| | |
|---|---|
| Document Title: | Validating Conclusion Scales in the Forensic Sciences |
| Author(s): | Thomas Busey |
| Document Number: | 304758 |
| Date Received: | May 2022 |
| Award Number: | 2018-DU-BX-0212 |

# Validating Conclusion Scales in the Forensic Sciences

Final Report

NIJ grant #2018-DU-BX-0212

Thomas Busey, PI    Indiana University, Bloomington, IN

# Executive Summary

This project combined two sets of studies to validate the conclusion scales in the fingerprint, footwear, and tool mark disciplines. The first study measured how fingerprint examiners and members of the general public interpreted different articulation statements. The second set of studies measured how fingerprint, footwear, and tool mark examiners would use articulation statements expressed in strength-of-evidence language rather than as source attribution statements. By combining across the two sets of studies, we demonstrate how statements are both used in casework-like comparisons as well as how the articulation language is interpreted by the consumers of forensic evidence. The two sets of studies have been submitted for publication: Busey and Klutzke (submitted), summarized as Section 1 in the current report; and Busey, Klutzke, Nuzzi, and Vanderkolk (submitted), summarized as Section 2 in the current report.

Pattern comparison disciplines use categorical statements to express conclusions. In the first study, we measured the strength of evidence for six different scales in members of the general public and fingerprint examiners. The statements came from different types of scales, 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 whatever statement was the highest in its 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.

The second set of studies compared traditional conclusion statements against statements phrased as strength of evidence for different propositions. In the pattern comparison disciplines such as fingerprints, footwear, and tool marks, the results of a comparison are communicated by examiners in the form of categorical conclusions such as Identification or Exclusion. These statements have been criticized as being overinterpreted by laypersons, and so alternative statements based on strength of evidence language have been proposed as replacements. The current study compares traditional conclusion scales against strength of evidence scales to determine how these new statements might be used by examiners in casework. Each participant completed 60 comparisons within their discipline that were designed to approximate casework conditions, using either a traditional or a strength of evidence conclusion scale. The scale used on each trial was randomly assigned, and participants knew the scale for that trial as they began the comparison. We found that fingerprint examiners redefined the term Identification when the scale was expanded to include Support for Common Source, using Identification less often than when the traditional scale was assigned. Fingerprint examiners were also much less likely to use Extremely Strong Support for Common Source than Identification. Footwear examiners treated the traditional and strength of evidence scales similarly, but toolmark examiners were much less likely to use Extremely Strong Support for Common Source than Identification, similar to fingerprint examiners. The results demonstrate that examiners reserve Extremely Strong Support for Common Source for only the comparisons with the most evidence for the common source proposition.

# Section 1: Calibrating the Perceived Strength of Evidence of Forensic Testimony Statements

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 (Swofford & Cino, 2017) and perhaps too strong given the error rates observed in error rate (black box) studies (Ulery, Hicklin, Buscaglia, & Roberts, 2011).

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 (Friction Ridge Subcommittee & OSAC, 2018). 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 over-interpreting the strength of the forensic evidence. While an examiner may qualify some conclusions on the stand during testimony (IAI, 2010), the vast majority of cases do not go to trial. Instead, these qualifications or hedges may be ignored or misunderstood by a 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 (Eldridge, 2019; SWGFAST, 2013b). Various organizations have criticized categorical conclusions as either prone to overinterpretation or implying absolute certainty (National Research Council of the National Academies of Science, 2009; PCAST, 2016; Swofford & Cino, 2017).

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. These are widely used in forensic DNA applications where probabilistic genotyping software provides a numerical result (Butler & Butler, 2010), but Morrison (2012) has argued that the likelihood ratio could be based on the expert's subjective evaluation. This approach is widely used in Europe (Berger, Buckleton, Champod, Evett, & Jackson, 2011) 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. In the former case, these would be considered a posterior, because it can is a statement about a proposition. In the latter case, this is 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 (and we see evidence for this in our data as well). They 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 (Evett, 1998).

Work on juror understanding of evidence has focused on whether categorical scales or numerical likelihood ratios are better understood by members of the jury (Martire, Kemp, & Newell, 2013) and calls for a unified scale across disciplines (Nordgaard, Ansell, Drotz, &

Jaeger, 2012). Thompson and Newman (2015) 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 (2020). The choice of wording will also matter; Howes, Kirkbride, Kelty, Julian, and Kemp (2013) found that reports from forensic glass analysis would be difficult for a lay audience to comprehend. Martire et al. (2013) 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 (2017) 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 (2009) found that qualitative statements were more damaging to the defense than quantitative statements. However, Thompson, Kaasa, and Peterson (2013) 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 (Garrett & Mitchell, 2013). The perceived reputation of the examiner and the sophistication of the methods may actually play a greater role than the testimony itself (Koehler, Schweitzer, Saks, & McQuiston, 2016).

Within the fingerprint discipline, Garrett, Mitchell, and Scurich (2018) 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.

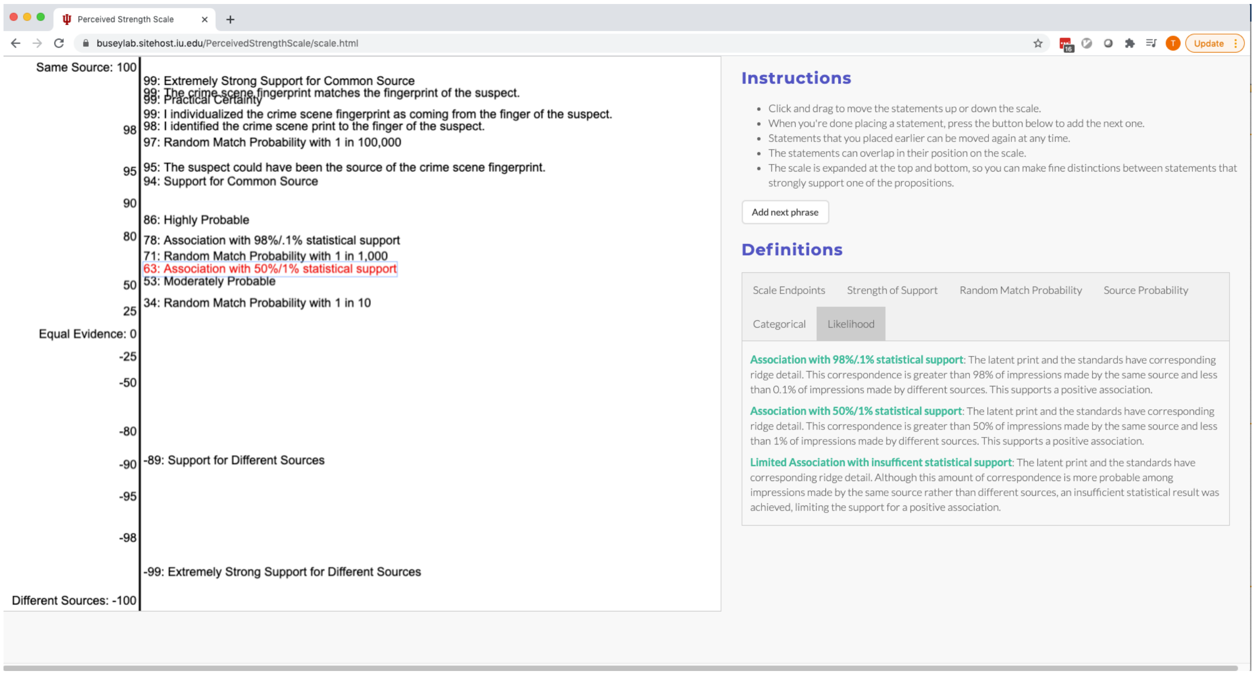
Thompson, Grady, Lai, and Stern (2018) addressed this question with a very straightforward design. They presented pairs of statements to members of the public (Amazon Mechanical Turk workers) and asked the participants: "Which of the following two conclusions would seem STRONGER if you heard it, meaning more convincing to you that the suspect is the source of the print?" (Thompson et al., 2018, p. 139). This is a time-consuming process because all possible pairs must be compared, but 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 (Carter, Vogelsang, Vanderkolk, & Busey, 2020). 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. 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 suffer from this overinterpretation.

To compliment the Thompson et al. (2018) 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 Figure 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 (2002). 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 et al. (2013) also took advantage of both numerical and visual displays to provide accurate estimation of proportions by their participants.

We designed the visual interface shown in Figure 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). A similar scale was used by Martire, Kemp, Sayle, and Newell (2014), with the exception that their scale ranged from -10,000 to 10,000. The interface in Figure 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.



99: Extremely Strong Support for Common Source
99: The crime scene fingerprint matches the fingerprint of the suspect.
99: Practical Certainty
99: I individualized the crime scene fingerprint as coming from the finger of the suspect.
98: I identified the crime scene print to the finger of the suspect.
97: Random Match Probability with 1 in 100,000
95: The suspect could have been the source of the crime scene fingerprint.
94: Support for Common Source
86: Highly Probable
78: Association with 98%/.1% statistical support
71: Random Match Probability with 1 in 1,000
63: Association with 50%/1% statistical support
53: Moderately Probable
34: Random Match Probability with 1 in 10
-89: Support for Different Sources
-99: Extremely Strong Support for Different Sources

Instructions

- Click and drag to move the statements up or down the scale.
- When you're done placing a statement, press the button below to add the next one.
- Statements that you placed earlier can be moved again at any time.
- The statements can overlap in their position on the scale.
- The scale is expanded at the top and bottom, so you can make fine distinctions between statements that strongly support one of the propositions.

Definitions

**Association with 98%/.1% statistical support**: The latent print and the standards have corresponding ridge detail. This correspondence is greater than 98% of impressions made by the same source and less than 0.1% of impressions made by different sources. This supports a positive association.

**Association with 50%/1% statistical support**: The latent print and the standards have corresponding ridge detail. This correspondence is greater than 50% of impressions made by the same source and less than 1% of impressions made by different sources. This supports a positive association.

**Limited Association with insufficent statistical support**: The latent print and the standards have corresponding ridge detail. Although this amount of correspondence is more probable among impressions made by the same source rather than different sources, an insufficient statistical result was achieved, limiting the support for a positive association.

Figure 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).

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) (Swanson, 2020). Note that this language has recently changed from the original formulation (DFSC, 2017) 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.

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 Figure 11. Further details of the methods are found below.

| 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 100000 | RMP 1 in 100000 |
| | 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%/.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 |

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.

## 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.

### *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 jury-eligible 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 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 et al. (2018). 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 about 5000, and location in the United States. These participants were compensated $2 for their participation.

Table 3 in the Supplementary Information has details on age distributions for the three groups, and Table 4 in the Supplementary Information has details on the education distribution for the three groups.

### *Instructions*

To address the information gap between fingerprint examiners and members of the general public, we produced an 8 minute 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 in the Supplementary Information. We also included a sorting tasks that demonstrated to participants the nature of each scale as described below.

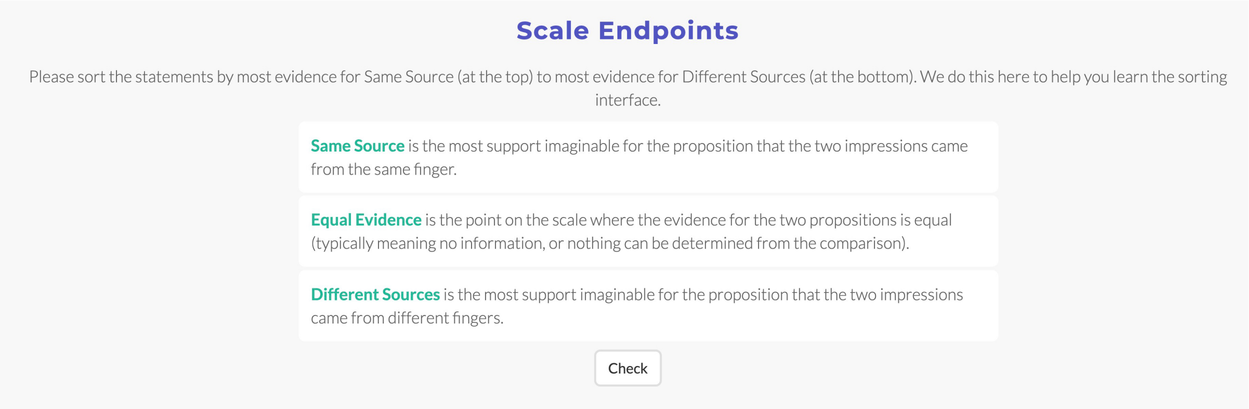
### *Procedure*

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 (Cohen et al., 2002). 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 Figure 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.



Scale Endpoints

Please sort the statements by most evidence for Same Source (at the top) to most evidence for Different Sources (at the bottom). We do this here to help you learn the sorting interface.

**Same Source** is the most support imaginable for the proposition that the two impressions came from the same finger.

**Equal Evidence** is the point on the scale where the evidence for the two propositions is equal (typically meaning no information, or nothing can be determined from the comparison).

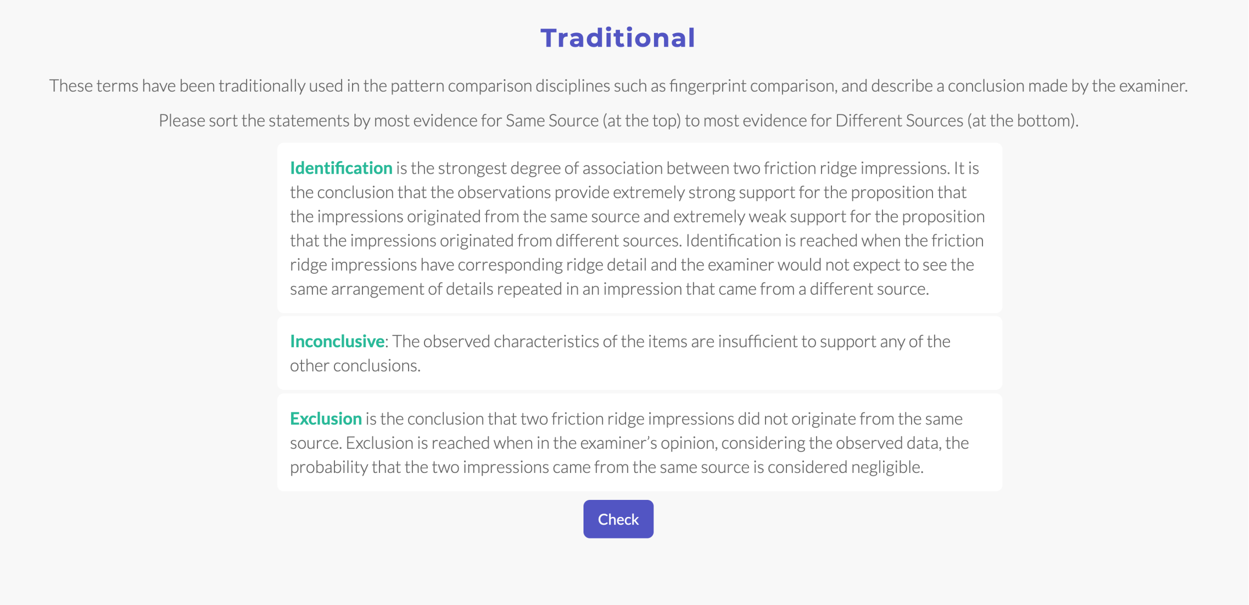
**Different Sources** is the most support imaginable for the proposition that the two impressions came from different fingers.

Figure 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.

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 Figure 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. 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 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 Figure 3. The order of the 6 scales was randomized across participants, so that the traditional scale shown in Figure 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 two exceptions were the Categorical scale, where the ordering between the "I individualized" and "I identified" statements is unclear and we did not want to bias participants by enforcing a particular order, and the Source Probability scale where we judged that Practical Certainty and Highly Probable were ambiguous enough not to enforce a sort order between these two items. Figure 11 in the Supplementary Information shows all scales in correct sort order.



Traditional

These terms have been traditionally used in the pattern comparison disciplines such as fingerprint comparison, and describe a conclusion made by the examiner.

Please sort the statements by most evidence for Same Source (at the top) to most evidence for Different Sources (at the bottom).

**Identification** is the strongest degree of association between two friction ridge impressions. It is the conclusion that the observations provide extremely strong support for the proposition that the impressions originated from the same source and extremely weak support for the proposition that the impressions originated from different sources. Identification is reached when the friction ridge impressions have corresponding ridge detail and the examiner would not expect to see the same arrangement of details repeated in an impression that came from a different source.

**Inconclusive**: The observed characteristics of the items are insufficient to support any of the other conclusions.

**Exclusion** is the conclusion that two friction ridge impressions did not originate from the same source. Exclusion is reached when in the examiner's opinion, considering the observed data, the probability that the two impressions came from the same source is considered negligible.

Figure 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.

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.

### *Participant Exclusion*

This experiment requires careful thought and logical thinking to appreciative 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 seconds, 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 2 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 2 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 2 but were not used to exclude participants. Numbers in paratheses 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 2 because these were repeat subjects).

| | 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 2. Violation counts for the three types of participants, with unique number of participants in paratheses. 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.

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 general public, 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.

## Results

We will present data aggregated across the two novice groups for comparison with the Fingerprint Examiners, and also provide separate comparisons between the two novice 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, which also contains the data and analysis files for the companion paper:

https://osf.io/xmwqg/?view_only=f1b996eee77d45d0907ecebdaa27437d

### *Raw Values*

Our first analysis presents the distribution of responses for each conclusion statement. Figure 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…” 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.

