a participant, a demographic questionnaire asked about age, level of education, experience with forensic examinations, primary forensic discipline, association with the justice system, and personal interactions with the justice system.

2.4. Participant exclusion

This experiment requires careful thought and logical thinking to appreciate both the meaning of each statement as well as its relation to other statements. If participants were to respond randomly, this would add noise to our data, which is compounded by the fact that our scale is bounded at −100 and 100. This means that any noise will be asymmetric, as it will tend to draw values away from the extremes. Rather than rely on the central tendency as the sole way to average out noise, we instead applied a series of criteria to evaluate subject inclusion as discussed below.

First, we applied a minimum time for adjusting each statement on the scale. If the minimum time between two successive clicks on the Add Next Phrase button was less than 2 s, we assumed that the participant was rushing through the experiment and we excluded that participant. We were particularly concerned about the Mechanical Turk participants, and the recruitment screen on the Amazon Turk site included the following paragraph:

Caution: This experiment requires careful thought and has built-in consistency checks. If you rush through the experiment (the data is timestamped) and respond without thinking, your data will not be useful to us. You will still be paid, but will be excluded from future studies from our group. Please do not continue unless you can take the time to make thoughtful judgments.

Second, our sorting task for each scale made it clear the order in which certain statements should maintain (with the exception of the Categorical scale). For example, we expect Identification to be placed above Inconclusive, and Inconclusive placed above Exclusion. Any violation of these relations was cause for exclusion. We adopted the same criterion for Extremely Strong Support for Common Source, Support for Common Source, Support for Different Sources, and Extremely Strong Support for Different Sources; any violation of this ordering was grounds for exclusion.

Finally, we noted violations of three other scales that tend to be confusing, but did not exclude participants based on these violations. These were the Likelihood, Random Match Probability, and Source Probability scales, and these are noted in Table 4 because they bear on the level of understanding of each scale (more confusing scales may have produced more violations even from conscientious participants). Note that a given participant could have more than one reason for exclusion.

This screening resulted in the exclusion of 4 of the 126 fingerprint examiners, 7 of the 45 Bloomington community members, and 51 out of the 143 Mechanical Turk participants. Table 4 lists the overall number of violations that lead to these exclusions, although the reader is cautioned that these numbers represent violations, not subjects, and a given subject could have produced multiple violations on a given scale by, for example, placing Exclusion above Inconclusive, and Inconclusive above Identified, which would have produced 3 violations. Violations for Likelihood, Random Match Probability, and Source Probability scales are shown in Table 4 but were not used to exclude participants. Numbers in parentheses indicate the number of unique participants who had at least one violation in that scale. In addition to these exclusions, we also excluded the second run of 12 Mechanical Turk participants who participated a second time despite instructions to avoid doing so (these 12 are not included in the 143 count in Table 4 because these were repeat subjects).

An early version of the code inadvertently failed to save the final placement of the last statement placed on the final conclusion scale. This issue was quickly corrected, and affected 3 members of the Bloomington community, 10 fingerprint examiners, and zero Mechanical Turk participants. Recall that the order of the six conclusion scales was random, so the missing data point for each of the 13 participants above was distributed across the six scales. This missing data does not otherwise affect the analyses reported below and only represents one out of the 20 statements placed by the affected participants. This missing data is easily accommodated by the GLM code because it does not need a full dataset from each participant to form the dominance matrix that serves as input to the GLM.

3. Results

We will present data aggregated across the two general public groups for comparison with the fingerprint examiners, and also provide separate comparisons between the two general public groups to demonstrate that they are quite similar despite different recruitment and selection procedures. While we will present raw distributions for visual inspection, the bulk of our statistical conclusions will come from the analysis of ordinal-transformed values as described in a subsequent section. We will also conduct targeted statistical analyses to address questions motivated by possible policy changes, but avoid blanket hypothesis testing due to the large number of possible comparisons and the alpha inflation that would result.

All data and analysis code is available at the OSF repository: https://osf.io/xmwqg/?view_only=f1b996eee77d45d0907ecebdaa27437d.

3.1. Raw values

Our first analysis presents the distribution of responses for each conclusion statement. Fig. 4 illustrates the distribution of responses for examiners and members of the general public (Mechanical Turk and Bloomington community participants combined). The abscissa is shown on the same log-transformed scale that the original interface used. The distributions reveal the following notable differences:

Examiners tend to place Identification at higher values than members of the general public, which tends to be true for other scales as well. There are large differences between the two groups for "I Identified…"

Table 4
Violation counts for the three types of participants, with unique number of participants in parentheses. Bold headings indicated violations that were sufficient to exclude a participant. The number of total violations counts violations, not subjects, and a given subject could have contributed more than one violation per scale. For example, the Strength of Support has 4 statements, which gives it more opportunity to produce violations from participants responding randomly, but the unique participant count in parentheses counts each participant only once despite multiple possible violations for that conclusion scale. Note that adding up the unique participants in a row will not equal the number of excluded participants because a given participant could have produced more than one type of violation.

	Number of Total Violations (Unique Participants)					
Subject Type	**Traditional (ID, Inc, Ex)**	**Strength of Support**	Likelihood	Random Match Probability	Source Probability	**Minimum Time Too Fast**
Fingerprint Examiners	0(0)	3(3)	7(7)	13(10)	9(7)	1
Mechanical Turk	45(30)	111(37)	53(31)	99(47)	54(35)	4
Bloomington Community	2(1)	3(2)	0(0)	8(4)	5(2)	4

Table 5
Transcript of Video Instructions. This transcript was auto-captioned from the video with light editing for transcription errors. Consult the full video for imagery and intonation.

This study looks at communicating evidence in forensics. Before we get started, I'd like to say a few words about the task, the interface you'll use, and why we feel this is important. Fingerprint examiners compare fingerprints obtained from crime scenes similar to these, because the fingerprints are often degraded, the impressions are compared by humans, not by computers. Fingerprints are unique, but so is every impression made by a finger. The job of a fingerprint examiner is to look at the latent impression collected from a crime scene and compare it against an exemplar impression collected from a suspect or retrieved from a computer database.
The fingerprint examiner must decide whether there is enough evidence to conclude that the two impressions were made by the same finger or whether there's enough evidence to conclude that the two impressions were made by different fingers. The amount of evidence is accumulated in the mind of the examiner, supported by charts and notes. The examiner has to communicate the results of that comparison to a detective, judge, or jury.
An examiner accumulates evidence in support of two propositions or hypotheses. The first is same source, the two impressions came from the same finger, and the second is different sources, the two impressions came from different fingers. Note that we typically never know which of these two propositions are actually correct, but we can accumulate evidence in support of each.
This evidence scale describes a range of support that different conclusions might imply. The top of the scale is the same source proposition, which is the most support imaginable for the proposition that the two impressions came from the same finger. The bottom of the scale is a different sources proposition, which is the most support imaginable for the proposition that the two impressions came from different fingers. In the middle is equal evidence, which is the point on this scale where the evidence for the two propositions is equal.
Different comparisons might result in different levels of support for the two propositions. If the crime scene fingerprint is distorted or only a partial copy of the finger, there may not be much detail to work with when doing the comparison similar to these. Other impressions might be higher quality, and this might result in more evidence in support of one of the two propositions.
To communicate the results of the examination, the fingerprint examiner typically relies on a conclusion scale, which has various statements that communicate different levels of support for the two propositions. For example, the two fingerprints below are obviously different, suggesting more support for the different sources proposition than the same source proposition. The one on the left is a whorl. The one on the right is a left loop. In other cases, there might be a lot of detail in agreement between the two fingerprint impressions, suggesting more support for the same source proposition than the different sources proposition as shown with these images here.
Fingerprint examiners have various phrases to express the strength and support for the two propositions. It is important that the phrase they use is interpreted properly by others, such as detectives, judges, or jury members. The goal of this study is to allow you to express how you interpret the meaning of different phrases if spoken by a fingerprint examiner.
We're going to show you different phrases and ask you to place them on an evidence scale. Here we've added numbers where 100 represents the strongest evidence imaginable for the same source proposition. Minus 100 represents the strongest evidence imaginable for the different sources. Zero represents equal support for the same source and different sources propositions. We will use this scale to help express how much support you believe each conclusion statement implies about the two propositions. Note that the scale has stretched at the endpoints to help you make fine judgments about different statements that are close to each proposition.
To get started, imagine that you were on a jury and the fingerprint examiner has presented fingerprint evidence along with a specific phrase that expresses their conclusion. We're going to show you a series of phrases and asked you to tell us how you would interpret the level of support each phrase implies for the two propositions, each were spoken by a fingerprint examiner.
Let's go through the interface and I'll explain how it works. Once you've finished that video, you'll see a screen that looks like this. This is our sorting interface that allows you to read each one of our statements, as well as the definitions for each of those statements. And then to sort them in terms of the order for most evidence for same source at the top, two most evidence for different sources at the bottom. So I'll move same source up here and then different sources down here. And now they're in the correct order. And this is just for practice to learn this interface. And then you'll click the Check button. And if it's correct, you'll get to see this screen right here. Click the Start button, and then move the same source statement up to the top here. This is again just for practice to learn our interface. Me, move the IPO evidence to here, and then move the different sources all the way down here. So next you go on to the next scale. Your scale might look different than this one. But what we'd like you to do is to read each statement and then the definitions, and then sort the statements by most evidence for same source at the top to most evidence for different sources at the bottom. So I'll move this one up here. That seems to sort them there, and then click the Check button. And if they're correct, then you'll move on to the next screen. This is where the experiment actually begins.
So what I'd like you to do is to read this statement, review the definition if you need to, and then think about the location of this statement along the evidence axis from same source proposition, two different sources proposition. Move this statement to a location that corresponds to the strength of the evidence for same or different sources that you believe that statement implies if stated by a fingerprint examiner in court. So I won't bias you by telling you where I would place this. I would say that just move it to a location that satisfies that strength of the evidence that they feel like this implies a cup. And once you've placed that, click the Add next phrase button. Then you'll move this one to the correct location, the correct location that you infer from this statement, referring back to the definition, if you need to, a couple of things about using this scale. First of all, the different statements can overlap. That's certainly fine. The second thing is that you should preserve the order. So if you feel like one statement is slightly higher in terms of strength of the evidence, you should place it above another statement. And you can go back and move different statements if you need to, even though they're no longer red.
We would like you to treat this as a scale that goes from a 100, which is most evidence for same source that you could ever imagine, to minus 100, which is most evidence you could ever imagine for different source proposition. 50 is midway between equal evidence and same source and minus 50 is about midway between different sources and equal evidence. Use that scale as you like. When you're done with the phrases for a particular scale, it will go on automatically to the next scale. Once you've worked your way through all of the scales, there'll be a screen with some demographics and you can work your way through those, and then you'll be done with the experiment. We feel like this experiment is really important in terms of helping forensic examiners think about how to make a conclusion that is interpreted properly by a judge or jury, or a detective, and does so in a way that accurately represents the strength of the evidence. I appreciate you thinking carefully about the definitions of each statement and thinking about where would buy on the evidence axis from evidence for the different sources proposition all the way up to evidence in favor of this same source proposition. Thank you so much for your help with this.