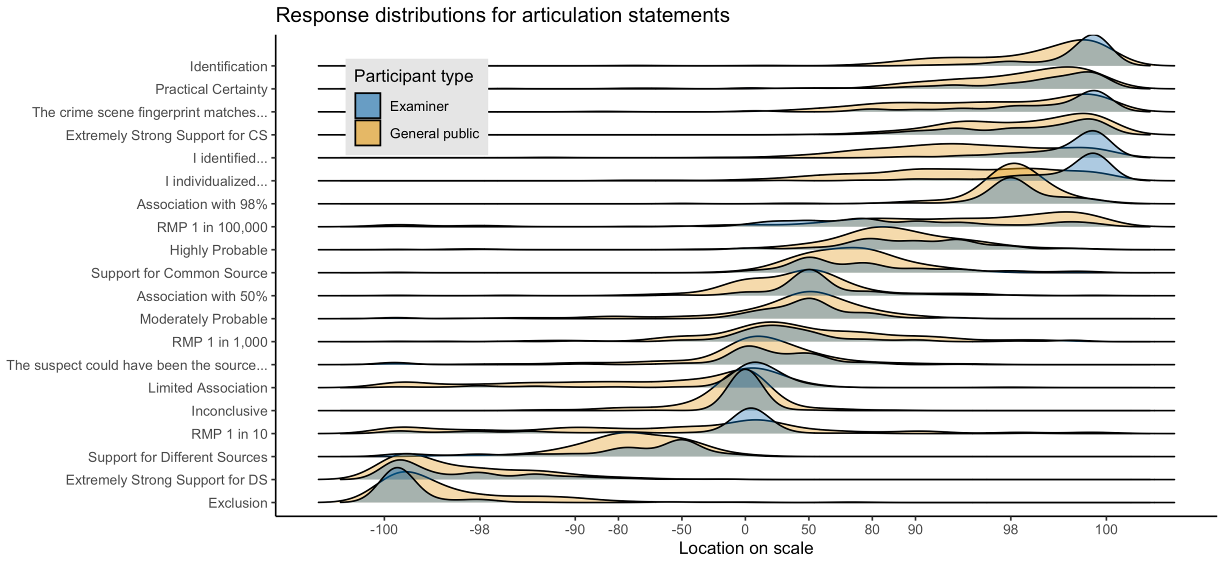


Figure 4. Ridge plot comparing Examiners to members of the general public. The different conclusion statements are summarized on the left, and the distribution of responses is illustrated with the colored ridge plots. Note that the evidence axis scale is expanded to mimic the scale used by participants (see Figure 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.

The two groups that constitute the members of the general public performed remarkedly consistently, as illustrated in Figure 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.

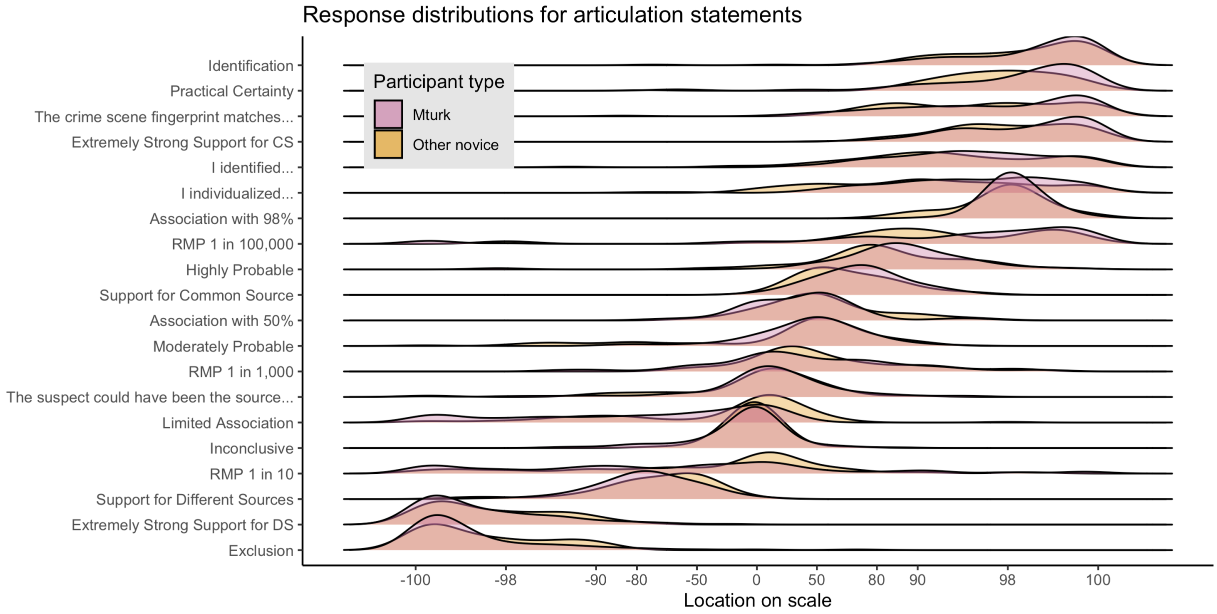


Figure 5. Ridge plot distributions for the two groups that constitute the members of the general public.

The variance (standard deviation) of the placement of each statement across participants is a measure of the (in)consistency across participants. Figure 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.

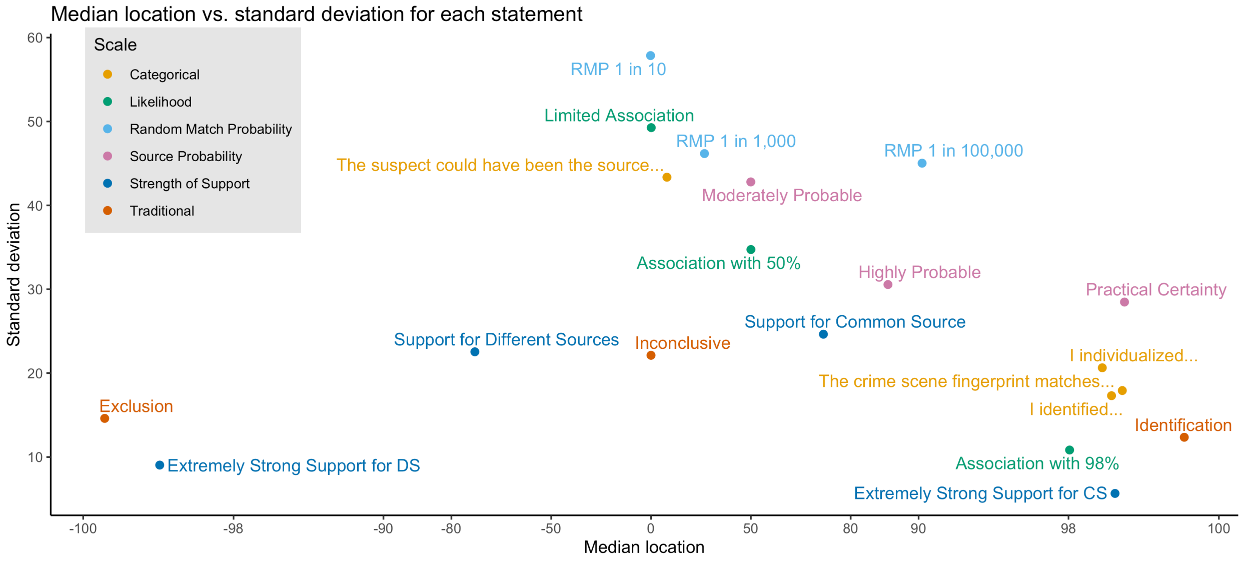


Figure 6. Scatterplot comparing the median for each conclusion statement against the associated standard deviation for that term, combined across all participants. Values higher in the graph are associated with greater variability. Some terms toward the ends of the scale have low variability and therefore 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.

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, 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 (Busey et al., submitted), 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 in the companion paper.

### *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 et al. (2018) when they directly compared pairs of individual statements. The authors were kind 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 correspond to the overall strength of evidence inferred from the dominance matrix for that statement (see Thompson et al. (2018) 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 Figure 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. Figure 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 Figure 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 Common Source. It may be that a numerical estimate tends to reduce the amount of support for common source implied by the statement.

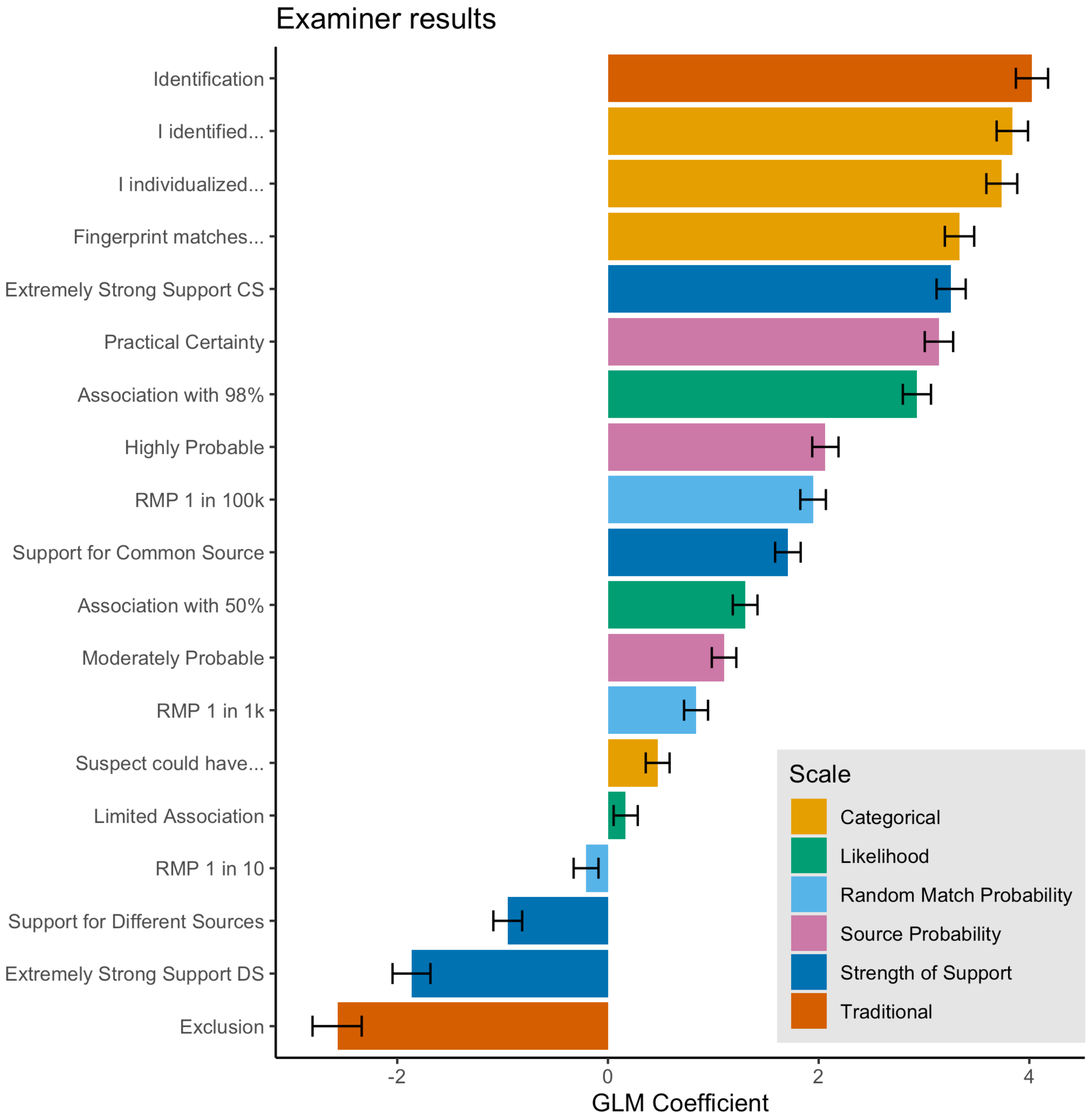


Figure 7. Generalized Linear Model (GLM) coefficients for each statement for fingerprint examiners. Error bars represent 95% confidence intervals around the point estimate for each coefficient.

Figure 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.

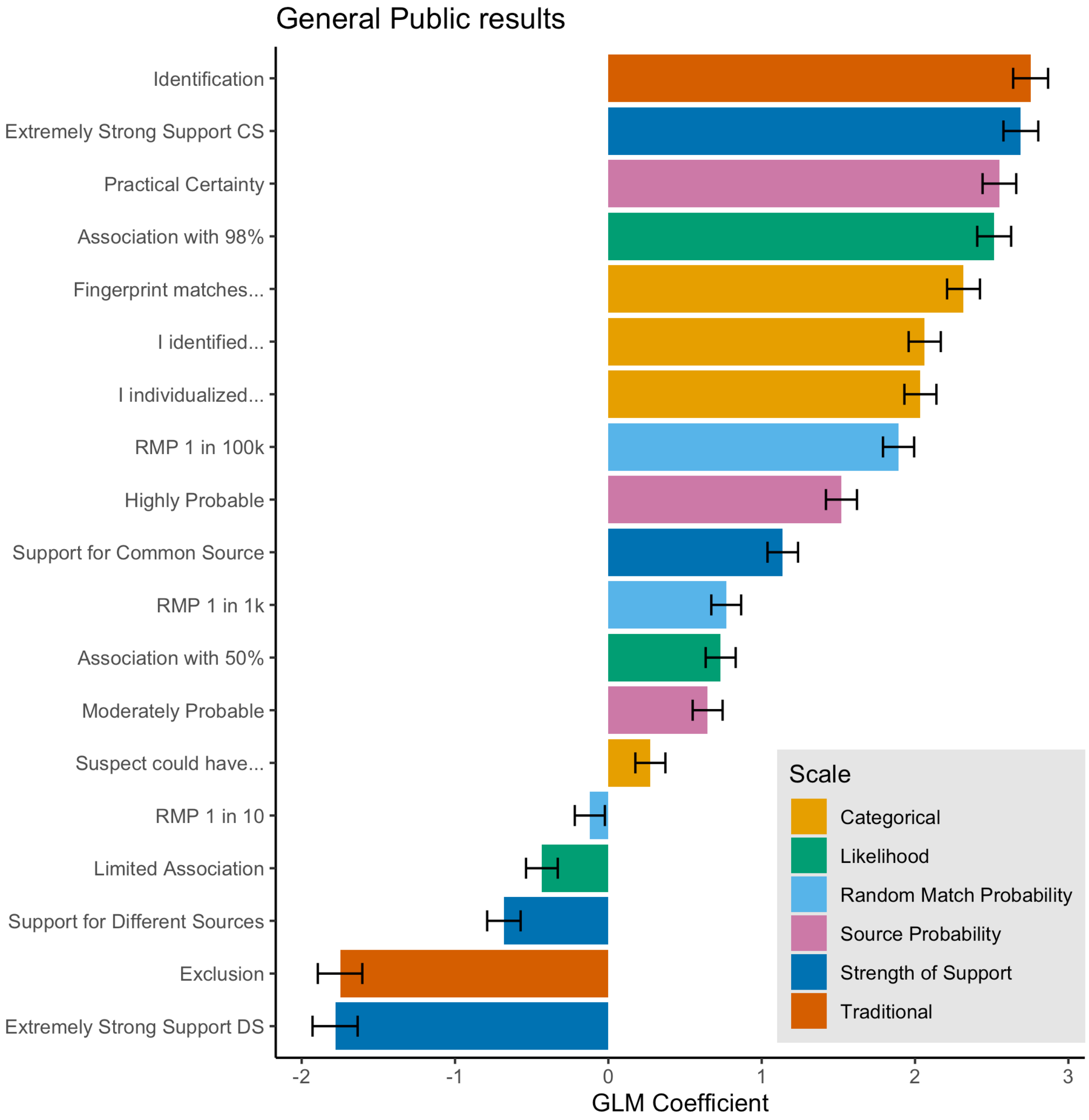


Figure 8. Generalized Linear Model (GLM) coefficients for each statement for members of the general public. Error bars represent 95% confidence intervals around the point estimate for each coefficient.

For direct comparison between the two participate types, Figure 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 100k 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 Pubic 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.

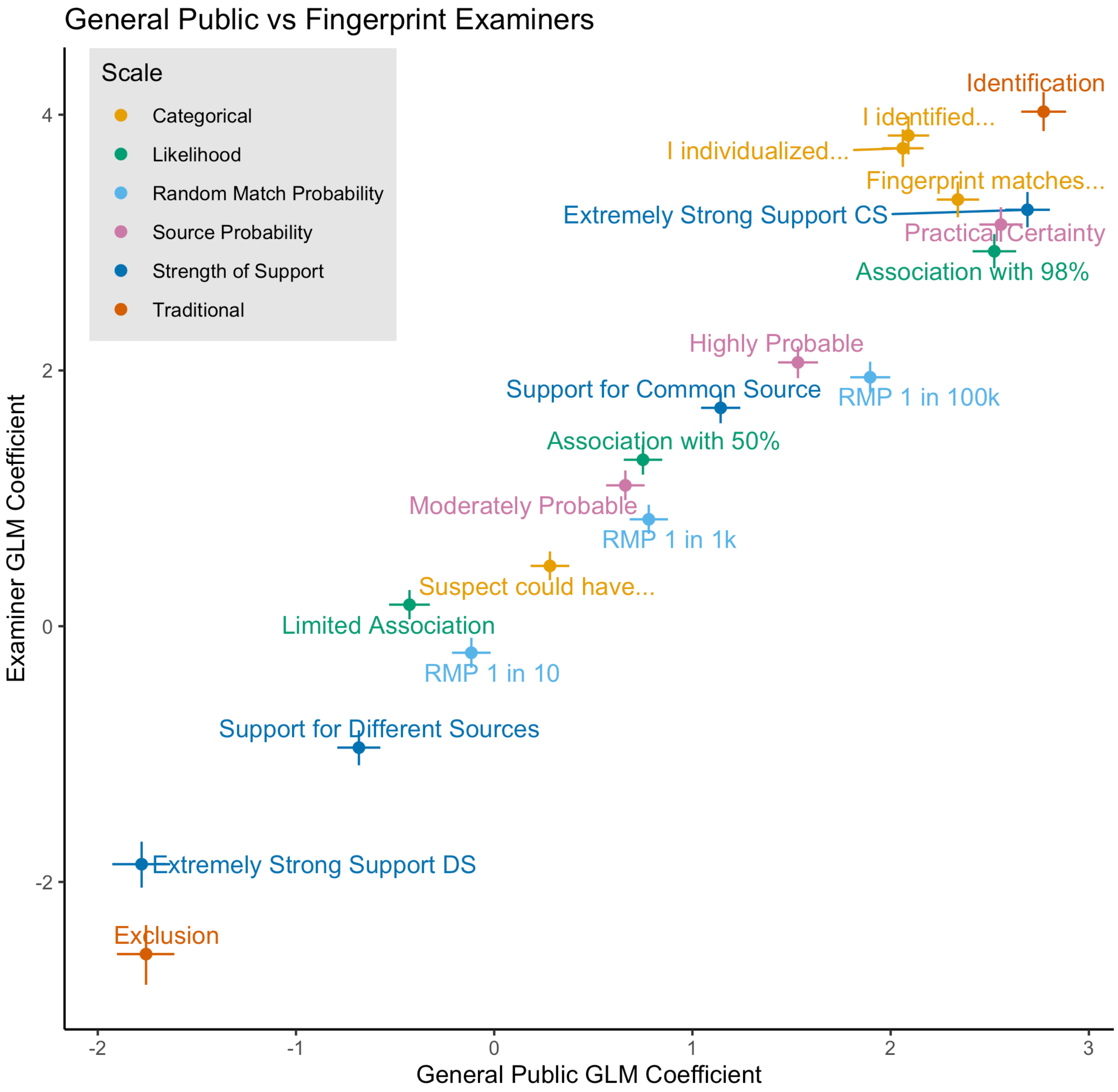


Figure 9. Scatterplot comparing the coefficients of members of the general public (abscissa) against the examiners (ordinate). Error bars represent 95% confidence intervals around each coefficient estimates.