and "I Individualized…", which may be related to our sorting task and how participants treat the highest statement in each scale as discussed in the Discussion section.

The two groups that constitute the members of the general public performed remarkedly consistently, as illustrated in Fig. 5. There appear to be few systematic differences between the two groups, which suggests that, despite the differences in recruitment strategies, the overall behavior of members of the general public is fairly consistent. For all further analyses we have aggregated these two participant types into the general public group.

The variance (standard deviation) of the placement of each statement across participants is a measure of the (in)consistency across participants. Fig. 6 plots the standard deviation of each measure combined overall all participants against the median value for that statement (we produced a version that separates examiners from the general public, but the graph is hard to interpret and not very illuminating). Low values on the ordinate indicate high consistency. Some low values are expected by the endpoints of the scale because phrases such as Exclusion and Identification are almost always placed near the endpoints and this will give low standard deviation values for these terms. The standard deviation for Inconclusive should also be low because it typically is placed in the middle of the scale. Higher values reveal marked disagreements between participants, including all of the Random Match Probability statements, as well as Limited Association from the Likelihood Ratio scale. However, Support for Common Source and Support for Different Sources demonstrate fairly good consistency, which makes them good candidates for inclusion in scales designed for casework.

The fingerprint community is currently contemplating a change in terminology from Identification to Extremely Strong Support for Common Source. To determine whether these two phrases are interpreted as the same or different, we conducted Kolmogorov-Smirnov tests on the distribution of responses for each term. We found that Examiners readily distinguished between these two statements ($D = 0.395$, $p < 0.0001$), demonstrating that they agree that Identification implies stronger evidence for same source than Extremely Strong Support for Common Source. However, members of the general public do not share that view,

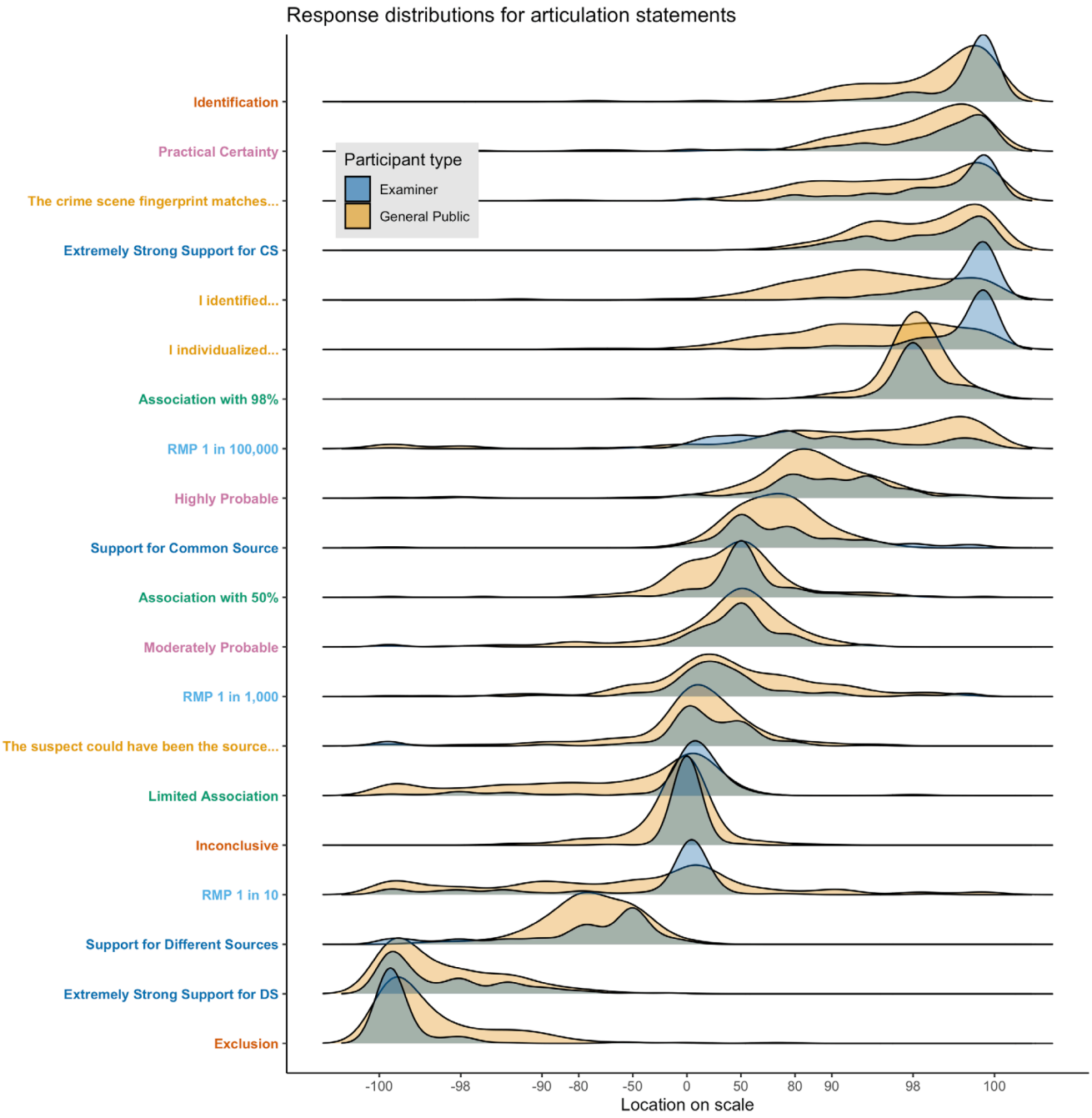

Fig. 4. Ridge plot comparing Examiners to members of the General Public. The different conclusion statements are summarized on the left, and the distribution of responses is illustrated with the colored ridge plots. Note that the evidence axis scale is expanded to mimic the scale used by participants (see Fig. 1). The statements are sorted by the median of each statement across all groups. The data is smoothed with a Gaussian kernel, which is why there are values above 100 and below −100.

and demonstrate little evidence that they interpret these two statements differently ($D = 0.148$, $p = 0.12$). Thus, it appears that members of the general public view these two statements as implying approximately equal strength of evidence despite fingerprint examiners' belief that Identification implies stronger evidence for same source than Extremely Strong Support for Common Source.

In a companion paper [28], we tested examiners on casework like comparisons using either Identification or Extremely Strong Support for Common Source, and found that examiners were *less* likely to use Extremely Strong Support for Common Source than Identification. The data from casework seems to suggest that examiners believe that Extremely Strong Support for Common Source should only be reserved for the pairs with the most support for common source, which appears to contradict their beliefs when placing statements on the present interface. This contradiction is discussed more fully below.

3.2. *Ordinal-transformed values*

The raw values presented in the previous section focus on the absolute placement of each value along the evidence scale, but different participants may have interpreted this scale differently yet preserved ordinal relations relative to other participants. Arguably, what is important is the *relative* placement of each statement, which can be captured by the ordinal relations of the items for each participant. This approach was used by Thompson, Grady, Lai and Stern [26] when they directly compared pairs of individual statements. The authors were kind

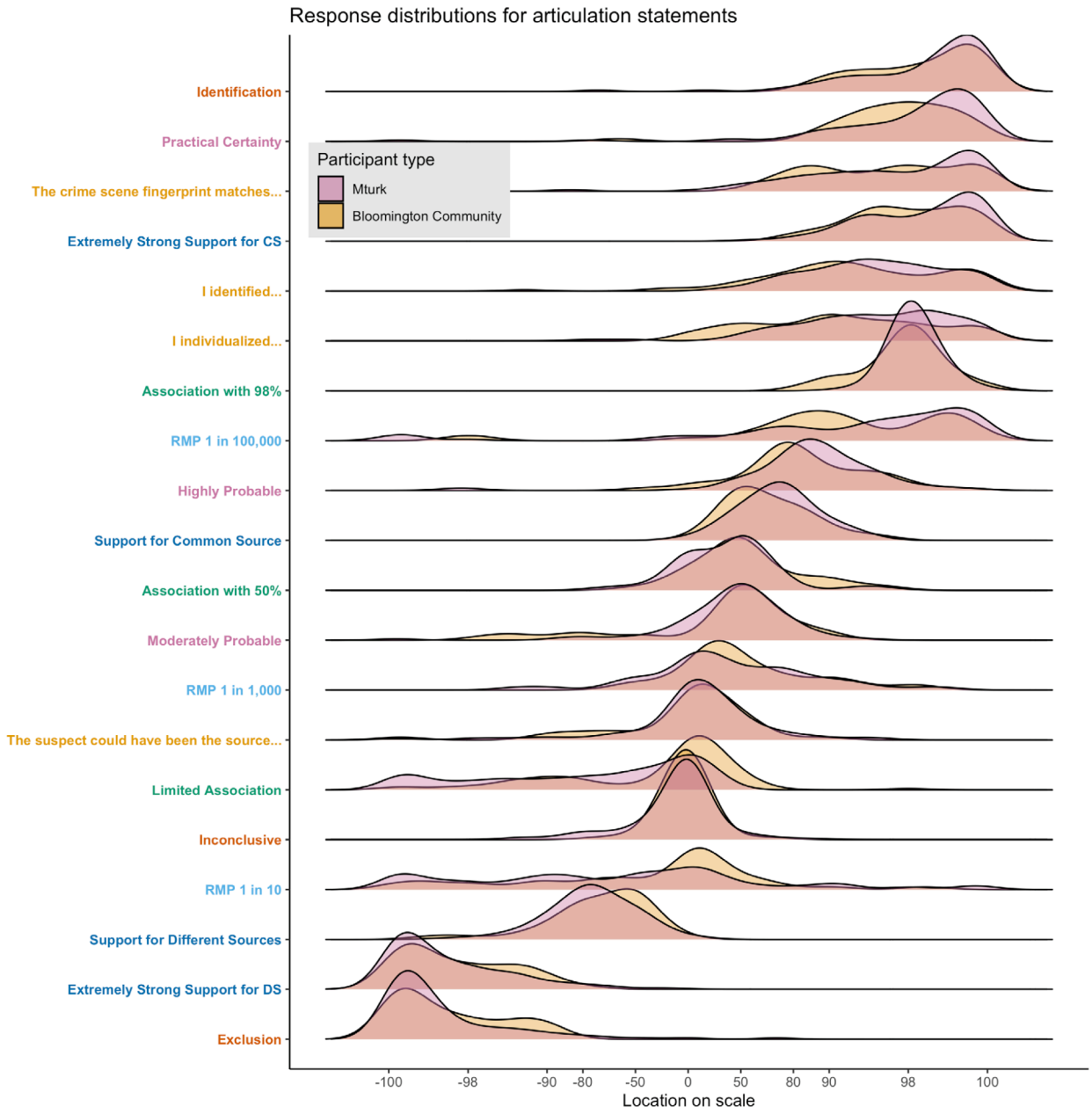

Fig. 5. Ridge plot distributions for the two groups that constitute the members of the general public.

enough to share their analysis code, and we adopted this approach to analyze our ordinal relations as well.

To convert the ordinal relations to a ratio-scale response metric, we first used the raw values of each participants to create a *dominance matrix* across participants in each group. This matrix counts the number of times a given statement is placed above any other statement. With 20 statements, this produces a 20x20 matrix with blanks on the diagonal, and each cell is a count of the number of participants who placed the statement for that row above the statement for that column. This procedure is performed separately for each participant type. This matrix is then fit using a Thurstone–Mosteller model, which is implemented as a variant of a general linear model. This model produces a parameter estimate for each statement that corresponds to the overall strength of evidence inferred from the dominance matrix for that statement (see Thompson, Grady, Lai and Stern [26] for more details on this approach).

This approach relies solely on the dominance (ordinal) relations for each participant, and bootstraps these relations into a ratio-scale metric that represents the inferred strength of evidence for each statement. This method requires one statement to act as a reference point, and for this we chose the Inconclusive statement as it is centrally located along the scale and relatively non-controversial in its placement. It also showed marked consistency in Fig. 6 as measured by the standard deviation of placement by participants.

The results of the analysis is a General Linear Model (GLM) coefficient that represents the inferred strength of evidence for same source as measured by the dominance matrix. Fig. 7 illustrates the coefficients for fingerprint examiners, sorted by the value of the coefficients. Identification is seen as implying the most evidence for common source, with Extremely Strong Support for Common Source much lower. This is consistent with the statistical analysis of the raw results described in the previous section, along with the data shown in Fig. 4. Fingerprint examiners consistently place Identification above Extremely Strong Support for Common source.

Numerical scales such as the Random Match Probability statements and the Likelihood statements (e.g. Association with 98 %) were placed consistently below Identification and Extremely Strong Support for

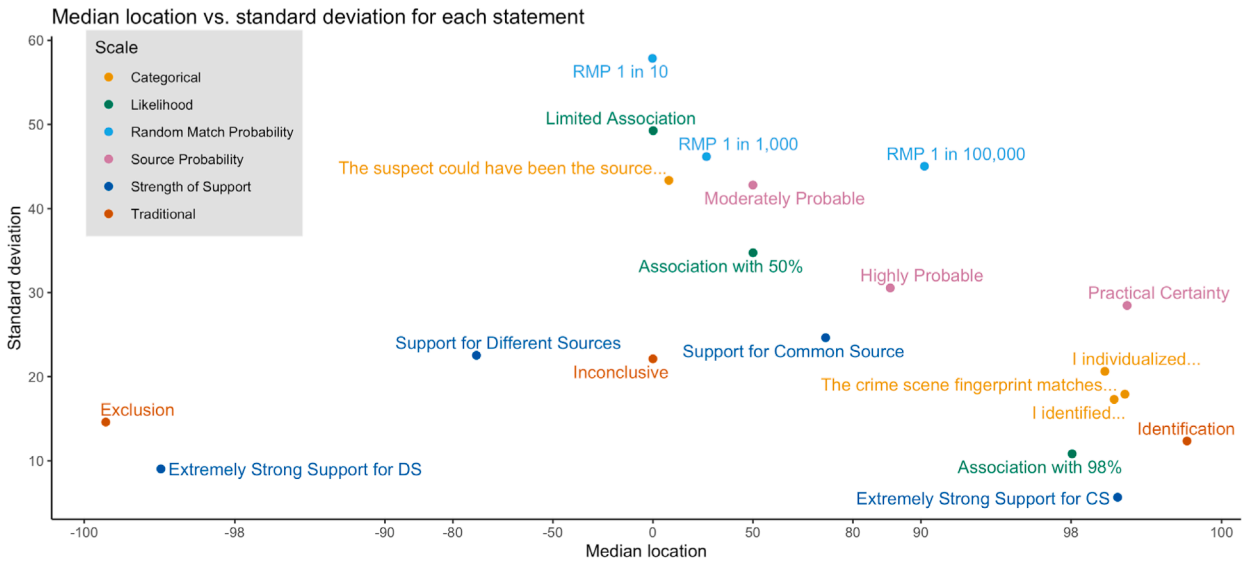

Fig. 6. Scatterplot comparing the median for each conclusion statement against the associated standard deviation for that term, combined across all participants. Values higher in the graph are associated with greater variability. Some terms toward the ends of the scale have low variability and therefore fairly high agreement across participants. Terms in the Random Match Probability, Likelihood, and Source Probability scales tend to have higher variance, suggesting that participants did not agree with each other on these terms.