Figure 10 compares the two types of novices. 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.

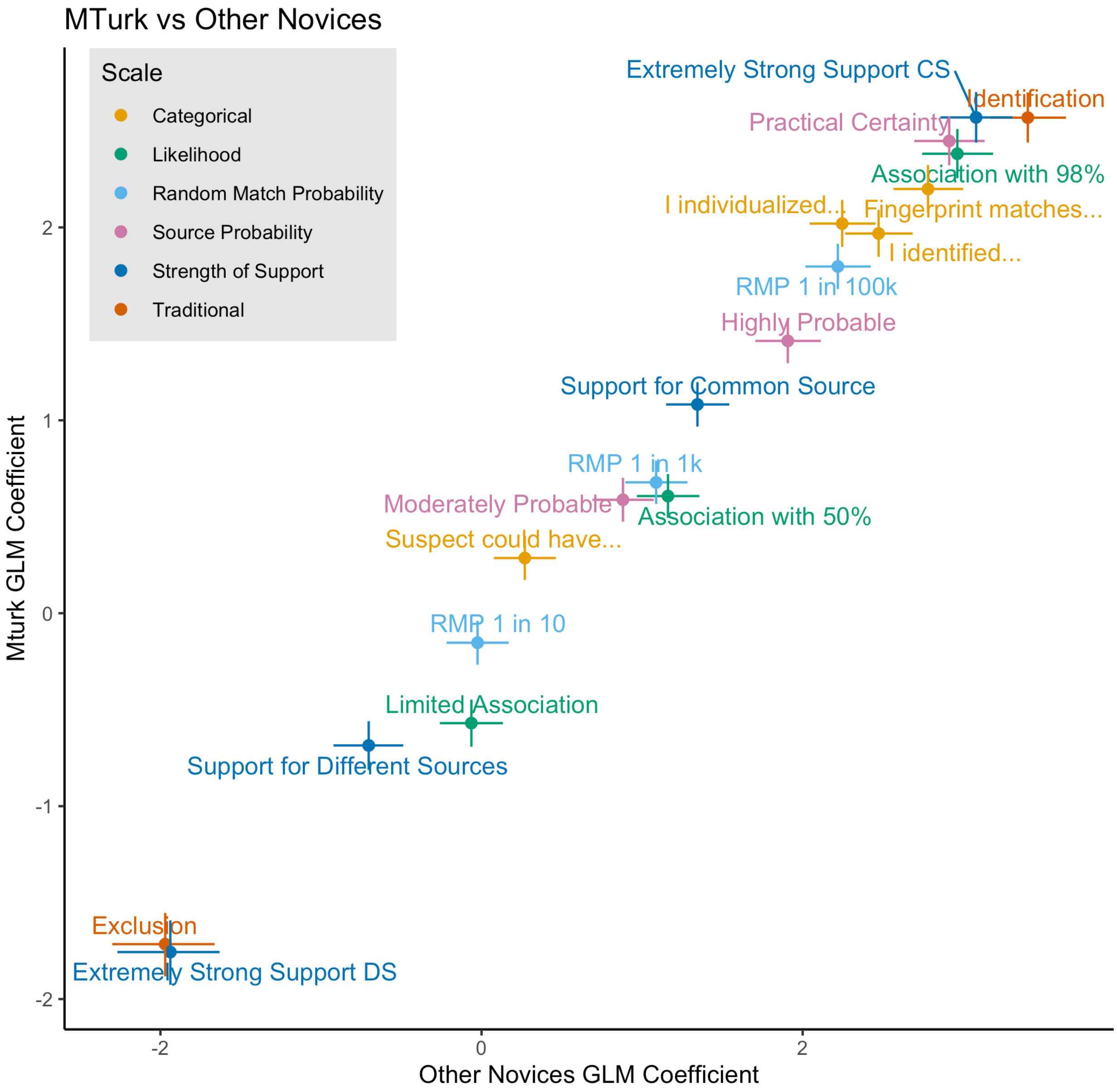


Figure 10. Scatterplot comparing the coefficients of the two types of members of the general public. Error bars represent 95% confidence intervals around each coefficient estimates.

## Discussion

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 Figure 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' much lower than 'Identification', demonstrating that they view it as implying less evidence overall when it is used. This potentially represents an overinterpretation of the 'Extremely Strong Support for Common Source' articulation statement if the goal is to move away from absolutist language. This conclusion is supported by the K-S tests discussed above. However, as discussed in the companion paper (Busey et al., submitted), examiners tend to use Extremely Strong Support for Common Source less often than Identification in casework-like comparisons, and so the chance for overinterpretation in casework may be minimized.

2) Both subject groups 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", despite the fact that error rate studies that use 'Identification' as a conclusion demonstrate a false identification rate of 1 in 1000 (Ulery et al., 2011). However, it is notable that "Association with 98%" and "Random Match 1 in 100,000" both contain numerical estimates, which may reduce the implied support for common source despite the fact that the numbers offer more specificity.

3) The DFSC language is seen as implying less evidence than '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 (Marquis et al., 2016). 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 obtained by Garrett et al. (2018), 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 et al. (2018) use the original DFSC statements DFSC (2017) whereas we used updated statements (Swanson, 2020), or possibly due to the differences in methods.

4) The results from the Categorical Scale from the general public (see Figure 8) are perhaps a bit surprising. Both novice groups placed “I Identified” and “I Individualized” below “Identification”, despite the fact that the wording is almost identical (see the yellow terms in Figure 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 Figure 7). One difference between Identification and I Identified/I individualized comes from the sorting task, where we did not enforce a sort for the Categorical terms because it is not a true conclusion scale. These terms were added as a consistency check, and the fact that members of the general public were not consistent with their placement of these terms suggests that the sorting task played a role in their interpretation of each term. In the cases of “I identified…” and “I individualized…” the sorting task did not indicate which term provided the most support for common source (recall that a sort order was not enforced for this scale), and therefore members of the general public may have been unsure of how to interpret these phrases. The terms in the Categorical scale also did not have definitions associated with them, and tended to focus on the examiner. This focus may have introduced some doubt or ambiguity on the part of novice participants, causing them to lower their estimate of the strength of the evidence for these terms.

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 offered by the Defense Forensic Science Center is perhaps the best conclusion scale, because it appears less likely to be overinterpreted by members of the general public due to the inclusion of numerical values, and the use of the term "Association" as opposed to "Identification". However, both fingerprint examiners and members of the general public were somewhat naïve in their interpretation of the statement, tending to place the statement at a value of 98 (see the Association with 98%/.1% statement in Figure 4) and the 50%/1% statement at a value of 50. We describe this as numerically naïve because 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 .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 .5%). We considered, but rejected, adding an additional phrase that included a 98%/.5% comparison to see how participants would treat this new statement, which logically would be placed below the 98%/.1% statement, but decided that this would be too confusing to participants.

The Defense Forensic Science Center articulation language has the additional advantage that it explicitly considers competing hypotheses because it provides separated estimates of the support for both same source and different sources propositions. We suggest exploring the possibility of applying this approach more broadly to the pattern comparison disciplines, even if quantification can be difficult to achieve in a particular domain.

Ultimately 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 may also help guard against overinterpretation by detectives of investigative leads such as Support for Common Source when the scale also includes Strong Support or Extremely Strong Support for Common Source.

## Acknowledgements

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

## Supplementary Information

Table 3. Age Demographics for the three groups.

| Group | 18-24 | 25-34 | 35-44 | 45-54 | 55-64 | 65-74 | 75+ | Decline |
|---|---|---|---|---|---|---|---|---|
| General Public | 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 4. Education Demographics for the three groups. Highest degree obtained.

| Group | Decline | Bachelor's | College Student | High School | Masters | PhD | Professional | Some College |
|---|---|---|---|---|---|---|---|---|
| General Public | 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 |

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.

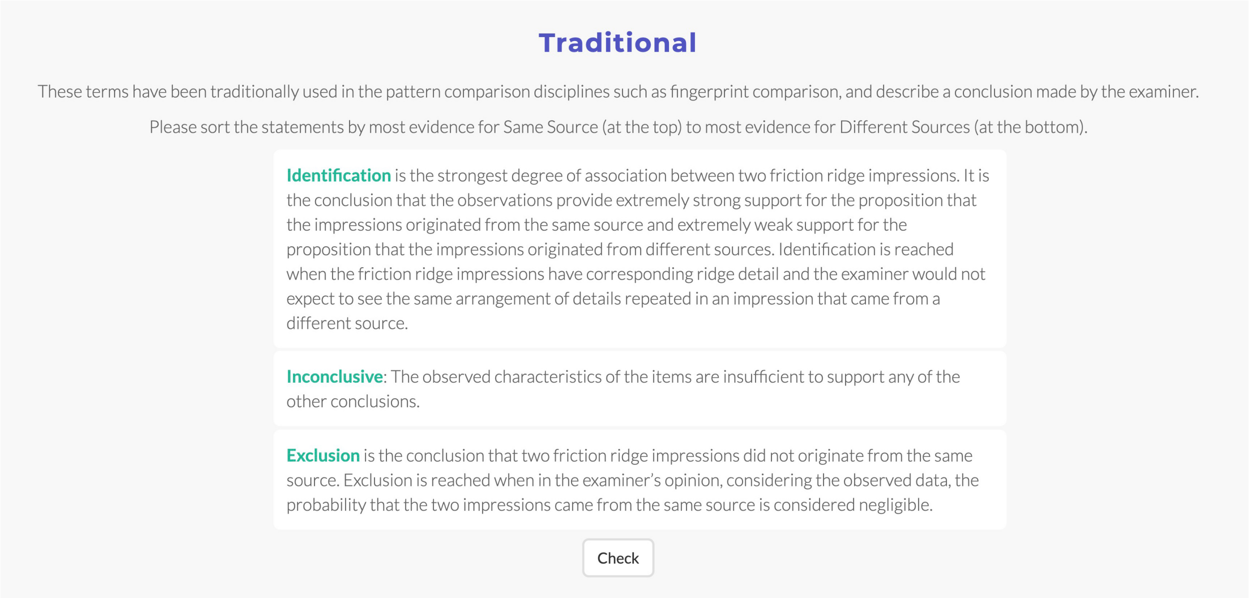
We're going to show you different phrases and ask you to place them on an evidence scale. Here we've added numbers where 100 represents the strongest evidence imaginable for the same source proposition. Minus 100 represents the strongest evidence imaginable for the different sources. Zero represents 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.



### Traditional

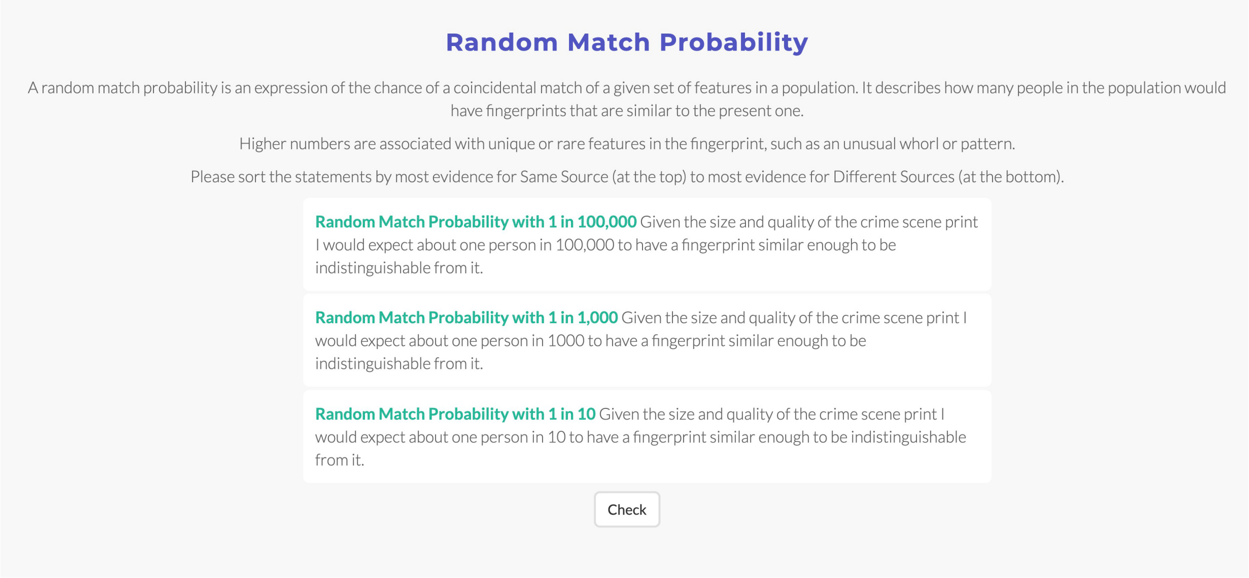
These terms have been traditionally used in the pattern comparison disciplines such as fingerprint comparison, and describe a conclusion made by the examiner.

Please sort the statements by most evidence for Same Source (at the top) to most evidence for Different Sources (at the bottom).

**Identification** is the strongest degree of association between two friction ridge impressions. It is the conclusion that the observations provide extremely strong support for the proposition that the impressions originated from the same source and extremely weak support for the proposition that the impressions originated from different sources. Identification is reached when the friction ridge impressions have corresponding ridge detail and the examiner would not expect to see the same arrangement of details repeated in an impression that came from a different source.

**Inconclusive**: The observed characteristics of the items are insufficient to support any of the other conclusions.

**Exclusion** is the conclusion that two friction ridge impressions did not originate from the same source. Exclusion is reached when in the examiner's opinion, considering the observed data, the probability that the two impressions came from the same source is considered negligible.



### Random Match Probability

A random match probability is an expression of the chance of a coincidental match of a given set of features in a population. It describes how many people in the population would have fingerprints that are similar to the present one.

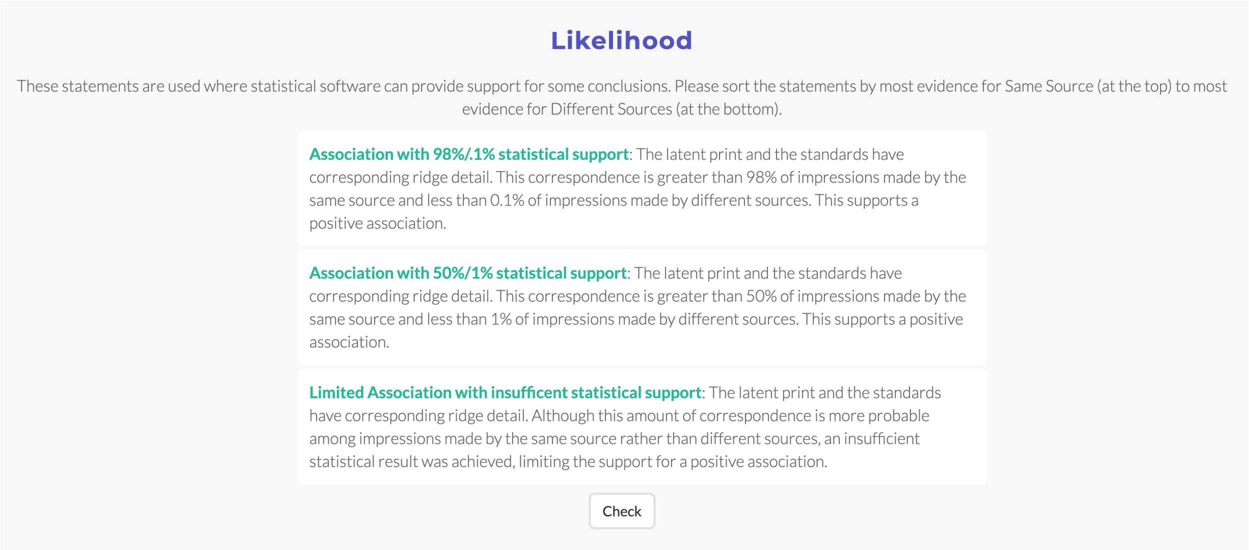
Higher numbers are associated with unique or rare features in the fingerprint, such as an unusual whorl or pattern.

Please sort the statements by most evidence for Same Source (at the top) to most evidence for Different Sources (at the bottom).

**Random Match Probability with 1 in 100,000** Given the size and quality of the crime scene print I would expect about one person in 100,000 to have a fingerprint similar enough to be indistinguishable from it.

**Random Match Probability with 1 in 1,000** Given the size and quality of the crime scene print I would expect about one person in 1000 to have a fingerprint similar enough to be indistinguishable from it.

**Random Match Probability with 1 in 10** Given the size and quality of the crime scene print I would expect about one person in 10 to have a fingerprint similar enough to be indistinguishable from it.



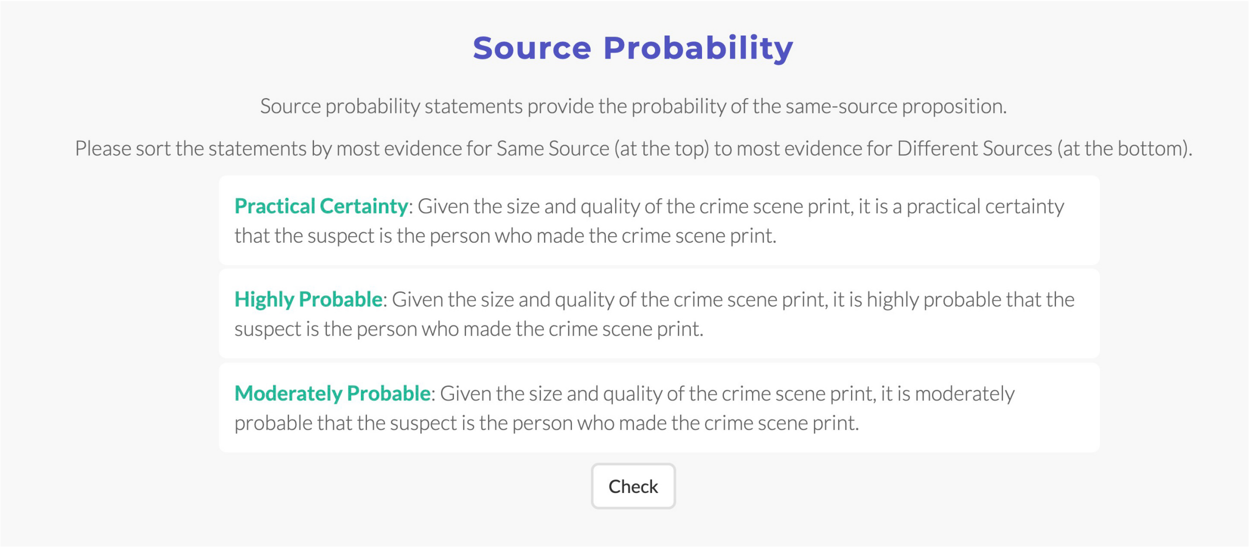
## Likelihood

These statements are used where statistical software can provide support for some conclusions. Please sort the statements by most evidence for Same Source (at the top) to most evidence for Different Sources (at the bottom).

**Association with 98%/.1% statistical support**: The latent print and the standards have corresponding ridge detail. This correspondence is greater than 98% of impressions made by the same source and less than 0.1% of impressions made by different sources. This supports a positive association.

**Association with 50%/1% statistical support**: The latent print and the standards have corresponding ridge detail. This correspondence is greater than 50% of impressions made by the same source and less than 1% of impressions made by different sources. This supports a positive association.

**Limited Association with insufficent statistical support**: The latent print and the standards have corresponding ridge detail. Although this amount of correspondence is more probable among impressions made by the same source rather than different sources, an insufficient statistical result was achieved, limiting the support for a positive association.



## Source Probability

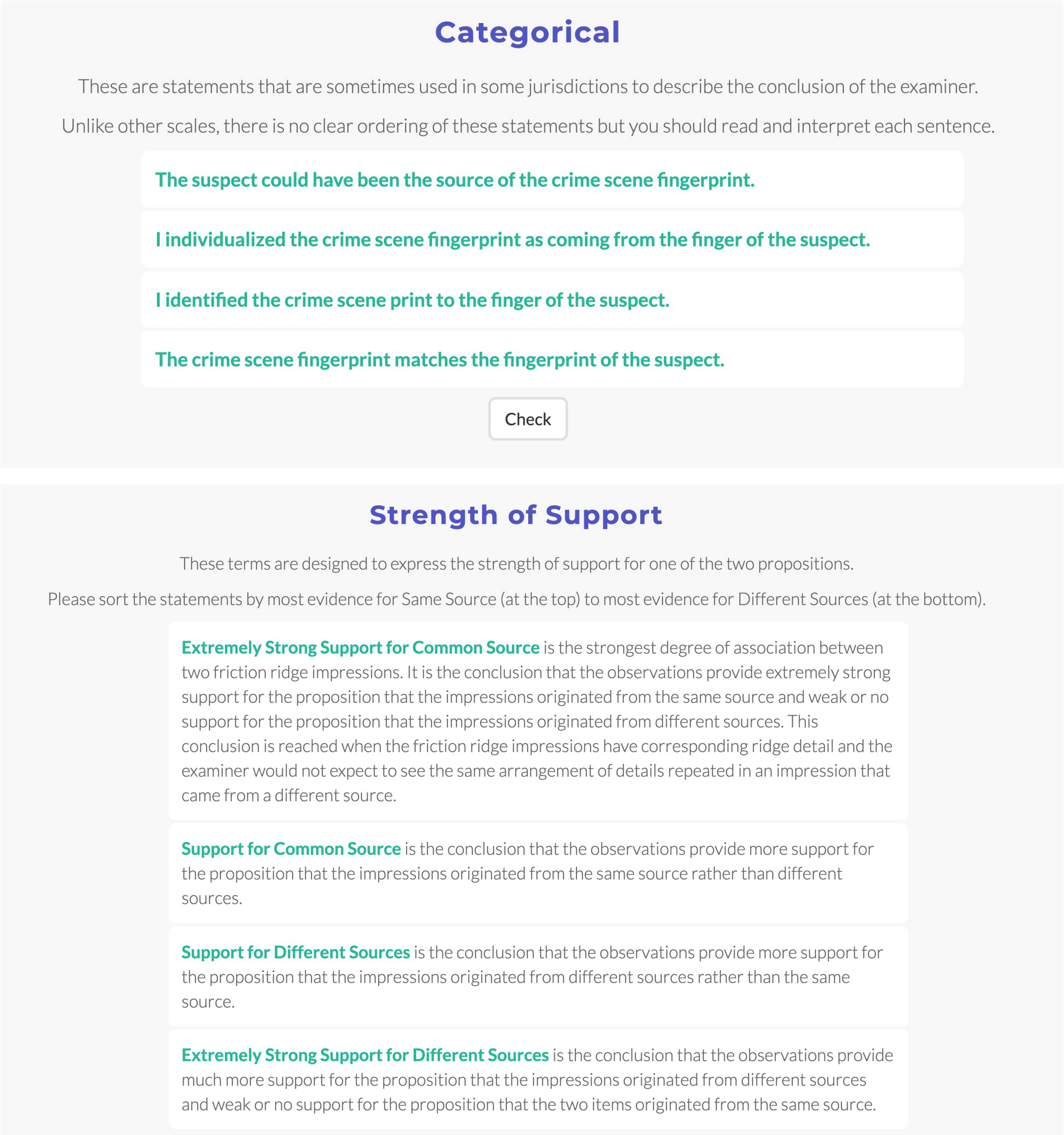
Source probability statements provide the probability of the same-source proposition.

Please sort the statements by most evidence for Same Source (at the top) to most evidence for Different Sources (at the bottom).

**Practical Certainty**: Given the size and quality of the crime scene print, it is a practical certainty that the suspect is the person who made the crime scene print.

**Highly Probable**: Given the size and quality of the crime scene print, it is highly probable that the suspect is the person who made the crime scene print.

**Moderately Probable**: Given the size and quality of the crime scene print, it is moderately probable that the suspect is the person who made the crime scene print.



**Categorical**

These are statements that are sometimes used in some jurisdictions to describe the conclusion of the examiner.

Unlike other scales, there is no clear ordering of these statements but you should read and interpret each sentence.

**The suspect could have been the source of the crime scene fingerprint.**

**I individualized the crime scene fingerprint as coming from the finger of the suspect.**

**I identified the crime scene print to the finger of the suspect.**

**The crime scene fingerprint matches the fingerprint of the suspect.**

**Strength of Support**

These terms are designed to express the strength of support for one of the two propositions.

Please sort the statements by most evidence for Same Source (at the top) to most evidence for Different Sources (at the bottom).

**Extremely Strong Support for Common Source** is the strongest degree of association between two friction ridge impressions. It is the conclusion that the observations provide extremely strong support for the proposition that the impressions originated from the same source and weak or no support for the proposition that the impressions originated from different sources. This conclusion is reached when the friction ridge impressions have corresponding ridge detail and the examiner would not expect to see the same arrangement of details repeated in an impression that came from a different source.

**Support for Common Source** is the conclusion that the observations provide more support for the proposition that the impressions originated from the same source rather than different sources.

**Support for Different Sources** is the conclusion that the observations provide more support for the proposition that the impressions originated from different sources rather than the same source.

**Extremely Strong Support for Different Sources** is the conclusion that the observations provide much more support for the proposition that the impressions originated from different sources and weak or no support for the proposition that the two items originated from the same source.

Figure 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.

# Section 2: Validating Expanded Conclusion Scales for Fingerprints, Toolmarks, and Footwear Impressions

Practitioners working in the pattern comparison disciplines in the US traditionally express their conclusions using an articulation statement drawn from a small set of approved statements. Quantitative tools for assessing the strength of evidence exist in some disciplines, but the vast majority of pattern comparisons are performed by human experts, who conduct manual comparisons between two or more impressions to accumulate evidence of whether the two patterns might share a common origin. In friction ridge comparisons, internal evidence is usually converted to a categorical conclusion such as “Identification”, “Inconclusive”, and “Exclusion/Elimination” and then provided to the consumer such as a detective or the court. Other pattern comparison disciplines use similar conclusions scales but with additional categories, in part because the manufacturing process creates features that tend to be similar (i.e. ‘repeated’ features as opposed to ‘unique’ features that are created through use or wear).

To accurately represent the strength of the evidence, the language that is used to describe the conclusion must be *calibrated*, much like any other measurement system or device. If a conclusion scale is not calibrated, consumers such as detectives, defense attorneys, or jurors may misinterpret a conclusion even if the original comparison was conducted appropriately. However, the translation of evidence from the examiner’s comparison all the way to the consumer has multiple places where information can be lost and inaccuracies can occur. To illustrate how errors in calibration might occur when reporting evidence, Figure 12 illustrates the flow of information during a forensic comparison. In the top row Figure 12, the evidence that is analyzed in the pattern disciplines accumulates in an examiner’s working memory during the comparison. This internal evidence, which may be thought of as on a scale of *perceived detail in agreement* between the two impressions, is eventually translated to a conclusion through the translation function Θ (middle row). In the friction ridge discipline this involves exclusion, inconclusive, and identification categories, while other disciplines use expanded scales that were developed by committees working to standardize conclusion scales.

Note that the translation function Θ maps a continuous internal evidence scale into a small number of discrete conclusions. This essentially throws away information, because some comparisons produce evidence that is close to the boundary between Inconclusive and Identification and the conclusion terminology does not reflect this borderline state. While it is true that an examiner may qualify some conclusions on the stand during testimony (IAI, 2010), the vast majority of cases do not go to trial. Instead, these qualifications or hedges may be ignored or misunderstood by a prosecutor or public defender, who may encourage a suspect to take a plea deal when the evidence may not support one.

An additional source of mis-calibration between the evidence and its use by the justice system can occur during the mapping $\Psi$ between the Examiner's conclusion and the assessment of the nature and strength of the evidence by the Consumer. This third scale requires the consumer to weigh the evidence along an Exculpatory/Inculpatory axis, and the strength of that evidence (or the risk in accepting the examiner's conclusion) must be accurately interpreted. For example, the general consensus in the friction ridge community is that the term *Identification* does not mean to the exclusion of all others (SWGFAST, 2013a). However, recent work by Swofford and Cino (2017) assessed the beliefs of potential jurors, and found that 71% of those surveyed interpreted "identification" to imply "to the exclusion of all others." Thus, there appears to be a disconnect between what examiners intend and how their conclusions are interpreted. In this case, jurors interpret the evidence as stronger than was originally intended by the examiner.

A final source of potential mis-calibration between the true strength of the evidence and how that evidence is interpreted occurs with the choice of terms used by examiners. As each discipline has become more rigorous through error rate studies that characterize the typical evidentiary strength of a discipline, we now have an opportunity to validate the conclusion language against the true strength of the evidence. This was demonstrated recently by (Thompson et al., 2018), who noted that the term "identified" was viewed as equivalent in strength to a random match probability of 1 in 100,000. However, the black box studies in friction ridge found that the empirical error rate was in the range of 1 in 100 or 1 in 1000, which is off by a factor of 100 to 1000. The authors concluded that the language used to express conclusions in friction ridge may overstate the true strength of fingerprints. In response to this criticism, disciplines have proposed new scales that express *strength of evidence* (e.g. Extremely Strong Support for Common Source) rather than absolutist language such as Identification. These new statements are still posteriors because they express support for propositions such as common source rather than observations given propositions such as the probability of the observations given a common source proposition. However, strength of evidence may be preferred over traditional scales because they might not be overinterpreted as in the Swofford and Cino (2017) study.

This article and the companion work (Busey & Klutzke, submitted) together provide a calibration of various conclusion scales by measuring the behavior of forensic practitioners on casework-like comparisons. The understanding of the various conclusion scales is then measured in laypeople to validate the scale through the translations expressed in Figure 12. Together the two sets of studies provide the data to calibrate scales and foster the development of new articulation language.

There is a robust debate in the literature about the use of definitive conclusions such as Identification. For example, the International Association for Identification held a symposium at their annual meeting in Atlanta in 2017 that considered whether the term Identification should be used as a conclusion. While that symposium produced no consensus or strong momentum for change within the forensic community, others outside the community have called for a shift away from definitive statements such as Identification to statements that express the strength of the evidence such as likelihood ratios or verbal equivalents (Aitken et al., 2011; Assoc Forensic Sci Providers, 2009; Martire et al., 2014). Some authors have argued that forensic examiners should not even make sole-source statements (e.g. Evett, 1998; Robertson, Vignaux, & Berger, 2011), and instead argued for a strength-of-evidence framework for conclusions, with language such as 'Supports' rather than definitive conclusions such as 'Identified'. However, whether jurors can understand more complex statistical terms such as likelihood ratios and random match probabilities is an empirical question. In addition, jurors might benefit from having examiners make definitive statements, because examiners may have a better understanding of the context of the examination. There have been cogent arguments on both sides, but the literature contains relatively few direct tests of how examiners would change their behavior with different forms of conclusion statements should they be asked to adopt a new set of conclusion statements.

Prior work on this question by our group revealed a surprising result: the interpretation of the term Identification depended on the scale in which it was embedded (Carter et al., 2020). We presented 60 casework-like comparisons to 27 latent print examiners and asked them use either the traditional 3-conclusion scale, or an expanded one that included *support for common source* and *support for different sources*. Examiners knew on each trial which scale they would use, and we fit the data using an extension of Signal Detection Theory (Macmillan & Creelman, 2004), which provides separate estimates of examiner ability (through an estimate of sensitivity) and examiner response bias (through an estimate of decision criteria). Response bias can be thought of as how risk averse a participant is when making decision: an examiner who makes more Identification conclusions than their colleagues will make more correct decisions but also has an elevated risk erroneous identification outcomes (Mannering, Vogelsang, Busey, & Mannering, 2021). Results from both the raw data and the signal detection modeling fits demonstrated that examiners used the Identification conclusion less often when the Support for Common Source conclusion was available. This means that examiners redefined what was sufficient for an Identification conclusion when they had the Support for Common Source conclusion available as part of the conclusion scale. They also redefined the Inconclusive term, using it less often when Support for Common Source and Support for Different Sources was available.

The present work is a extension and generalization of this approach, including both a direct replication with more realistic time deadlines, as well as comparisons between a traditional scale and a pure strength of evidence conclusion scale in Fingerprint, Footwear, and Toolmark disciplines. Strength of evidence scales have the potential advantage that they focus on the evidence rather than the conclusion of an examiner. A traditional conclusion such as Identification has a definitive nature to the statement and is interpreted as such (Swofford & Cino, 2017). In fields such as DNA where likelihood ratios are common, the evidence is framed in terms of the probability of the observations *given* the hypothesis of same source and the probability of the observations *given* the hypothesis of different sources. This allows the jury to evaluate the evidence in conjunction with the rest of the case without the expert making the decision that is really in the domain of the jury. Strength of support statements also have the potential to seem less definitive, which may be appropriate given the low but non-zero erroneous identification error rates (Ulery et al., 2011). As a result, governing bodies such as the Organization of Scientific Area Committees for Forensic Science (OSAC) are considering new language that relies on strength of evidence statements such as Extremely Strong Support for Common Source (Friction Ridge Subcommittee & OSAC, 2018). The traditional scale for likelihood ratios typically maps this verbal statement to likelihood ratios of 1000-10,000 in the

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

forensic disciplines (Assoc Forensic Sci Providers, 2009), and this is consistent with black box studies that show error rates of around 0.1% for fingerprints (Ulery et al., 2011).

The goal of the present work is to validate proposed strength of evidence scales, to determine how examiners would use new scales in casework-like situations, and use the results in combination with those from a companion article to determine how the statements might be interpreted by laypersons (Busey & Klutzke, submitted).

## Method

Participating forensic examiners each conducted 60 casework-like comparisons in their discipline of expertise. On each trial they were give either the traditional scale from their discipline or a scale based on strength of evidence language. Some fingerprint examiners participated in a conceptual replication of the  Carter et al. (2020) study. The present study was conducted using web-based interfaces written in Javascript, with data stored remotely in a MySQL server. All data was collected according to a Human Subject protocol approved by Indiana University.

### *Participants*

For the fingerprint portion of the study, 66 latent print examiners from forensic facilities participated. They were required to be eligible to testify in the United States. This portion of the study had two groups of participants. 32 examiners compared the traditional scale with an expanded traditional scale, while 34 examiners compare the traditional scale with a pure strength of evidence scale. Of the participants who completed demographic surveys, 50 were female, 13 were male, and one declined to answer. The median age was 40, with an age range of 27 to 62. 21 had no eye correction, 12 had contacts, 27 had glasses, and 4 had Lasik. 12 worked in Federal agencies, 20 worked in local agencies, 10 worked in metro/county agencies, 17 worked in state agencies, 3 worked in other agency types, and 2 preferred not to answer.

For the footwear portion of the study, 32 footwear examiners from forensic facilities participated. They were required to be eligible to testify in the United States. Of the participants who completed demographic surveys, 25 were female and 6 were male. The median age was 43, with an age range of 27 to 71. Eight had no eye correction, 8 had contacts, 10 had glasses, and 5 had Lasik. One worked in a Federal agency, 5 worked in local agencies, 4 worked in metro/county agencies, 20 worked in state agencies and 1 worked in other agency type.

For the toolmark portion of the study, 20 toolmark examiners from forensic facilities participated. This subject count is lower than fingerprint and footwear datasets due to the Covid-19 pandemic, but still sufficient for data analysis. Toolmark examiners were required to be eligible to testify in the United States. Of the participants who completed demographic surveys, 7 were female and 10 were male. The median age was 42, with an age range of 28 to 54. Five had no eye correction, 4 had contacts, 7 had glasses, and 2 had Lasik. Four worked in local agencies, 6 worked in metro/county agencies, and 8 worked in state agencies.

### Stimuli

All stimuli were collected under the supervision of a subject matter expert (Vanderkolk) with the goal of making the trials similar to casework. All images used in the study are available for inspection from the OSF site linked below.

#### *Fingerprints*

Fingerprint impressions were selected from a 3,000-print database collected from volunteering Indiana University staff and students. Each exemplar print was labeled with an anonymized participant code and the hand and finger the print was from, then scanned into an editing software. All exemplar prints were tapped or rolled ink prints. The latent prints were black powder, ninhydrin, black powder on galvanized metal, or ink prints. The latent prints were also labeled with a participant code and the hand and finger, then scanned into the same editing software to create the database. Images were scanned using an Epson Perfection 4870 scanner at 4800 pixels per inch and downsampled to 750 pixels per inch. The final display resolution was dependent on the size of the participant's monitor, but a higher resolution could be accessed through a zoom function.

The latent prints chosen for the study contained various sources of noise such as distortion, scarring, smearing, medium, contrast, and percentage of print present, while the exemplar prints were typically of high quality. Our goal was to create a test set of stimuli that were similar to other error rate studies (e.g. Ulery et al. (2011)), although we do not consider this study to measure error rates, but instead provide a comparison of two reporting scales under conditions that are similar to casework. To that end, we selected our non-mated images using left-right reversed impressions from the opposite hand of the donor individual. We used a subject matter expert (Vanderkolk) to select both mated and non-mated pairs that were similar in difficulty to what examiners experienced during typical casework. Thus, our exemplar impressions for non-mated pairs were designed to be challenging exclusions that for the most part bore superficial similarity to the latent impression.