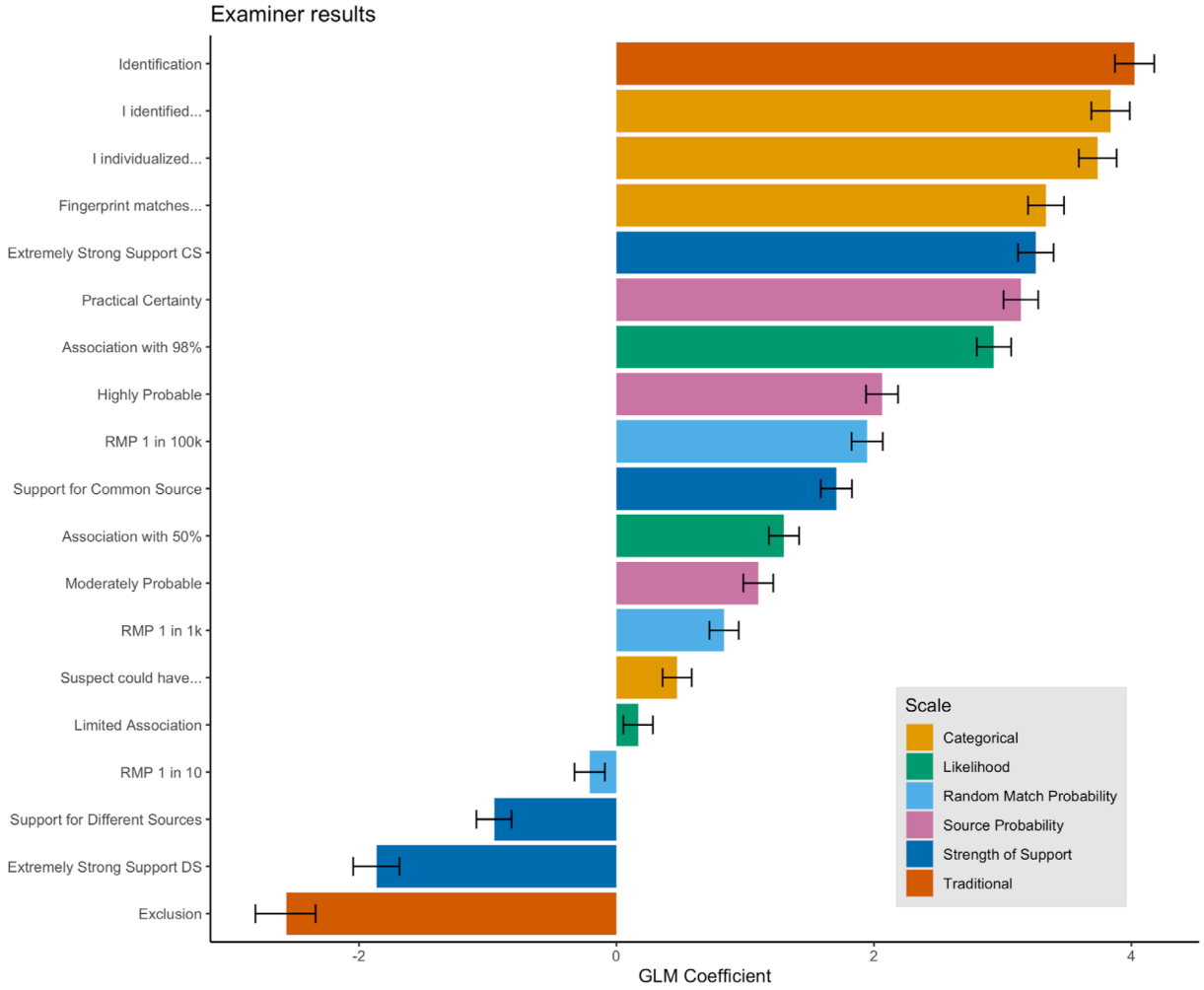

Fig. 7. Generalized Linear Model (GLM) coefficients for each statement for fingerprint examiners. Error bars represent 95 % confidence intervals around the point estimate for each coefficient.

Common Source. It may be that a numerical estimate tends to reduce the amount of support for common source implied by the statement.

Fig. 8 presents the coefficients for members of the general public. Identification and Extremely Strong Support for Common Source are seen as implying the most support for common source and appear virtually identical in terms of that support. As with the previous analysis, members of the general public do not seem to distinguish between these two statements in terms of the strength of support they offer for common support.

For direct comparison between the two participant types, Fig. 9 provides a scatterplot of the coefficients for each scale from both groups, with error bars representing 95 % confidence intervals around the coefficient estimates. If the two groups interpreted the statements equivalently, all points would lie on the diagonal. Instead, we see some notable deviations. First, Extremely Strong Support for Common Source, Practical Certainty, RMP 1 in 100 k and Association with 98 % are all higher for the general public than for examiners. Second, "I Identified…" and "I Individualized…" are both lower for the general public than for examiners, despite the fact that they are treated as virtually equivalent by examiners (and probably should be, given the wording of the statements). In the Discussion section we develop a general set of (somewhat speculative) explanations that may address these differences across participant types.

Fig. 10 compares the two types of general public. In general, we find very close correspondence between the two groups, as evidenced by the tight grouping of the points along the diagonal. There appear to be no notable deviations from the diagonal, which validates our aggregation of the two types of general public participants in comparisons with examiners.

4. Discussion

It is important to reiterate that this study looks only at perceptions of relative strength of various articulation statements and conclusion scales, and does not consider normative comparisons with the actual strength of evidence (with the exception of indirect inferences from error rate studies). The conclusions below bear on the ongoing policy discussions on how different phrases are likely to be interpreted by fingerprint examiners and members of the general public. The results from both the analysis of the raw data as well as the general linear model fits are fairly consistent, and there are four conclusions that we consider most important.

1) There are large differences between examiners and members of the general public in terms of their interpretation of Extremely Strong Support for Common Source. As illustrated in Fig. 9, members of the general public view this statement as virtually identical in strength to Identification. However, examiners place Extremely Strong Support for Common Source (ESSCS) much lower than Identification, demonstrating that they view ESSCS as implying less evidence overall than Identification when it is used. However, as discussed in the companion paper [28], examiners tend to use Extremely Strong Support for Common Source less often than Identification in casework-like comparisons, which