*Footwear*

Footwear stimuli were collected under the guidance of our subject matter expert (Vanderkolk). We used a collection of shoes drawn from different sources. Half of the trials contained shoes that were the same make and model purchased by a runner who wore them to approximately the same wear level. The other half of the trials contained shoes and light hiking boots that were chosen because at least two pairs of the same make and model were available, and there were 9 different models, some with multiple exemplars. All shoes had been moderately worn. Similar soles with similar amounts of wear were used to produce challenging impressions for the study. Only heel impressions were used for the study because image acquisition proved to be easier to manage and more reliable. All images used in the study are available from the OSF site linked below.

Shoe prints, or impressions, were made using different techniques to produce images bearing various qualities and quantities of details. One technique consisted of applying extremely light to somewhat heavy mixtures of petroleum jelly and black finger print powder to the soles of the shoes. This mixture was applied to gloved fingers then gently rubbed onto the sole. Then, the soles of the shoes were pressed, rolled, or slapped onto pieces of white paper. The other technique consisted of applying melted chocolate ice cream to the soles of the shoes. Melted chocolate ice cream was chosen to produce a dark viscous matrix that dried quickly for the impressions. With the melted chocolate ice cream on a flat paper plate, the soles were tapped into the ice cream. Then, the soles of the shoes were pressed, rolled, or slapped onto pieces of white paper. Some of the impressions either had been made from areas of the soles that were relatively clean or areas that had been previously used to make the petroleum jelly/powder impressions. These techniques produced various qualities and quantities of recorded details in the impressions. Areas from the prints were selected to produce comparison pairs that ranged from easy to difficult.

Once dried, the impressions were scanned at 600 pixels per inch using an Epson V600 scanner. Images were then downsampled to 200 pixels per inch. Photographs of the known shoe image were taken with a Sony α7IIIr camera with a FE 1.4/24 GM lens and downsampled to match the pixels per inch of the scanned images. The final display resolution was dependent on the size of the participant's monitor, but a higher resolution could be accessed through the zoom function.

Mated pairs were created by pairing the questioned images from the simulated crime scene methods with the gel lifts and photographs of the same shoe. Nonmated pairs were created by using impressions from the same make and model shoe yet different shoes. The chosen shoes were potentially quite difficult due to the fact that not only were they the same make and model, but some had been worn by the same individual and therefore likely subject to the similar, but still different, wear patterns. Because of this, we do not consider this dataset to accurately represent error rates in the field. Instead, the goal is to identify differences between the two scales under comparison, and so we merely require a dataset that is somewhat similar to casework. All images used in the study are available from the OSF site linked below.

*Toolmarks*

Striated Toolmarks were collected from 15 quarter-inch screwdrivers (Craftsman 9-41584 1/4" x 6" Slotted Screwdriver) and 15 quarter-inch wood chisels (TEKTON 67551 1/4-Inch Wood Chisel). These were purchased in the same order from Amazon, although we have no control over the batch origin. We constructed a custom 3D printed jig that was used to produce striated scrapings in heavy-duty aluminum foil (see Figure 13), and while we collected scrapings at 5 different angles (10, 20, 35, 55, and 80 degrees), we judged the 20 and 35 degree angles to be most representative of what might be produced by tools used on metal window and door frames to gain access to property. Thus we used only scrapings collected using either of these two angles. Each tool was used to create three separate scrapings at each angle, and care was taken to mark the tool face used to create the marks because flat screwdrivers are double-sided. Chisel blades were single sided. The toolmarks were lit with oblique lighting using a fiber optic light source and photographed with a Sony α7IIIr camera and FE 2.8/90 Macro G OSS lens with two Kenko DG extension tubes totaling 26 mm extension to improve image capture size. The images were then downsampled to 1500 pixels per inch. The final display resolution was dependent on the size of the participant’s monitor, but a higher resolution could be accessed through the zoom function.

Mated pairs were created by selecting one scraping from each tool at either 20 or 35 degrees, and then presenting the 20 and 35 degree images from the same tool from a different scraping. Nonmated pairs were created by selecting similar-looking scrapings from different tools.

While it is difficult to determine whether the task difficulty was comparable to typical casework, the goal of the experiment is to compare two scales, and thus we simply need the task difficulty to be generally similar to casework. All images used in the study are available from the OSF site linked below.

### *Instructions*

The instructions for all three sets of participants included descriptions of the different scales. The definitions for each scale are found in Table 6 and Table 7 for fingerprint examiners, Table 8 for Footwear examiners, and Table 9 for Toolmark examiners. In addition to these definitions, the instructions included the following general statements:

*Fingerprints*

Within the field of latent print identification, various groups, including the Friction Ridge Subcommittee of OSAC, are contemplating changes to the way that conclusions are reported. The groups are proposing additional categories beyond the traditional Identification/Inconclusive/Exclusion conclusions that have historically been used. The goal of this experiment is to understand the consequences of moving to different conclusions scales, and we are testing scales that have some language in common with the Draft Standard for Friction Ridge Examination Conclusions as produced by the Friction Ridge Subcommittee of OSAC.

*Footwear*

Within the pattern comparison disciplines, various groups are contemplating changes to the way that conclusions are reported, including a shift to language that expresses conclusions according to the strength of support for one of two propositions (Common Source or Different Sources). The goal of this experiment is to understand the consequences of moving to conclusions scales that express conclusions according to strength of support for propositions. These strength-of-support statements are alternatives to definitive statements such as Identification or Exclusion. However, before we mandate new language, we need to understand the consequences of adopting new conclusion language, and thus this experiment.

*Toolmarks*

Within the field of firearm and toolmark comparisons, various groups, including the Firearms & Toolmarks Subcommittee of OSAC, are contemplating changes to the way that conclusions are reported. However, before any change is made we need to understand the consequences of such a change. The goal of this experiment is to understand the consequences of moving to different conclusions scales.

### Procedure

*Fingerprints*

The study was composed of 60 trials for each participant, and each trial consisted of one fingerprint comparison. The experiment was administered electronically using a custom Javascript interface designed to mimic the tools available during casework (see Figure 14). On each trial, the latent print was placed on the left side of the screen next to an exemplar print on the right side, as shown in Figure 14. The interface allowed the participants to zoom, rotate, and pan the individual images, as well as mark individual features with transparent digital markers.

Each trial began with an "of value" decision, which in casework allows the examiner to decide not to proceed with a comparison due to poor quality of the latent impression. However, while we recorded this response, we still required the participant to complete the trial. We made this decision because the interpretation of our results depend in part on model fits from signal detection theory, and it is difficult to fit models in which an initial quality threshold is assessed. Both scales included an 'inconclusive' category, and while we understand that in casework 'no value' and 'inconclusive' have different meanings, we considered the two to be approximately equal for the purposes of comparing the 3-conclusion and 5-conclusion scales. We also randomized the assignment of images to conditions (3-conclusion and 5-conclusion scales) across participants, and thus we would not expect a systematic bias of image quality on one of the two scales. Participants did not know on each trial which scale they would use until after they had made the 'of value' determination, and so we are unlikely to observe differences in the 'of value' rates between the two types of scales.

After making an 'of value' determination, the exemplar impression also became visible. Data collection was terminated after 30 minutes for expediency sake, with a "Pause" button available that hid the trial and paused the countdown timer until the "Resume" button was selected. At the end of the 30 minute mark or when the participant pressed the "Next" button, they were allowed to state their conclusion on a screen that hid the fingerprint comparison. Participants completed 30 trials using the 3-conclusion scale and 30 trials using the 5-conclusion scale. Images were randomly assigned to condition for each participant, and the order of the images and conditions was randomized for each participant. As a result of this randomization we would not expect our results to be affected by, say, difficult trials only being assigned to the traditional scale. Half of the trials were designated as mated pairs, and half were non-mated.

Participants received only the instructions and training provided by the text in either Table 6 or Table 7 depending on the comparison they were randomly assigned, and did not have extensive training on the new categories in the expanded or Strength of Evidence scales. We acknowledge that the behavior of examiners may change as they adapt to the use of novel statements if they were to be included in operational casework. For the participants in the comparison between Traditional and Expanded Traditional scales, both scales use the 'Identification' and 'Exclusion' categories as shown in Table 6, which examiners have had experience with and presumably should not change, although this is an empirical question. Participants in the Traditional and Strength of Evidence comparison condition were provided the definitions of the terms shown in Table 7.

There was one other difference between the two Fingerprint participant groups: The comparisons for the two groups are asking slightly different scientific questions. The first comparison conceptually asks "what would happen if we added two new categories to the existing 3-conclusion scale?" Because both scales included the Exclusion and Identification conclusions, we provided explicit definitions of these terms as illustrated in Table 6. However, the second comparison conceptually asks "if we switched to a Strength of Support conclusion scale, how would the response distribution change?" In this case we asked participants to use the definitions that they had refined during casework, and we did not provide definitions for the traditional terms. We will show that this difference across conditions did not meaningfully change how they interpreted the traditional terms (discussed in Figure 20 as part of the results section).

### *Footwear*

The instructions and procedure for the Footwear examiners was similar to those for Fingerprint examiners, but included the instructions shown in Table 10. The images could be rotated and the known image could be colorized and dragged over the questioned impression. Either image could be toggled on and off to aid in the comparison process. No marks were allowed in this interface. The design included a 30 minute timer and a pause button.

*Toolmarks*

The instructions and procedure for the Toolmark examiners was similar to those for Fingerprint examiners, but included the instructions shown in Table 11. The interface included a split screen controlled by a slider. The questioned image (on the left in Figure 16) could be dragged around, as could the known impressions on the right in Figure 16. The images could be rotated and zoomed. No participant markings were allowed in this interface. The design included a 30 minute timer and a pause button.

## Results

All images, data, and analysis code are available at the OSF repository, which also contains the data and analysis files for the companion paper:

https://osf.io/xmwqg/?view_only=f1b996eee77d45d0907ecebdaa27437d

Table 12 through Table 15 provide the response distributions for the various conditions, which are discussed below. In principle, the each row in these tables will sum to a multiple of the number of participants multiplied by 15, which was the number of trials per participant in that row. However, a rare data saving problem of unknown origin (likely due to intermittent network problems) resulted in 8 fingerprint examiners with 59 trials (four in each participant group), 3 footwear examiners with 59 trials, and no toolmark examiners with missing data. In addition, one toolmark mated pair was incorrectly identified as nonmated when images were assigned to participants. This error was corrected during the analysis and resulted in one additional mated pair and one fewer nonmated pair trial assigned to each participant. While these issues are regrettable, the modeling section (described below) is almost completely unaffected by this unequal distribution of responses across mated and nonmated trials, because we are comparing across the two scales which were both affected by these issues, and because the modeling relies on frequencies not raw trial counts.

### *Response Distributions*

*Fingerprint Examiners*

Table 12 illustrates the response distribution for participants who compared the traditional scale with the expanded traditional scale. Table 13 illustrates the response distribution for participants who compared the traditional scale with the strength of support scale. There were six erroneous identification responses across the two groups. Four pairs had one erroneous identification outcome, while one pair had two erroneous identification outcomes. In each case the ground truth was verified against the original scans to verify the nonmated status of each of the five nonmated pairs that produced erroneous identification outcomes. Combining over both participant groups and scale types, there were 15 erroneous Support for Common Source outcomes and 86 erroneous exclusion or erroneous extremely strong support for different source outcomes (30 and 26 erroneous exclusions from the traditional scale from the two participant groups, 22 erroneous exclusions from the expanded traditional scale, and 8 erroneous extremely strong support for different source outcomes from the strength of support scale).

Consistent with Carter et al. (2020), the number of correct identification outcomes dropped when the scale was expanded. This was somewhat pronounced in the Traditional/Expanded Traditional comparison (231 to 199 in Table 12), and more pronounced in the Traditional/Strength of Support comparison (250 to 180 in Table 13). These results are again consistent with the finding that examiners redefine the definition of the term Identification when the scale is expanded (see Table 12) or are less likely to use Extremely Strong Support for Common Source than the term Identification (see Table 13).

The number of Inconclusive responses to mated pairs also dropped as the traditional scale was expanded. These dropped from 220 to 131 in Table 12, and from 244 to 179 in Table 13. These results demonstrate that the Support for Common Source response to mated pairs is a mixture of what would have been Inconclusive and Identification responses in the Traditional scale.

The number of Exclusion responses to nonmated pairs also dropped for the expanded scales. These responses dropped from 272 to 225 in Table 12 and 265 to 195 in Table 13. This suggests that examiners become risk averse for the expanded scales on the exclusion side.

### *Footwear Examiners*

Table 14 provides the response distributions for footwear examiners. There were five total erroneous identification outcomes distributed across five different nonmated sets. In each case, the ground truth was verified by accessing the raw images collected from the scanner or photography rig that contained the shoe pair number, and in each case these nonmated pairs were verified as coming from different shoes. However, as previously noted, many of these shoes are not only of the same make and model, but were worn by the same individual and retired with similar wear because of the nature of the running activity. This difficulty may not be fully representative of casework as a result, but the results still allow for comparisons across scales.

Both the Traditional and Strength of Support scales have six categories, and the question of interest is whether the response distribution changes as the scale changes. We might find, for example, that examiners are reluctant to use a particular statement such as Extremely Strong Support for Common Source. However, Table 14 demonstrates that there were no large differences between the two conclusion scales, with perhaps a slight drop between High Degree of Association and Strong Support for Common Source. Thus these scales seem to be treated more or less equivalently by participants.

### *Toolmark Examiners*

Table 15 provides the response distributions for the toolmark examiners. There were four erroneous identification outcomes, which were distributed across four different nonmated pairs. In each case the ground truth was verified against the original scans to verify the nonmated status of each of the four pairs. It is difficult to establish the task difficulty of these comparisons relative to casework, although the fact that the toolmarks were created by tools of the same make and model does make this a particularly challenging task. However, the response distributions do allow for comparisons across scales, which is the intent of the study.

Both the Traditional scale and the Strength of Support scale have 5 statements, and the question of interest is whether the response distributions are similar across the two scales. That is, do examiners treat the two scales in the same way? We find that Extremely Strong Support for Common Source demonstrates a degree of risk aversion, because participants used this response category for mated pairs much less often than the Identification response (136 vs 178). Most of these responses that might have been Identification responses in the Traditional scale appear to have been moved to the Support for Common Source, because the outcomes grew from 47 for Insufficient for Identification to 84 in the Support for Common Source.

Evidence for risk aversion in the exclusion outcomes is evident for the Strength of Support scale, because the number of correct exclusions drops from 84 in the Traditional scale to 45 in the Strength of Support scale.

### Estimating Decision Criterion

The response distributions presented in Table 12 through Table 15 can be summarized using extensions to Signal Detection Theory (Macmillan & Creelman, 2004). As in Carter et al. (2020), we fit the response distributions with a model that assumes that the result of each comparison produces a unidimensional value on an internal evidence axis, which is then mapped to a categorical statement using a set of decision criteria. The distribution of nonmated and mated pairs along this evidence axis are summarized using Gaussian distributions, and separate decision criteria are fit to each scale. Further details are found in Carter et al. (2020), but for the present work we fit the combined data across all subjects using the brms library (Bürkner, 2017) in R (Team, 2013).

The goal of signal detection theory is to separate ability (as measured by d') from response aversion/bias (as measured by the decision criteria). Although it is possible for sensitivity (d') to differ across scales, prior work found no evidence for this, and thus we first established a common d' and standard deviation value for the mated pair distribution, and then fit separate decision criteria for each scale.

The results of the modeling for each participant group are provided below. We fit the combined data for each group rather than individual participants, because we are interested in how the field as a whole would respond if the conclusion scale were changed. In our earlier work with a very similar design, we fit individual participants in addition to group data and found similar results across the two types of fits (Carter et al., 2020) .

*Fingerprint Examiners*

The sensitivity (d') value for fingerprint examiners across the two participant groups was 2.39, with a standard deviation for the mated pair distribution of 1.48. This difficulty level is consistent with other error rate studies (Ulery et al. (2011); see Mannering et al. (2021)) and thus the task difficulty appears similar to that of casework. Figure 17 illustrates the results of this modeling, and shows the location of different decision criteria for the two scales. The color bands represent 95% confidence intervals around the decision criterion estimates. As was suggested by the response distributions in Table 12, the decision criteria for Identification in the Expanded scale is shifted to the right of the Identification decision criteria for the Traditional scale. This is consistent with prior results (Carter et al., 2020) and provides evidence that examiners become more risk averse with expanded scales. A similar result is found with Exclusion, were examiners are less likely to use this response category in the expanded scale, thus pushing the Exclusion decision criteria from the Expanded scale to the left.

Examiners become even more risk averse when asked to use the Extremely Strong Support for Common Source conclusion, as shown in Figure 18. They are also less likely to use the Extremely Strong Support for Different Sources conclusion than the Exclusion conclusion. In each case, examiners become more risk averse with the expanded scale. The Inconclusive area in the traditional scale also shrinks when the scale is expanded. These results demonstrate that examiners reserve conclusion statements that include Extremely Strong Support for only those conclusions with the highest amount of support.

Although we did not compare the Expanded Traditional and Strength of Support scales directly, we can do a virtual comparison across participant groups because our model fits rely on a common d' and mated distribution standard deviation across the two participant groups. Figure 19 illustrates the decision criteria for the two five-item scales, and demonstrates that examiners are more risk-averse when using the strength of support scale than the expanded traditional scale. This is again consistent with the above result that suggests that any conclusion statement that contains Extremely Strong Support is reserved for comparisons with the highest amount of support.

Finally, note that there were subtle differences in the instructions given to the two participant groups with respect to the use of the Identification, Inconclusive, and Exclusion terms. This was done deliberately, because the data from the two groups is being used to address slightly different scientific questions. However, we see that the two groups performed very similarly with respect to the placement of their decision criteria, as shown in Figure 20. The Identification criteria are almost identical, and the Exclusion criteria are also quite similar. Thus we feel that the subtle differences in instructions still allow for comparisons across the two groups as in Figure 19.

### *Footwear Examiners*

The fitted values of d' is 2.14 and the standard deviation for the mated pairs was 1.13. The response distributions across the two scales shown in Table 14 demonstrated that there were no large shifts in responses across the two scales. Consistent with this result, Figure 21 demonstrates that the fitted decision criteria across the two scales were very similar, with perhaps a slight difference between the High Degree of Association and Strong Support for Common Source decision criteria. Thus it appears that footwear examiners treat these two scales in a very similar manner.