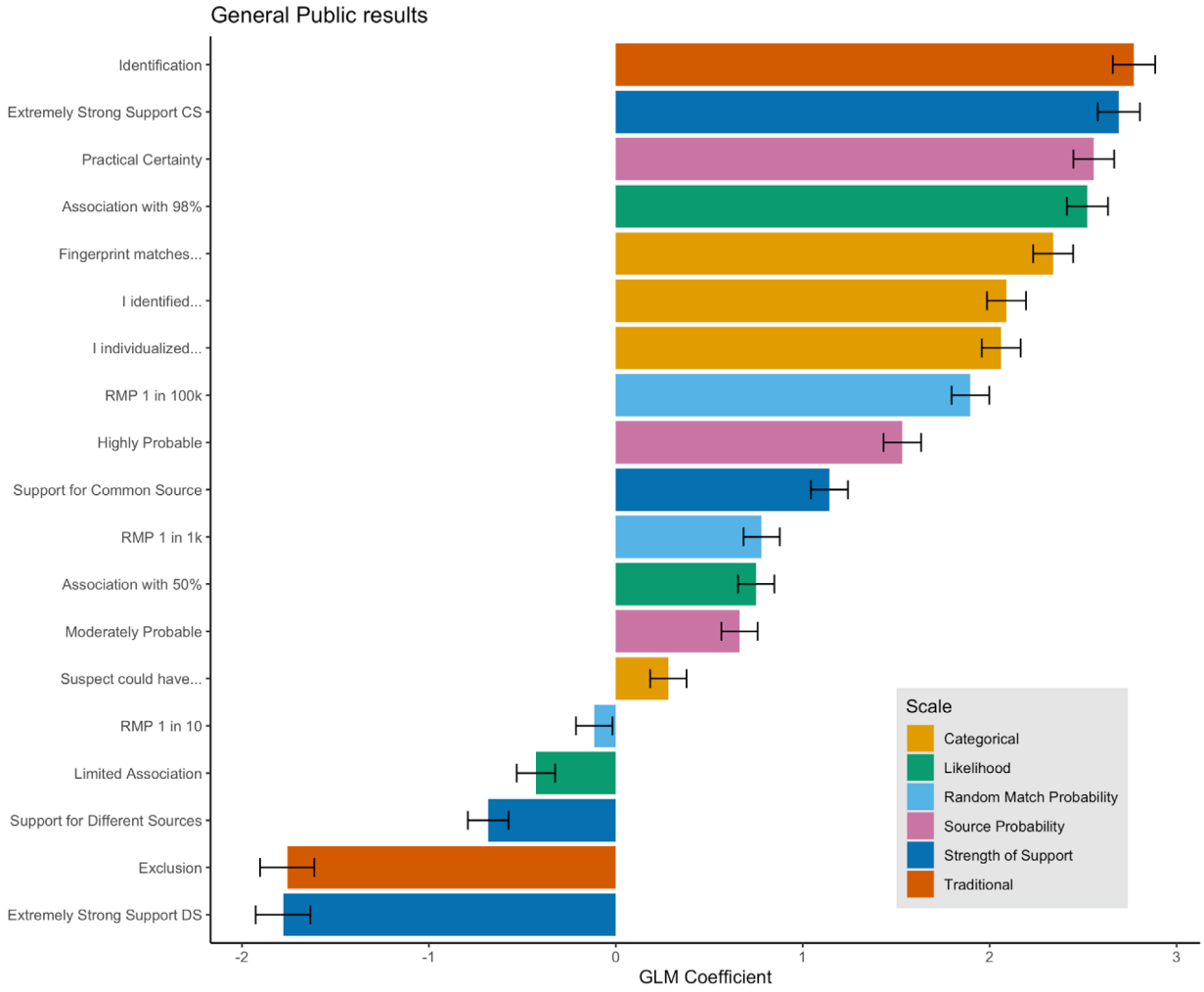

Fig. 8. Generalized Linear Model (GLM) coefficients for each statement for members of the general public. Error bars represent 95 % confidence intervals around the point estimate for each coefficient.

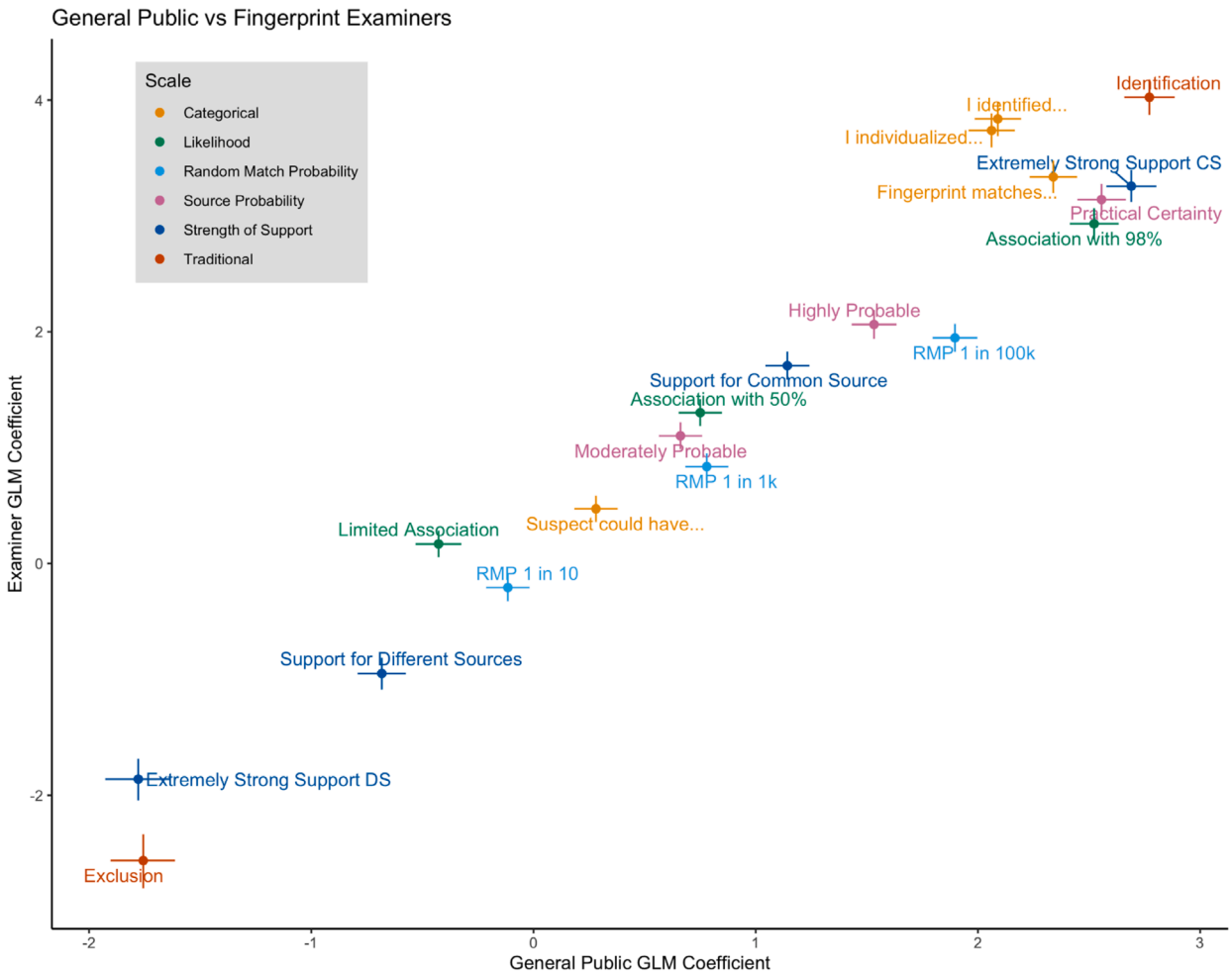

Fig. 9. Scatterplot comparing the coefficients of members of the General Public (abscissa) against the Examiners (ordinate). Error bars represent 95 % confidence intervals around each coefficient estimates.

represents a contradiction between how Extremely Strong Support is interpreted (in the present study) and how it is used in casework-like comparisons (in the companion paper [28]). The unfamiliarity of the phrase may contribute to its sparse use with comparisons, or the full gravity of the phrase may only become apparent when doing comparisons. Alternatively, in the present study, some examiners may cling to the anachronistic belief that Identification reflects "to the exclusion of all others", and therefore place Identification above ESSCS. Of course, neither phrase has been calibrated against the actual strength of the evidence and the current study only addresses relative strength.

2) Both examiners and members of the general public readily distinguished between the Identification, Association with 98 %, and Random Match 1 in 100,000. Although these are the top statement of each scale, the fact that both subject groups distinguished between them illustrates that they were capable of interpreting the statements and didn't just place the highest statement from each scale at the top of the evidence axis. These likelihood-ratio style statements appear to be interpreted as implying less evidence than Identification. However, error rate studies demonstrate a false identification rate of 1 in 1000 [2], which we would roughly interpret as a likelihood ratio of around 1000. If we convert a random match probability of 1 in 100,000 to a likelihood ratio, we obtain a likelihood ratio of 100,000. This would suggest that both members of the general public and fingerprint examiners give less weight to an RMP of 1 in 100,000 than the term Identification, where the error rate data itself does not seem to support an RMP of 1 in 100,000. Any comparison between categorical statements and those that include numerical values will of course depend on the exact numerical values, but the values used in the present experiment represent fairly strong evidence relative to most friction ridge evidence given error rates of 0.1 % [2], and therefore we find it surprising that categorical conclusions were placed above numerical values. Some of these differences may come from the fact that RMP values are perhaps confusing to both members of the general public and fingerprint examiners, where such values are not traditionally used to articulate the results of a comparison.

3) Examiners and the public perceive the strength of fingerprint evidence to be less when it is accompanied by the DFSC language than when it is accompanied by categorical statements claiming an identification. However, participants tended to place Associated with 98 % at a value of 98, and Associated with 50 % at a value of 50. This suggests that they adopted only a very superficial understanding of these conclusion statements. It is unclear where exactly these statements should fall on the scale, because the strength of the evidence depends on both the sensitivity and specificity values given in the statement and have no direct relation to the numerical values on our scale. This suggests that further explication is required for a consumer to understand the statement, as the confusion above was shared by both examiners and members of the general public. However, presenting only a verbal equivalent is not advised [33]. Numerical approaches (likelihood ratio and RMP) tend to be viewed as weaker than categorical conclusions or statements that do not include numerical values. Clearly this depends on the numerical values used, but a RMP of 1 in 100,000 seems to exceed the error rates found in fingerprint error rate studies (with erroneous identification rates of 0.1 %). Note that this result is different than the one

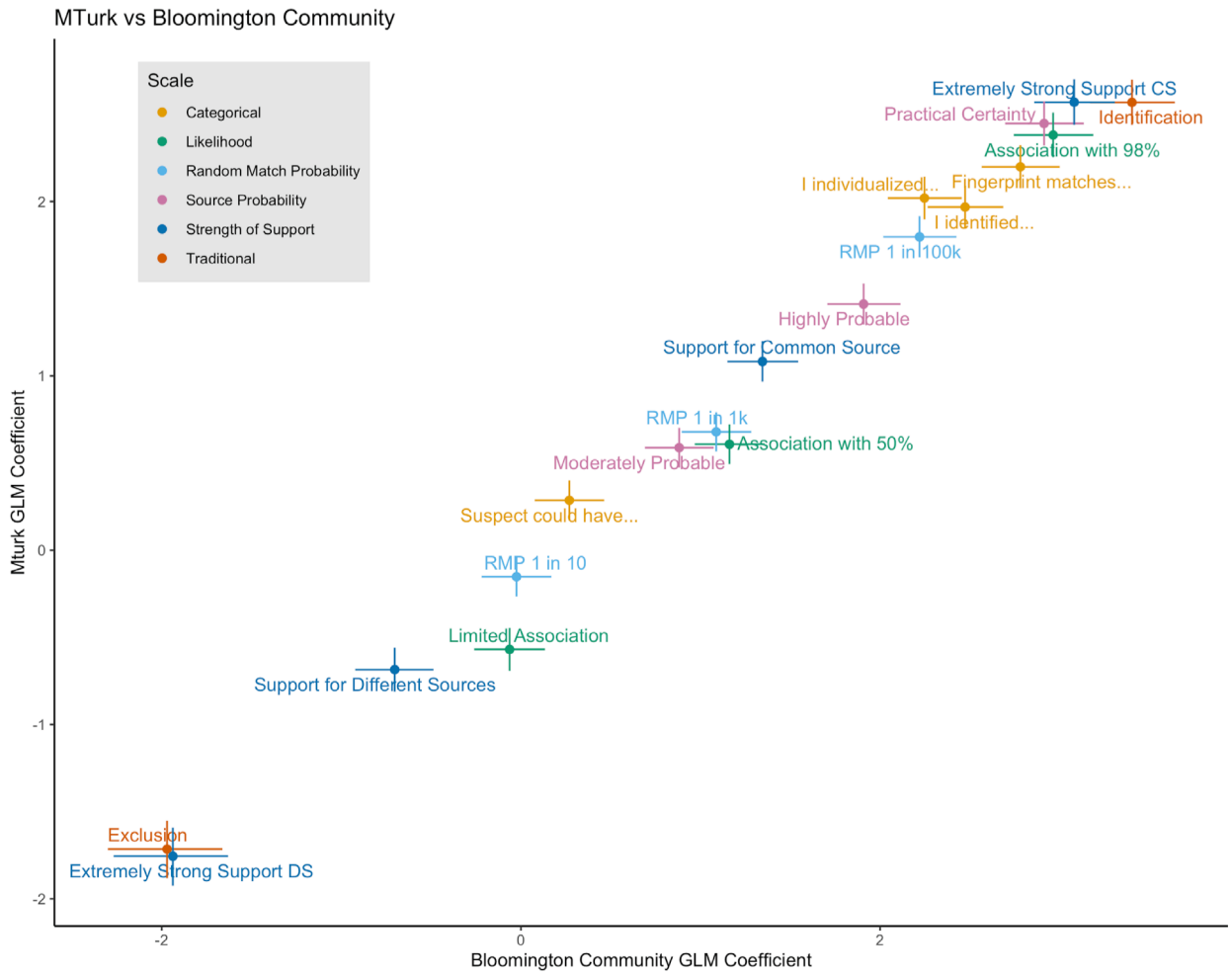

Fig. 10. Scatterplot comparing the coefficients of the two types of members of the General Public. Error bars represent 95 % confidence intervals around each coefficient estimates.

obtained by Garrett, Mitchell and Scurich [25], who found that strong probabilistic statements were seen as equivalent to categorical statements. It is unclear whether the differences are due to the language changes (Garrett, Mitchell and Scurich [25] use the original DFSC statements [32] whereas we used updated statements [31], or possibly due to the differences in methods.

4) The results from the Categorical Scale from the general public (see Fig. 8) are perhaps a bit surprising. Both General Public groups placed I Identified and I Individualized below Identification, despite the fact that the wording is almost identical (see the yellow terms in Fig. 10 for an example of the consistency of this finding). Examiners, on the other hand, placed I identified and I Individualized on par with or slightly below Identification (see Fig. 7). One difference between Identification and I Identified/I individualized comes from the slight phrase differences that highlight the fact that the latter statements are personalized to the individual examiner (e.g. "I…) as opposed to simply the result of the comparison process. This difference has been described as a difference between internal mode where the statement is based on the expert's personal knowledge, and external mode where the source of uncertainty is out in the world. In the literature, listeners give higher belief to internal mode statements [34–36]. However, this explanation cannot account for the present results, because our participants placed the external mode statement ("Identified") above the internal mode statements ("I identified…). In our data, either the participants were taking cues from the sorting task where internal mode statements were not sorted, or they were discounting the internal mode statements because they were seen as more opinion-based because they are attached to the examiner rather than being expressed as the result of a comparison. Note that members of the general public placed "I identified" above "I individualized" 65 times, and reversed this ordering 62 times, so they view these two statements as more or less equivalent, at least collectively. Fingerprint examiners placed "I identified" above "I individualized" 50 times, and reversed this ordering 49 times, so they also view these two statements as quite similar.

We offer one general speculation that may account for all of the results we observe. First, members of the general public may assume that the highest term in each scale should be essentially equivalent and placed near the top of the scale. This would explain why Extremely Strong Support for Common Source was treated as equivalent to Identification by members of the general public. However, statements that include numerical values (the Likelihood and Random Match Probability scales) tend not to follow this pattern, suggesting that numerical qualifications of the strength of the evidence reduce the implied support for common source. When given a phrase but no indication of which term is the highest as in the Categorical scale, general public participants exhibit more variability in their interpretation of the strength of evidence.

We conclude with a final set of recommendations. We believe that the approach proposed by the Defense Forensic Science Center offers many strengths, but also some weaknesses. A strength of the approach is that it is grounded in the physical evidence. In addition, the use of the term "Association" as opposed to "Identification" implies strength of support rather than a posterior conclusion. However, both fingerprint examiners and members of the general public were somewhat naïve in