*Toolmark Examiners*

The fitted values of d’ is 2.49 and the standard deviation for the mated pairs was 1.77. The response distributions in Table 15 illustrated that toolmark examiners grew more risk-averse when using the Strength of Support scale. As shown in Figure 22, the fitted decision criteria for the Extremely Strong Support for Common Source is to the right of the Identification decision criterion, demonstrating increased risk aversion for the Strength of Support scale. This is also true for the Exclusion/Extremely Strong Support for Different Sources comparison. It appears that toolmark examiners become much more risk-averse when using the Strength of Support scale. This result is consistent with that observed with Fingerprint examiners, although Footwear examiners did not show evidence of such a shift.

## Discussion

The present work provides four clear conclusions:

1) In fingerprint comparisons, participants redefined the term Identification when Support for Common Source was included in the conclusion scale, relative to the traditional scale. This is a direct replication of the Carter et al. (2020) result. The Support for Common Source category absorbed some of the weaker Identification conclusions from the traditional scale, as well as some of the stronger Inconclusive conclusions from the traditional scale. We view both of these as positive outcomes, because perhaps some of the weaker identifications may have been borderline and arguably at the boundary of sufficiency for Identification, and some of the stronger Inconclusive conclusions probably merited an investigative lead.

2) Fingerprint examiners also became more risk averse when moving from the traditional scale to the strength of support scale. Surprisingly, they show a strong reticence to use the Extremely Strong Support for Common Source conclusion relative to their use of Identification in the traditional scale. We view this as surprising, because in a companion article (Busey & Klutzke, submitted) we found that examiners viewed Extremely Strong Support for Common

Source as implying *less* evidence than Identification for the proposition of common source when comparing the two on a visual scale. This disconnect perhaps reduces the utility of the Extremely Strong Support for Common Source conclusion as a policy recommendation, because examiners might use it less often, yet think it means something less than it does. Members of the general public, however, interpret it at equivalent to Identification (Busey & Klutzke, submitted), further reducing the utility of this term as a conclusion scale statement.

One possibility that we did not test is whether Strong Support for Common Source (without the term 'Extremely') might be a more appropriate endpoint to a strength of support scale. Examiners may show less risk aversion to using this phrase, and this phrase may be more justified given the error rate studies that show an erroneous identification rate of .1% (Ulery et al., 2011).

3) In footwear comparisons, the behavior examiners may not change if a shift is made to a strength of support conclusion scale. The guiding principles for such a shift might be whether these statements accurately reflect the typical strength of the evidence in casework. Thompson and Newman (2015) found that prior beliefs about a discipline affect evidence interpretation by mock jurors, and so even if members of the general public interpret the highest categories on different scales as equivalent, they will probably contextualize this result given their understanding of the individual discipline.

4) Toolmark examiners exhibited strong risk aversion when using the strength of support conclusion scale, similar to that observed with fingerprint examiners (see Figure 22). As with fingerprint comparisons, we suggest that perhaps Extremely Strong Support for Common Source is too strong relative to the strength of the evidence in a discipline, and that the discipline might consider Strong Support for Common Source as the highest category of conclusion statements.

In general, we view expanded conclusion scales as an improvement over scales with just three statements, as expanded scales lose less information at the border between two categories, and provide investigative leads with some of the weaker conclusions. However, the consumer must be taught to interpret the conclusion scale properly, which should include saying what could have been concluded but was not. There are also other operational considerations when considering a change of scales. For example, examiners may have to work longer to reach an Identification conclusion than a Support for Common Sources conclusion, and may decide to terminate the examination process earlier if they can make a less definitive conclusion when using an expanded scale. This interacts with the lab's current backlog and how consumers of that agency use less definitive conclusions, and therefore our data don't bear directly on what might happen operationally should an expanded scale be introduced. We suggest that each lab conduct its own validation studies to determine the possible effect of expanded scales given their own policies and constraints. For example, labs with large backlogs may benefit from making relatively rapid investigative lead conclusions if full Identification conclusions would take extensive time and effort.

Strength of Support scales in two of the three disciplines resulted in a shift of examiner behavior toward becoming more risk averse, because participants used the Extremely Strong Support for Common Source conclusion less often than Identification. Thus while strength-of-support language may focus on the evidence rather than the examiner (i.e. 'the evidence supports' as opposed to 'I identified'), the words Extremely Strong Support may not be justified by the error rate studies in a given discipline, and the examiners may have understood this by using this term infrequently and only for the strongest cases. Consideration should be given to how the consumers might interpret various statements, and readers should consult the companion article (Busey & Klutzke, submitted) for details on how members of the general public view candidate articulation statements.

## Tables

| Traditional Scale | Expanded Traditional Scale |
|---|---|
| **Identification**: Identification is the strongest degree of association between two friction ridge impressions. It is the conclusion that the observations provide extremely strong support for the proposition that the impressions originated from the same source and extremely weak support for the proposition that the impressions originated from different sources. Identification is reached when the friction ridge impressions have corresponding ridge detail and the examiner would not expect to see the same arrangement of details repeated in an impression that came from a different source. | **Identification**: Identification is the strongest degree of association between two friction ridge impressions. It is the conclusion that the observations provide extremely strong support for the proposition that the impressions originated from the same source and extremely weak support for the proposition that the impressions originated from different sources. Identification is reached when the friction ridge impressions have corresponding ridge detail and the examiner would not expect to see the same arrangement of details repeated in an impression that came from a different source. |
|  | **Support for Common Source**: Support for Same Source is the conclusion that the observations provide more support for the proposition that the impressions originated from the same source rather than different sources; however, there is insufficient support for an Identification. |
| **Inconclusive**: The observed characteristics of the items are insufficient to support any of the other conclusions (including one of the 'support' conclusions if they are available). | **Inconclusive**: The observed characteristics of the items are insufficient to support any of the other conclusions (including one of the 'support' conclusions if they are available). |
|  | **Support for Different Sources**: Support for Different Sources is the conclusion that the observations provide more support for the proposition that the impressions originated from different sources rather than the same source; however, there is insufficient support for an Exclusion. |
| **Exclusion**: Exclusion is the conclusion that two friction ridge impressions did not originate from the same source. Exclusion is reached when in the examiner's opinion, considering the observed data, the probability that the two impressions came from the same source is considered negligible. | **Exclusion**: Exclusion is the conclusion that two friction ridge impressions did not originate from the same source. Exclusion is reached when in the examiner's opinion, considering the observed data, the probability that the two impressions came from the same source is considered negligible. |

Table 6. Traditional and Expanded Traditional statements that friction ridge examiners were asked to use during casework-like comparisons. In each trial, they knew which set of statements they would be required to use.

| Traditional | Strength of Support |
|---|---|
| **Identification** | **Extremely Strong Support for Common Source:** Extremely Strong Support for Common Source is the strongest degree of association between two friction ridge impressions. It is the conclusion that the observations provide extremely strong support for the proposition that the impressions originated from the same source and weak or no support for the proposition that the impressions originated from different sources. This conclusion is reached when the friction ridge impressions have corresponding ridge detail and the examiner would not expect to see the same arrangement of details repeated in an impression that came from a different source. |
| | **Support for Common Source**: Support for Common Source is the conclusion that the observations provide more support for the proposition that the impressions originated from the same source rather than different sources. |
| **Inconclusive** | **Inconclusive**: The observed characteristics of the items are insufficient to support any of the other conclusions. |
| | **Support for Different Sources**: Support for Different Sources is the conclusion that the observations provide more support for the proposition that the impressions originated from different sources rather than the same source. |
| **Exclusion** | **Extremely Strong Support for Different Sources**: Extremely Strong Support for Different Sources is the conclusion that the observations provide much more support for the proposition that the impressions originated from different sources and weak or no support for the proposition that the two items originated from the same source. |

Table 7. Traditional and Strength of Support statements that friction ridge examiners were asked to use during casework-like comparisons. Definitions were not provided for the traditional scale for this group of participants, but instead they read the following instructions: "For the traditional categories of Exclusion, Inconclusive, and Identification, we would like you to use the criteria that you use in casework. You may choose from Exclusion, Inconclusive, or Identification on each trial for your conclusion."

| Definitive Conclusions | Strength of Support |
|---|---|
| **Identification**: The footwear impressions correspond in physical size, design, class, wear, and randomly acquired characteristics. The likelihood of observing this quality and quantity of correspondence if the questioned impression was made by a different source is considered extremely low. | **Extremely Strong Support for Common Source**: The questioned impression and the impression from the known footwear share sufficient quality and quantity of agreement of class, wear, and randomly acquired characteristics. The observed characteristics provide extremely strong support for the proposition that the questioned impression was made by the known footwear **and** little to no support for the proposition that the questioned impression was made by a different source. |
| **High Degree of Association**: The footwear impressions appear to have strong associations; however, the quality and quantity of shared characteristics are insufficient for an identification. Other footwear with the same class characteristics as observed in the known impression are included in the population of possible sources only if they display similar wear and randomly acquired characteristics as observed in the questioned impression. | **Strong Support for Common Source**: The observed characteristics exhibit strong associations between the questioned impression and the impression from the known footwear. These characteristics offer stronger support for the proposition that the questioned impression came from the known footwear than for the proposition that the questioned impression came from another source. Other footwear with the same class characteristics as observed in the known impression are included in the population of possible sources only if they display similar wear and randomly acquired characteristics observed in the questioned impression. |
| **Limited Association**: The footwear impressions correspond in size and shape of class characteristics. Other footwear having similar class characteristics may be included as possible sources. | **Support for Common Source**: The questioned impression and the impression from the known footwear correspond in class characteristics. The observed characteristics of the items provide more support for the proposition that the questioned impression came from the known footwear than for the proposition that the questioned impression came from another source. Other footwear with the same class characteristics as observed in the known impression are included in the population of possible sources. |
| **Inconclusive**: Evaluation of the footwear impressions is inconclusive due to insufficient data to support an inclusion or exclusion conclusion of the shoe as a possible source. | **Indeterminate With Respect to Source**: The observed characteristics are insufficient or too ambiguous to support any other source conclusions, as defined in the other sections, or support the two competing propositions equally. |
| **Indications of Non-Association**: The footwear impressions have dissimilarities which indicate non-association; however, the details or features are not sufficient to permit an exclusion. | **Support for Different Sources**: The questioned impression exhibits dissimilarities when compared to the known footwear and provide stronger support for the proposition that the questioned impression came from a different source than the proposition that the questioned impression came from the known footwear. |

| | |
|---|---|
| **Exclusion**: The two impressions originated from different footwear. | **Extremely Strong Support for Different Sources**: Sufficiently significant differences were noted in class tread design or sufficiently significant differences were noted in the comparison of wear or randomly acquired characteristics between the questioned impression and the impression from known footwear to state that the known footwear is not capable of having made the questioned impression. (Such as, there is significantly different wear or randomly acquired characteristics between the impressions, especially when there is more wear or randomly acquired characteristics in the questioned impression than the known impression.) |

Table 8. Definitive conclusion and strength-of-evidence statements used by footwear examiners.

| Definitive Conclusions | Strength of Support |
|---|---|
| **Identification:** Agreement of all discernible class characteristics and sufficient agreement of a combination of individual characteristics where the extent of agreement exceeds that which can occur in the comparison of toolmarks made by different tools and is consistent with the agreement demonstrated by toolmarks known to have been produced by the same tool. | **Extremely Strong Support for Common Source:** Extremely Strong Support for Common Source is the strongest degree of association between two tool marks. It is the conclusion that the observations provide extremely strong support for the proposition that the tool marks originated from the same source and weak or no support for the proposition that the tool marks originated from different sources. This conclusion is reached when the tool marks have corresponding detail and the examiner would not expect to see the same arrangement of details repeated in a tool mark that came from a different source. This conclusion implies agreement of all discernible class characteristics and therefore the basis for this conclusion comes from the observed individual characteristics. |
| **Insufficient for Identification:** Agreement of all discernible class characteristics and some agreement of individual characteristics, but insufficient for an identification. | **Support for Common Source:** Support for Common Source is the conclusion that the observations provide more support for the proposition that the tool marks originated from the same source rather than different sources. This conclusion implies agreement of all discernible class characteristics and therefore the basis for this conclusion comes from the observed individual characteristics. |
| **Inconclusive**: Agreement of all discernible class characteristics without agreement or disagreement of individual characteristics due to an absence, insufficiency, or lack of reproducibility. | **Inconclusive:** The observed characteristics of the items are insufficient to support any of the other conclusions. This conclusion implies agreement of all discernible class characteristics and therefore the basis for this conclusion comes from the observed individual characteristics. |
| **Insufficient for Elimination**: Agreement of all discernible class characteristics and disagreement of individual characteristics, but insufficient for an elimination. | **Support for Different Sources**: Support for Different Sources is the conclusion that the observations provide more support for the proposition that the tool marks originated from different sources rather than the same source. This conclusion implies agreement of all discernible class characteristics and therefore the basis for this conclusion comes from the observed individual characteristics. |
| **Elimination**: Significant disagreement of discernible class characteristics and/or individual characteristics. | **Extremely Strong Support for Different Sources:** Extremely Strong Support for Different Sources is the conclusion that the observations provide much more support for the proposition that the tool marks originated from different sources and weak or no support for the proposition that the tool marks originated from the same source. This conclusion can be made on the basis of either class characteristics or individual characteristics. |

Table 9. Definitive conclusions and strength-of-evidence statements used by toolmark examiners.

You will be completing 60 footwear comparisons using an online interface we've developed for this purpose. Please make the following assumptions about the known shoe/boot:

1) The shoe was recovered almost immediately after the crime was committed, and so you should assume that there was *no opportunity* for wear or alteration to occur on the shoe.

2) Each trial consists of a questioned image on left, a gel test impression of the suspect's shoe in the middle, and a photograph of the suspect's shoe on the right.

3) You may observe differences due to variable pressure between the two impressions. This results from the fact that the technician who made the test impressions did not know what pressure the criminal used when placing the mark. There may also be slight distortion in the photographs from the vice used to hold the shoe for photography.

You will be using different scales on different trials, which will allow us to compare the two scales. We would like you to use one of the following two scales when making your conclusions, and we will tell you which scale you will use at the start of each trial.

The definitions for both conclusion scales are below. You are welcome to print this page if you would like these definitions to be available during your comparisons.

Table 10. Instructions given to Footwear examiners.

You will be completing 60 tool mark comparisons using an online interface we've developed for this purpose. Please make the following assumptions about the known tool:

- The tool was recovered almost immediately after the crime was committed, and so you should assume that there was *no opportunity* for wear or alteration to occur on the tool.
- There will be two test impressions in the comparisons, one at 20° and one at 35°. From the crime scene you are able to ascertain that the tool was used at an angle that falls within this range.
- You may observe differences due to variable pressure between the two impressions. This results from the fact that the technician who made the test impressions did not know what pressure the criminal used when using the tool.
- The tools included in the dataset are 1/4" screwdrivers and 1/4" chisels. All are impressed on heavy-duty aluminum foil. You should make no assumptions about the questioned impression, other than it is either a screwdriver or a chisel, nor should you assume that each trial contains *only* screwdrivers or chisels. Some trials may contain a questioned mark from a screwdriver, and test impressions from a chisel, for example. However, both test impressions were definitely made by the same tool, just at different angles.

Table 11. Instructions given to Toolmark examiners.

| | Traditional Scale | | | | |
|---|---|---|---|---|---|
| **Ground Truth** | Exclusion | | Inconclusive | | Identification |
| Nonmated | 272 | *NA* | 207 | *NA* | 2 |

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

| | | | | | |
|---|---|---|---|---|---|
| Mated | 30 | *NA* | 220 | *NA* | 231 |

**Expanded Traditional Scale**

| **Ground Truth** | Exclusion | Support for Different Sources | Inconclusive | Support for Common Source | Identification |
|---|---|---|---|---|---|
| Nonmated | 225 | 92 | 154 | 6 | 2 |
| Mated | 22 | 33 | 131 | 93 | 199 |

Table 12. Response distribution for Fingerprint participants in the experimental group that compared the traditional response scale to the expanded traditional scale.

**Traditional Scale**

| **Ground Truth** | Exclusion | | Inconclusive | | Identification |
|---|---|---|---|---|---|
| Nonmated | 265 | *NA* | 254 | *NA* | 1 |
| Mated | 26 | *NA* | 244 | *NA* | 250 |

**Strength of Support Scale**

| **Ground Truth** | Extremely Strong Support for Different Sources | Support for Different Sources | Inconclusive | Support for Common Source | Extremely Strong Support for Common Source |
|---|---|---|---|---|---|
| Nonmated | 195 | 127 | 185 | 9 | 1 |
| Mated | 8 | 37 | 179 | 117 | 180 |

Table 13. Response distribution for Fingerprint participants in the experimental group that compared the traditional scale to a pure strength-of-support scale.

| | Definitive Conclusions (Traditional) Scale | | | | | |
|---|---|---|---|---|---|---|
| Ground Truth | Exclusion | Indications of Non-Association | Inconclusive | Limited Association | High Degree of Association | Identification |
| Nonmated | 260 | 126 | 11 | 65 | 15 | 3 |
| Mated | 25 | 23 | 17 | 117 | 146 | 151 |
| | Strength of Support Scale | | | | | |
| Ground Truth | Extremely Strong Support for Different Sources | Support for Different Sources | Indeterminate with Respect to Source | Support for Common Source | Strong Support for Common Source | Extremely Strong Support for Common Source |
| Nonmated | 258 | 123 | 33 | 51 | 15 | 2 |
| Mated | 18 | 29 | 23 | 147 | 103 | 160 |

Table 14. Response distribution for footwear examiners.

| | Traditional Scale | | | | |
|---|---|---|---|---|---|
| Ground Truth | Elimination | Insufficient for Elimination | Inconclusive | Insufficient for Identification | Identification |
| Nonmated | 84 | 68 | 113 | 24 | 2 |
| Mated | 8 | 20 | 57 | 47 | 178 |
| | Strength of Support Scale | | | | |
| Ground Truth | Extremely Strong Support for Different Sources | Support for Different Sources | Inconclusive | Support for Common Source | Extremely Strong Support for Common Source |
| Nonmated | 45 | 119 | 110 | 13 | 2 |
| Mated | 11 | 19 | 61 | 84 | 136 |

Table 15. Response distribution for toolmark examiners.

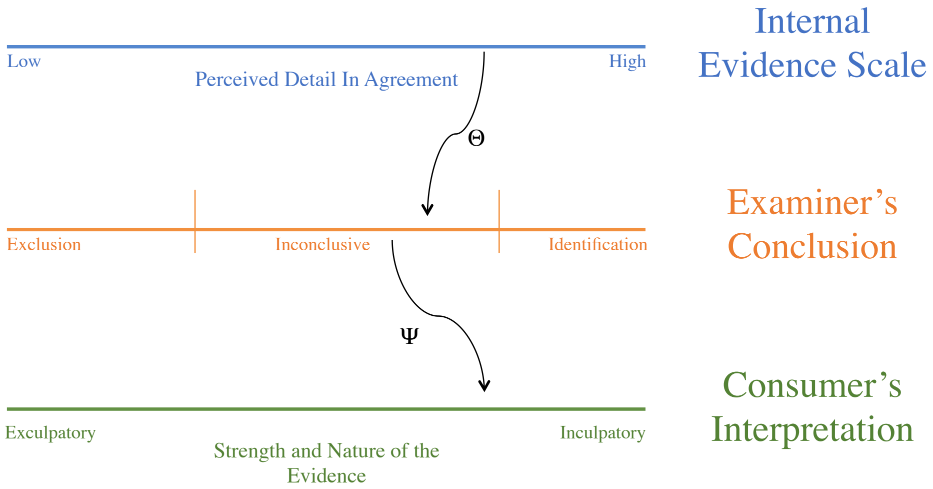


Figure 12. Evidence from a pattern comparison is accumulated internally by an examiner, which they then map to a conclusion scale using function Θ. This conclusion is then communicated to the consumer using articulation language, usually in the form of a set of verbal conclusions that may in some cases be supported by likelihood ratio models where available. The consumer (i.e. detective, prosecutor, defense attorney, judge, or juror) then interprets the conclusion statement, translating it into a separate Strength and Nature of the Evidence Scale using function Ψ. Both the Θ and Ψ  translations must be calibrated in order to accurately represent the true strength of the evidence.

This resource was prepared by the author(s) using Federal funds provided by the U.S.
Department of Justice. Opinions or points of view expressed are those of the author(s) and do not
necessarily reflect the official position or policies of the U.S. Department of Justice.

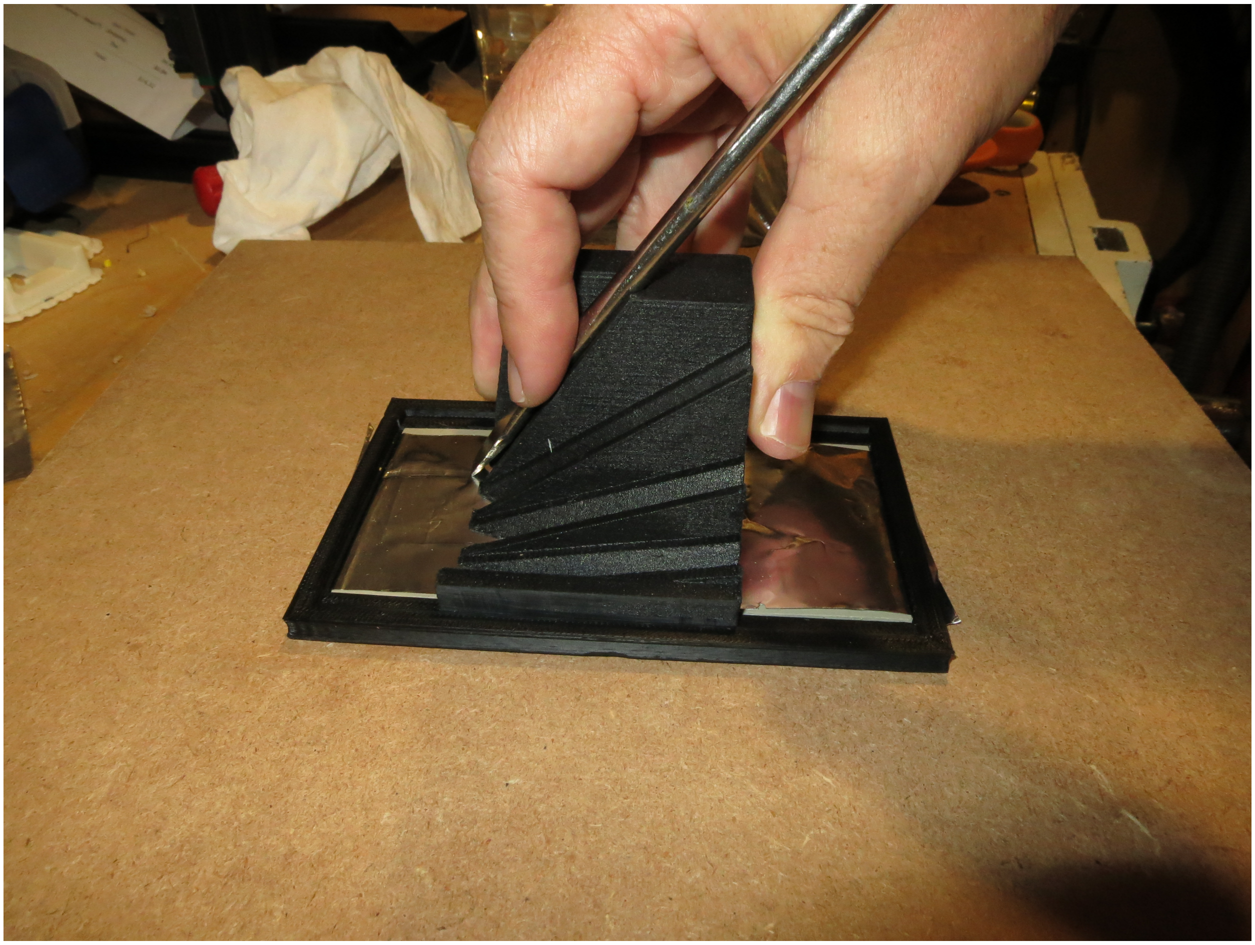


Figure 13. Custom 3D printed Jig for making toolmark impressions.

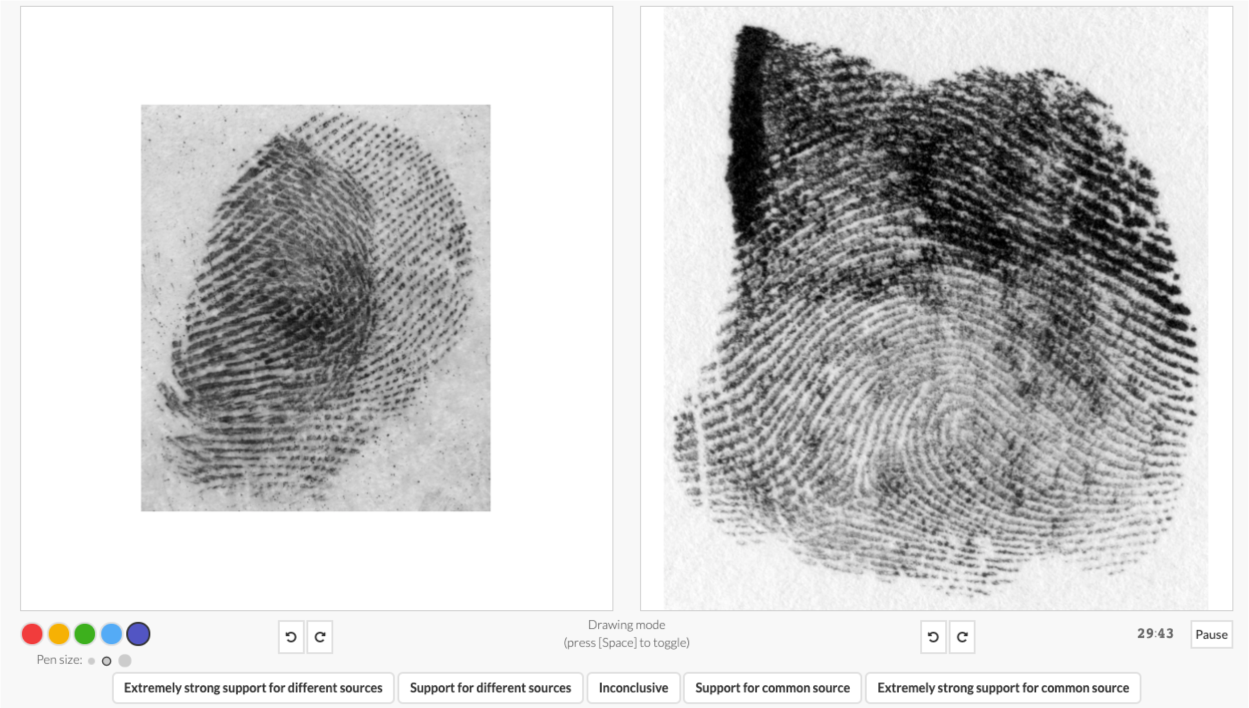
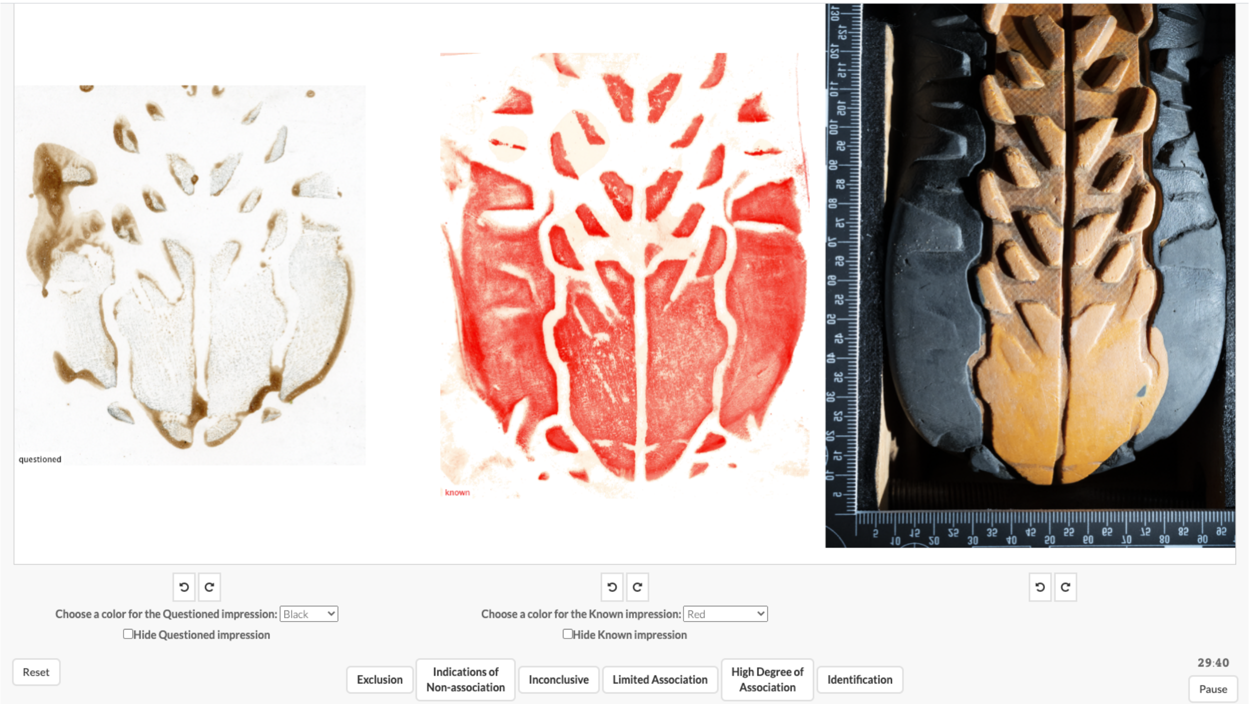

Drawing mode
(press [Space] to toggle)
Pen size:
29:43
Pause
Extremely strong support for different sources
Support for different sources
Inconclusive
Support for common source
Extremely strong support for common source

Figure 14. Interface used by fingerprint examiners to conduct casework-like comparisons.



Choose a color for the Questioned impression: Black

□Hide Questioned impression

Choose a color for the Known impression: Red

□Hide Known impression

Figure 15. Interface used by footwear examiners to conduct casework-like comparisons.

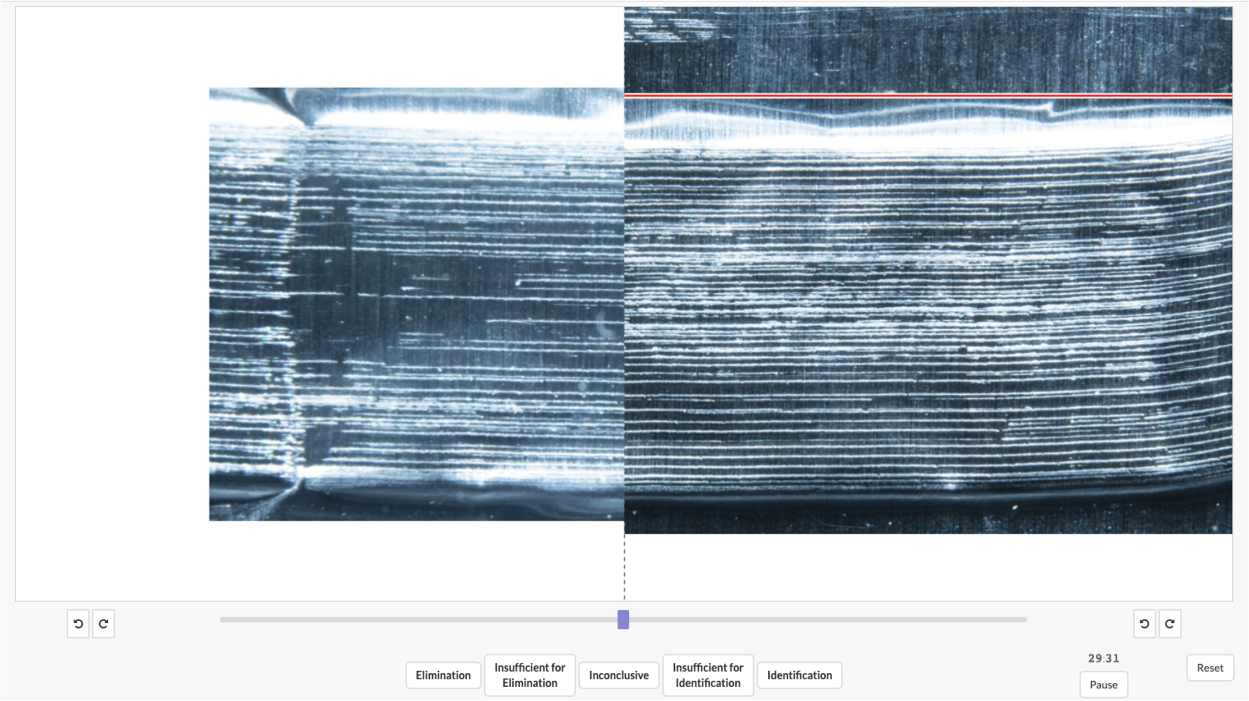


Figure 16. Interface used by toolmark examiners to conduct casework-like comparisons.

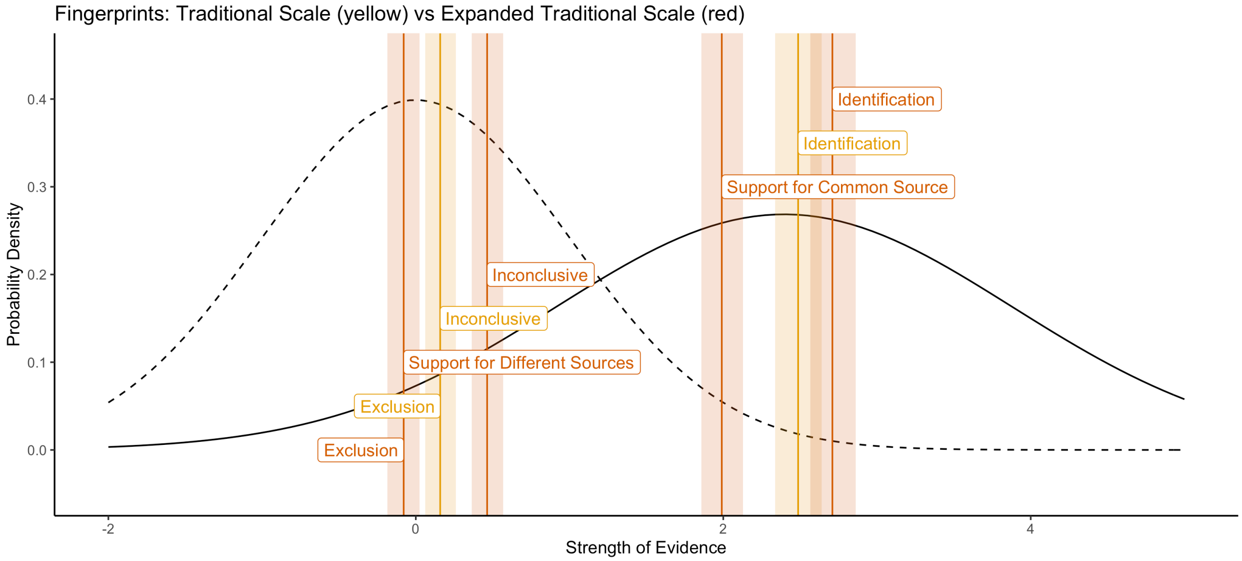


Figure 17. Estimates of the decision criteria for the comparison for Fingerprint examiners between the Traditional 3-item scale (Exclusion/Inconclusive/Identification) with the Expanded Traditional 5-item scale that included the Support For Different Sources and Support for Common Source categories. Color bands represent 95% confidence intervals. Note that the

Identification criterion shifts to the left for the expanded scale (red), indicating that examiners use this category less often than when they have only 3 categories to choose from. Examiners are also more risk-averse when making Exclusion conclusions when using the expanded scale.