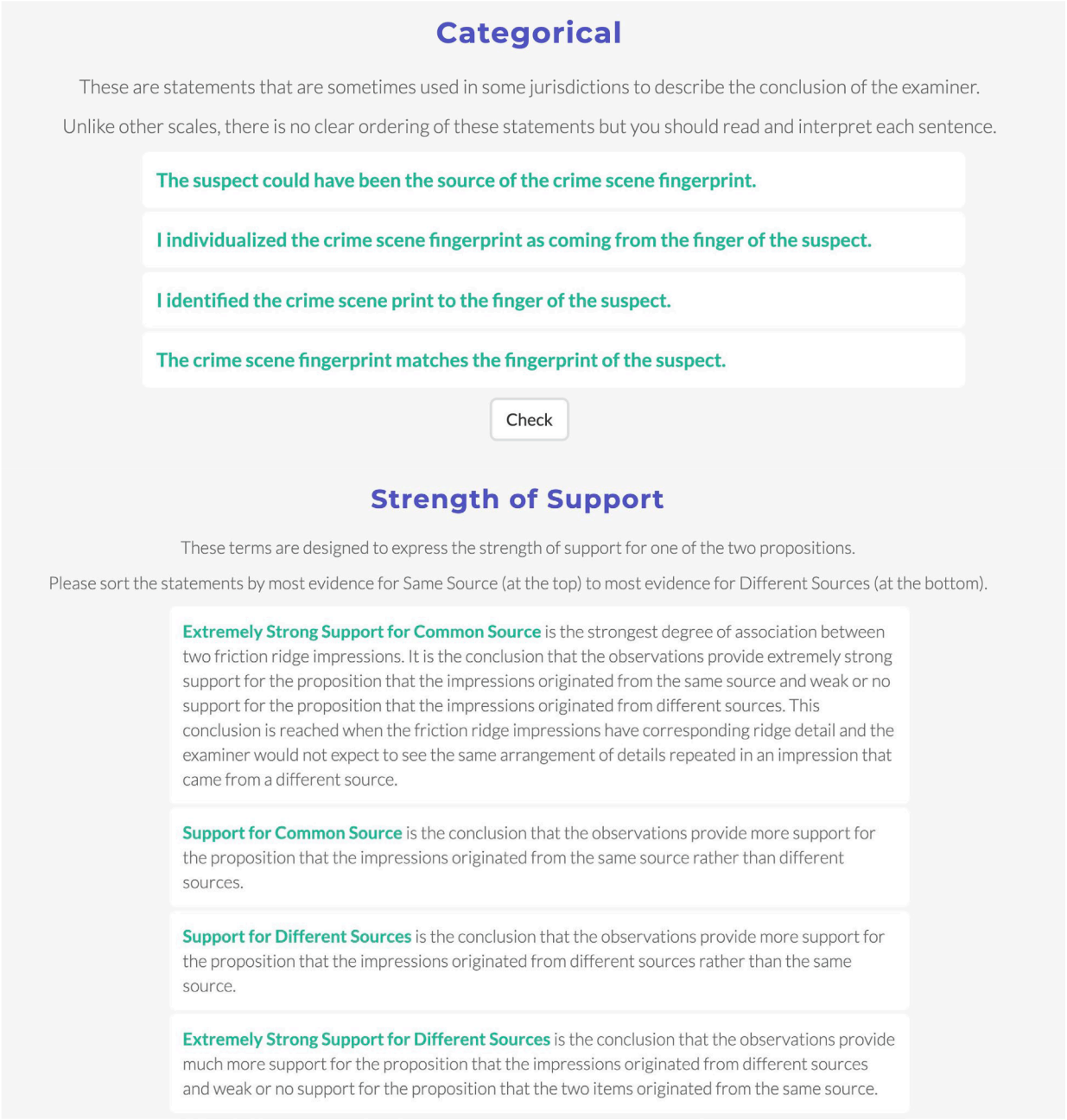

Fig. 11. All terms for each scale, correctly sorted. These scales were shown in random order for each participant, and most required the participant to correctly sort the items to demonstrate a general understanding of the terms. Note that the first scale was used as a tutorial for the sorting task.

their interpretation of the statement, tending to place the statement at a value of 98 (see the Association with 98 %/0.1 % statement in Fig. 4) and the 50 %/1 % statement at a value of 50. The statement include statistics for both common source and different sources propositions (the full definition includes "This correspondence is greater than 98 % of impressions made by the same source and less than 0.1 % of impressions made by different sources"). The two numbers somewhat independent, and different distributions of minutiae could give an identical first number and a different second number (e.g. 98 % and 0.5 %). However, all three types of participants tended to focus only on the first number when placing statements. We considered, but rejected, adding an additional phrase that included a 98 %/0.5 % comparison to see how participants would treat this new statement, which logically would be placed below the 98 %/0.1 % statement, but decided that this would be too confusing to participants. In hindsight, such an inclusion might have revealed the superficial nature of the scale placement for these statements, and it might have encouraged some participants to take a more nuanced approach to the Defense Forensic Science Center statements.

The Defense Forensic Science Center articulation language has the additional advantage that it explicitly considers competing hypotheses because it provides separate measurements of the support for both same source and different sources propositions. However, the numerical values that are produced are difficult for examiners and novices alike to interpret, in part because they are not pure likelihood ratios, yet the full strength of the evidence depends on both numbers that are reported. This highlights the need to develop true likelihood ratios, either based on the physical features of the impressions [37] or based on subjective examiner responses. These values, or verbal equivalents, should be calibrated against the actual strength of the evidence in a particular discipline, which is not currently the case with conclusion scales in the

pattern comparison disciplines.

Finally, we feel that the best approach to communicating the strength of the observations is to explain not only the conclusion that was obtained, but also state what conclusions *could have been made but were not*. This should also include explicitly stating that both the same source and different source propositions were considered, as well as the relative support for both propositions where possible.

Table 5. Transcript of Video Instructions. This transcript was auto-captioned from the video with light editing for transcription errors. Consult the full video for imagery and intonation.

This study looks at communicating evidence in forensics. Before we get started, I'd like to say a few words about the task, the interface you'll use, and why we feel this is important. Fingerprint examiners compare fingerprints obtained from crime scenes similar to these, because the fingerprints are often degraded, the impressions are compared by humans, not by computers. Fingerprints are unique, but so is every impression made by a finger. The job of a fingerprint examiner is to look at the latent impression collected from a crime scene and compare it against an exemplar impression collected from a suspect or retrieved from a computer database.

The fingerprint examiner must decide whether there is enough evidence to conclude that the two impressions were made by the same finger or whether there's enough evidence to conclude that the two impressions were made by different fingers. The amount of evidence is accumulated in the mind of the examiner, supported by charts and notes. The examiner has to communicate the results of that comparison to a detective, judge, or jury.

An examiner accumulates evidence in support of two propositions or hypotheses. The first is same source, the two impressions came from the same finger, and the second is different sources, the two impressions came from different fingers. Note that we typically never know which of these two propositions are actually correct, but we can accumulate evidence in support of each.

This evidence scale describes a range of support that different conclusions might imply. The top of the scale is the same source proposition, which is the most support imaginable for the proposition that the two impressions came from the same finger. The bottom of the scale is a different sources proposition, which is the most support imaginable for the proposition that the two impressions came from different fingers. In the middle is equal evidence, which is the point on this scale where the evidence for the two propositions is equal.

Different comparisons might result in different levels of support for the two propositions. If the crime scene fingerprint is distorted or only a partial copy of the finger, there may not be much detail to work with when doing the comparison similar to these. Other impressions might be higher quality, and this might result in more evidence in support of one of the two propositions.

To communicate the results of the examination, the fingerprint examiner typically relies on a conclusion scale, which has various statements that communicate different levels of support for the two propositions. For example, the two fingerprints below are obviously different, suggesting more support for the different sources proposition than the same source proposition. The one on the left is a whorl. The one on the right is a left loop. In other cases, there might be a lot of detail in agreement between the two fingerprint impressions, suggesting more support for the same source proposition than the different sources proposition as shown with these images here.

Fingerprint examiners have various phrases to express the strength and support for the two propositions. It is important that the phrase they use is interpreted properly by others, such as detectives, judges, or jury members. The goal of this study is to allow you to express how you interpret the meaning of different phrases if spoken by a fingerprint examiner.

We're going to show you different phrases and ask you to place them on an evidence scale. Here we've added numbers where 100 represents the strongest evidence imaginable for the same source proposition. Minus 100 represents the strongest evidence imaginable for the different sources. Zero represents equal support for the same source and different sources propositions. We will use this scale to help express how much support you believe each conclusion statement implies about the two propositions. Note that the scale has stretched at the endpoints to help you make fine judgments about different statements that are close to each proposition.

To get started, imagine that you were on a jury and the fingerprint examiner has presented fingerprint evidence along with a specific phrase that expresses their conclusion. We're going to show you a series of phrases and asked you to tell us how you would interpret the level of support each phrase implies for the two propositions, each were spoken by a fingerprint examiner.

Let's go through the interface and I'll explain how it works. Once you've finished that video, you'll see a screen that looks like this. This is our sorting interface that allows you to read each one of our statements, as well as the definitions for each of those statements. And then to sort them in terms of the order for most evidence for same source at the top, two most evidence for different sources at the bottom. So I'll move same source up here and then different sources down here. And now they're in the correct order. And this is just for practice to learn this interface. And then you'll click the Check button. And if it's correct, you'll get to see this screen right here. Click the Start button, and then move the same source statement up to the top here. This is again just for practice to learn our interface. Me, move the IPO evidence to here, and then move the different sources all the way down here. So next you go on to the next scale. Your scale might look different than this one. But what we'd like you to do is to read each statement and then the definitions, and then sort the statements by most evidence for same source at the top to most evidence for different sources at the bottom. So I'll move this one up here. That seems to sort them there, and then click the Check button. And if they're correct, then you'll move on to the next screen. This is where the experiment actually begins.

So what I'd like you to do is to read this statement, review the definition if you need to, and then think about the location of this statement along the evidence axis from same source proposition, two different sources proposition. Move this statement to a location that corresponds to the strength of the evidence for same or different sources that you believe that statement implies if stated by a fingerprint examiner in court. So I won't bias you by telling you where I would place this. I would say that just move it to a location that satisfies that strength of the evidence that they feel like this implies a cup. And once you've placed that, click the Add next phrase button. Then you'll move this one to the correct location, the correct location that you infer from this statement, referring back to the definition, if you need to, a couple of things about using this scale. First of all, the different statements can overlap. That's certainly fine. The second thing is that you should preserve the order. So if you feel like one statement is slightly higher in terms of strength of the evidence, you should place it above another statement. And you can go back and move different statements if you need to, even though they're no longer red.

We would like you to treat this as a scale that goes from a 100, which is most evidence for same source that you could ever imagine, to minus 100, which is most evidence you could ever imagine for different source proposition. 50 is midway between equal evidence and same source and minus 50 is about midway between different sources and equal evidence. Use that scale as you like. When you're done with the phrases for a particular scale, it will go on automatically to the next scale. Once you've worked your way through all of the scales, there'll be a screen with some demographics and you can work your way through those, and then you'll be done with the experiment. We feel like this experiment is really important in terms of helping forensic examiners think about how to make a conclusion that is interpreted properly by a judge or jury, or a detective, and does so in a way that accurately represents the strength of the evidence. I appreciate you thinking carefully about the definitions of each statement and thinking about where would buy on the evidence

axis from evidence for the different sources proposition all the way up to evidence in favor of this same source proposition. Thank you so much for your help with this.

CRediT authorship contribution statement

Thomas Busey: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Morgan Klutzke:** Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful to Hal Stern for sharing the Thurstone–Mosteller model code used in the Thompson et al. (2018) paper. This work was supported by a grant from the National Institute of Justice, Award 2018-DU-BX-0212 to Thomas Busey and Indiana University.

References

[1] H.J. Swofford, J.G. Cino, Lay Understanding of "Identification": How Jurors Interpret Forensic Identification Testimony, J. Foren. Identif. 68 (2017) 29–41.
[2] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Accuracy and reliability of forensic latent fingerprint decisions, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 7733–7738.
[3] Friction Ridge Subcommittee, OSAC, Standard for Friction Ridge Examination Conclusions, in, 2018.
[4] IAI, IAI Resolution 2010-18, in, International Association for Identification, 2010.
[5] H. Eldridge, Juror comprehension of forensic expert testimony: a literature review and gap analysis, Foren. Sci. Int.: Synergy 1 (2019) 24–34.
[6] SWGFAST, Document #10 Standards for Examining Friction Ridge Impressions and Resulting Conclusions (Latent/Tenprint), in, Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST) www.swgfast.org, 2013.
[7] National Research Council of the National Academies of Science, Strengthening Forensic Science in the United States: A Path Forward, National Academies of Science, Washington DC, 2009.
[8] PCAST, Ensuring Scientific Validity of Feature-Comparison Methods, in, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf 2016.
[9] J.M. Butler, J.M. Butler, Fundamentals of forensic DNA typing, Academic Press/Elsevier, Amsterdam; Boston, 2010.
[10] G.S. Morrison, The likelihood-ratio framework and forensic evidence in court: a response to R v T, Int. J. Evidence Proof 16 (2012) 1–29.
[11] C.E.H. Berger, J. Buckleton, C. Champod, I.W. Evett, G. Jackson, Re: Expressing evaluative opinions; A position statement Response, Sci. Justice 51 (2011).
[12] C. Aitken, A. Barrett, C. Berger, A. Biedermann, C. Champod, T. Hicks, J. Lucena-Molina, L. Lunt, S. McDermott, L. McKenna, ENFSI guideline for evaluative reporting in forensic science, (2015).
[13] W.C. Thompson, R.H. Grady, E. Lai, H.S. Stern, Perceived strength of forensic scientists' reporting statements about source conclusions, Law Probability & Risk 17 (2018) 133–155.
[14] I.W. Evett, Towards a uniform framework for reporting opinions in forensic science casework, Sci. Justice 38 (1998) 198–202.
[15] K.A. Martire, R.I. Kemp, B.R. Newell, The psychology of interpreting expert evaluative opinions, Aust. J. Forensic Sci. 45 (2013) 305–314.
[16] A. Nordgaard, R. Ansell, W. Drotz, L. Jaeger, Scale of conclusions for the value of evidence, Law, Probability and Risk 11 (2012) 1–24.
[17] W.C. Thompson, E.J. Newman, Lay understanding of forensic statistics: Evaluation of random match probabilities, likelihood ratios, and verbal equivalents, Law Hum. Behav. 39 (2015) 332.
[18] B. Garrett, W. Crozier, R. Grady, Error rates, likelihood ratios, and jury evaluation of forensic evidence, J. Forensic Sci. 65 (2020) 1199–1209.
[19] L.M. Howes, K.P. Kirkbride, S.F. Kelty, R. Julian, N. Kemp, Forensic scientists' conclusions: How readable are they for non-scientist report-users? Forensic Sci. Int. 231 (2013) 102–112.
[20] B.A. Spellman, Communicating forensic evidence: lessons from psychological science, Seton Hall L. Rev. 48 (2017) 827.
[21] D. McQuiston-Surrett, M.J. Saks, The testimony of forensic identification science: What expert witnesses say and what factfinders hear, Law Hum. Behav. 33 (2009) 436–453.
[22] W.C. Thompson, S.O. Kaasa, T. Peterson, Do jurors give appropriate weight to forensic identification evidence? J. Empir. Legal Stud. 10 (2013) 359–397.
[23] B. Garrett, G. Mitchell, How jurors evaluate fingerprint evidence: The relative importance of match language, method information, and error acknowledgment, J. Empir. Legal. Stud. 10 (2013) 484–511.
[24] J.J. Koehler, N. Schweitzer, M.J. Saks, D.E. McQuiston, Science, technology, or the expert witness: What influences jurors' judgments about forensic science testimony? Psychology, Public Policy, and Law 22 (2016) 401.
[25] B. Garrett, G. Mitchell, N. Scurich, Comparing categorical and probabilistic fingerprint evidence, J. Forensic Sci. 63 (2018) 1712–1717.
[26] W.C. Thompson, R.H. Grady, E. Lai, H.S. Stern, Perceived strength of forensic scientists' reporting statements about source conclusions, Law, Probability and Risk 17 (2018) 133–155.
[27] K.E. Carter, M.D. Vogelsang, J. Vanderkolk, T. Busey, The Utility of Expanded Conclusion Scales During Latent Print Examinations, J. Forensic Sci. (2020).
[28] T. Busey, M. Klutzke, A. Nuzzi, J. Vanderkolk, Validating strength-of-support conclusion scales for fingerprint, footwear, and toolmark impressions, J. Forensic Sci. 67 (2022) 936–954.
[29] D.J. Cohen, J.M. Ferrell, N. Johnson, What very small numbers mean, J. Exp. Psychol.-General 131 (2002) 424–442.
[30] K.A. Martire, R.I. Kemp, M. Sayle, B.R. Newell, On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect, Forensic Sci Int 240 (2014) 61–68.
[31] C.L. Swanson, USACIL DFSC conclusion scale, in: T. Busey (Ed.), Email correspondence, 2020.
[32] H.J. Swofford, A.J. Koertner, F. Zemp, M. Ausdemore, A. Liu, M.J. Salyards, A method for the statistical interpretation of friction ridge skin impression evidence: Method development and validation, Forensic Sci. Int. 287 (2018) 113–126.
[33] R. Marquis, A. Biedermann, L. Cadola, C. Champod, L. Gueissaz, G. Massonnet, W. D. Mazzella, F. Taroni, T. Hicks, Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings, Sci. Justice 56 (2016) 364–370.
[34] C.R. Fox, J.R. Irwin, The role of context in the communication of uncertain beliefs, Basic Appl. Soc. Psychol. 20 (1998) 57–70.
[35] C. Fox, B. Malle, On the communication of uncertainty: Two modes of linguistic expression, Unpublished manuscript, (1997).
[36] E. Løhre, K.H. Teigen, There is a 60% probability, but I am 70% certain: Communicative consequences of external and internal expressions of uncertainty, Thinking Reasoning 22 (2016) 369–396.
[37] C. Neumann, C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz, D. Meuwly, A. Bromage-Griffiths, Computation of likelihood ratios in fingerprint identification for configurations of three minutiae, J. Forensic Sci. 51 (2006) 1255–1266.