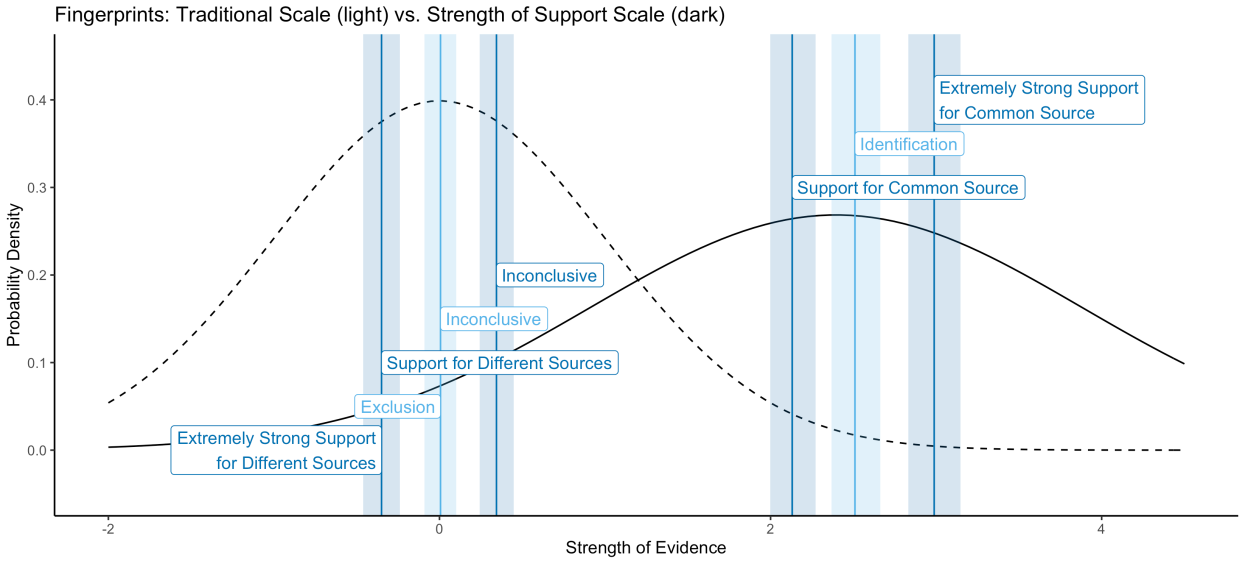


Figure 18. Estimates of the decision criteria for the comparison between the Traditional 3-item scale (Exclusion/Inconclusive/Identification) with the Strength of Support 5-item scale for Fingerprint examiners. Examiners become more risk-averse when using the expanded strength of support scale (see text for details).

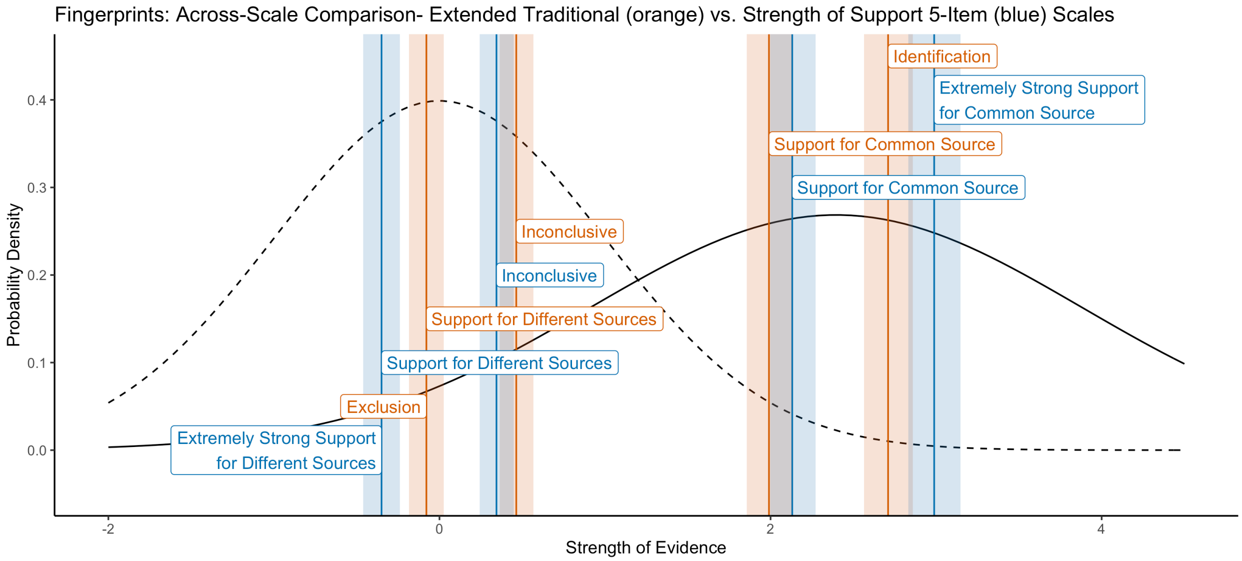


Figure 19. Across-scale comparison between the two 5-item scales for Fingerprint examiners. This comparison combines the data from both participant groups to estimate how each scale would be used if adopted for casework. Color bands represent 95% confidence intervals. The strength of support scale tends to make examiners more risk averse, because the  Extremely Strong Support for Common Source decision criterion is to the left of the Identification decision criteria. This results from the fact that examiners use Extremely Strong Support for Common Source less often than Identification.

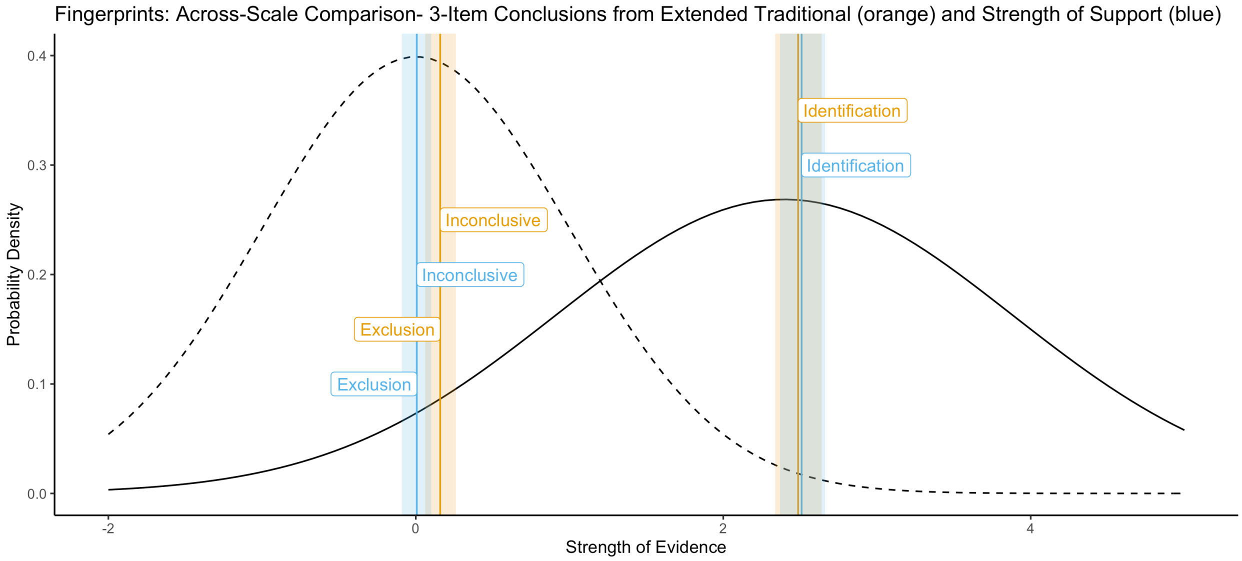


Figure 20. Comparison of estimates for decision criteria for the two 3-item scales for Fingerprint examiners. Color bands represent 95% confidence intervals. The two sets of participants had slightly different instructions for the 3-item scales (one provided explicit definitions, while the other asked them to use whatever criteria they would apply to a 3-item scale in casework). This graph illustrates that the estimates for the Identification criteria are almost identical, while there is slight variation in the Exclusion criteria.

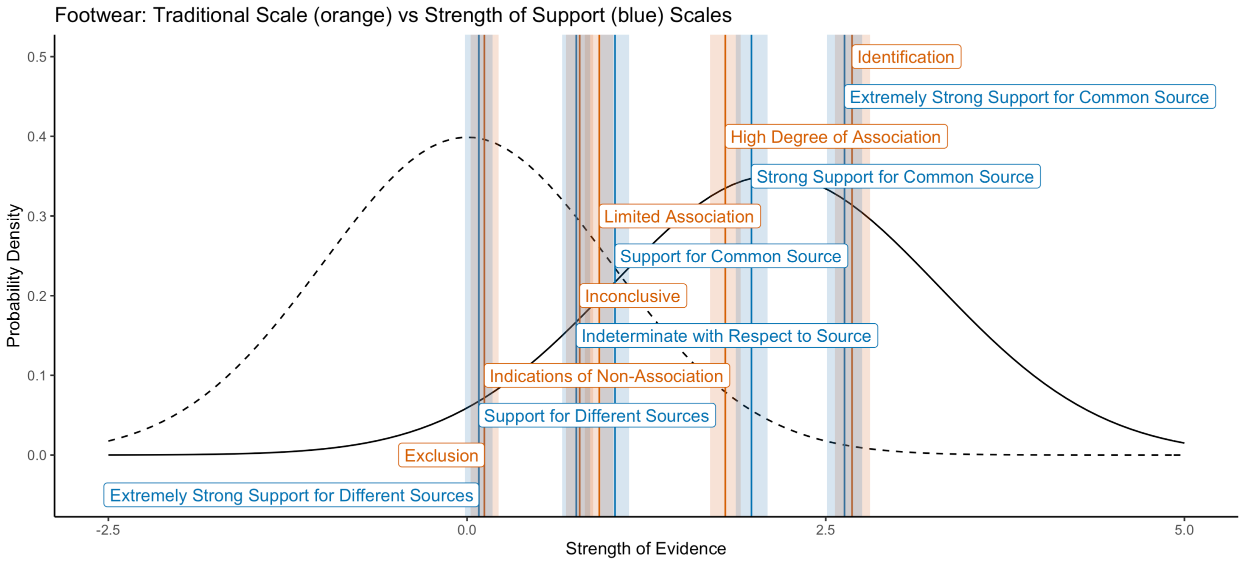


Figure 21. Estimates of the decision criteria for the comparison for the Traditional Scale and the Strength of Support Scale for Footwear examiners. Color bands represent 95% confidence intervals. The two scales appear to be used similarly, with perhaps a slight difference between High Degree of Association and Strong Support for Common Source.

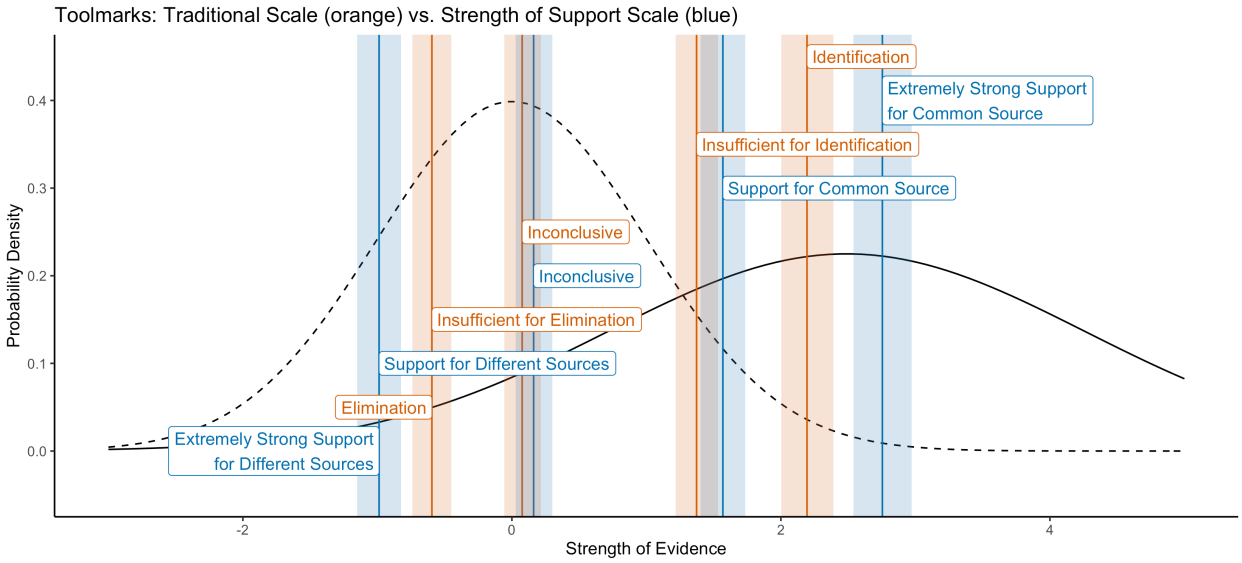


Figure 22. Estimates of the decision criteria for the Traditional Scale (orange) and  the Strength of Support Scale (blue) for Toolmark examiners. Color bands represent 95% confidence intervals. Examiners become more risk-averse when using the expanded strength of support scale (see text for details).

# References

Aitken, C., Berger, C. E. H., Buckleton, J. S., Champod, C., Curran, J., Dawid, A. P., . . . Zadora, G. (2011). Expressing evaluative opinions: A position statement. *Science & Justice, 51*(1), 1-2. doi:10.1016/j.scijus.2011.01.002

Assoc Forensic Sci Providers. (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice, 49*(3), 161-164. doi:10.1016/j.scijus.2009.07.004

Berger, C. E. H., Buckleton, J., Champod, C., Evett, I. W., & Jackson, G. (2011). Re: Expressing evaluative opinions; A position statement Response. *Science & Justice, 51*(4), 215-215. doi:10.1016/j.scijus.2011.09.006

Bürkner, P. (2017). Bayesian Regression Models using Stan. *R package version, 1*(0).

Busey, T., & Klutzke, M. (submitted). Calibrating the Perceived Strength of Evidence of Forensic Testimony Statements.

Busey, T., Klutzke, M., Nuzzi, A., & Vanderkolk, J. (submitted). Validating Expanded Conclusion Scales for Fingerprints, Toolmarks, and Footwear. *J Forensic Sci*.

Butler, J. M., & Butler, J. M. (2010). *Fundamentals of forensic DNA typing*. Amsterdam ; Boston: Academic Press/Elsevier.

Carter, K. E., Vogelsang, M. D., Vanderkolk, J., & Busey, T. (2020). The Utility of Expanded Conclusion Scales During Latent Print Examinations. *J Forensic Sci*. doi:10.1111/1556-4029.14298

Cohen, D. J., Ferrell, J. M., & Johnson, N. (2002). What very small numbers mean. *Journal of Experimental Psychology-General, 131*(3), 424-442. doi:10.1037//0096-3445.131.3.424

DFSC, D. F. S. C. (2017). Information paper: modification of latent print technical reports to include statistical calculations. Retrieved from https://osf.io/8kajs/download

Eldridge, H. (2019). Juror comprehension of forensic expert testimony: a literature review and gap analysis. *Forensic Science International: Synergy, 1*, 24-34.

Evett, I. W. (1998). Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice, 38*(3), 198-202. doi:Doi 10.1016/S1355-0306(98)72105-7

Friction Ridge Subcommittee, & OSAC. (2018). Standard for Friction Ridge Examination Conclusions. In.

Garrett, B., Crozier, W., & Grady, R. (2020). Error rates, likelihood ratios, and jury evaluation of forensic evidence. *Journal of Forensic Sciences, 65*(4), 1199-1209.

Garrett, B., & Mitchell, G. (2013). How jurors evaluate fingerprint evidence: The relative importance of match language, method information, and error acknowledgment. *Journal of Empirical Legal Studies, 10*(3), 484-511.

Garrett, B., Mitchell, G., & Scurich, N. (2018). Comparing categorical and probabilistic fingerprint evidence. *Journal of Forensic Sciences, 63*(6), 1712-1717.

Howes, L. M., Kirkbride, K. P., Kelty, S. F., Julian, R., & Kemp, N. (2013). Forensic scientists' conclusions: How readable are they for non-scientist report-users? *Forensic Science International, 231*(1-3), 102-112. doi:10.1016/j.forsciint.2013.04.026

IAI. (2010). IAI Resolution 2010-18. In (Vol. 2010): International Association for Identification.

Koehler, J. J., Schweitzer, N., Saks, M. J., & McQuiston, D. E. (2016). Science, technology, or the expert witness: What influences jurors' judgments about forensic science testimony? *Psychology, Public Policy, and Law, 22*(4), 401.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*: Psychology press.

Mannering, W. M., Vogelsang, M. D., Busey, T. A., & Mannering, F. L. (2021). Are forensic scientists too risk averse? *J Forensic Sci*. doi:10.1111/1556-4029.14700

Marquis, R., Biedermann, A., Cadola, L., Champod, C., Gueissaz, L., Massonnet, G., . . . Hicks, T. (2016). Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings. *Science & Justice, 56*(5), 364-370.

Martire, K. A., Kemp, R. I., & Newell, B. R. (2013). The psychology of interpreting expert evaluative opinions. *Australian Journal of Forensic Sciences, 45*(3), 305-314. doi:10.1080/00450618.2013.784361

Martire, K. A., Kemp, R. I., Sayle, M., & Newell, B. R. (2014). On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect. *Forensic Science International, 240*, 61-68. doi:10.1016/j.forsciint.2014.04.005

McQuiston-Surrett, D., & Saks, M. J. (2009). The testimony of forensic identification science: What expert witnesses say and what factfinders hear. *Law and human behavior, 33*(5), 436-453.

Morrison, G. S. (2012). The likelihood-ratio framework and forensic evidence in court: a response to R v T. *The international journal of evidence & proof, 16*(1), 1-29.

National Research Council of the National Academies of Science. (2009). *Strengthening Forensic Science in the United States: A Path Forward*. Washington DC: National Academies of Science.

Nordgaard, A., Ansell, R., Drotz, W., & Jaeger, L. (2012). Scale of conclusions for the value of evidence. *Law, Probability and Risk, 11*(1), 1-24. doi:10.1093/lpr/mgr020

PCAST. (2016). *Ensuring Scientific Validity of Feature-Comparison Methods*. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf

Robertson, B., Vignaux, G. A., & Berger, C. E. H. (2011). Extending the Confusion About Bayes. *Modern Law Review, 74*(3), 444-455. doi:10.1111/j.1468-2230.2011.00857.x

Spellman, B. A. (2017). Communicating forensic evidence: lessons from psychological science. *Seton Hall L. Rev., 48*, 827.

Swanson, C. L. (2020, November 20, 2020). [USACIL DFSC conclusion scale].

SWGFAST. (2013a). Document 19: Standard Terminology of Friction Ridge Examination (Latent/Tenprint). Version 4.1. In.

SWGFAST. (2013b). Document #10 Standards for Examining Friction Ridge Impressions and Resulting Conclusions (Latent/Tenprint). In. www.swgfast.org: Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST)

Swofford, H. J., & Cino, J. G. (2017). Lay Understanding of "Identification": How Jurors Interpret Forensic Identification Testimony *Journal of Forensic Identification, 68*(1), 29-41.

Team, R. C. (2013). R: A language and environment for statistical computing.

Thompson, W. C., Grady, R. H., Lai, E., & Stern, H. S. (2018). Perceived strength of forensic scientists' reporting statements about source conclusions. *Law, Probability and Risk, 17*(2), 133-155. doi:10.1093/lpr/mgy012

Thompson, W. C., Kaasa, S. O., & Peterson, T. (2013). Do jurors give appropriate weight to forensic identification evidence? *Journal of Empirical Legal Studies, 10*(2), 359-397.

Thompson, W. C., & 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.

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences of the United States of America, 108*(19), 7733-7738. doi:Doi 10.1073/Pnas.1018707108