Validating Strength of Support Conclusion Scales for Fingerprint, Footwear, and Toolmark Impressions

Thomas Busey, Indiana University
Morgan Klutzke, Indiana University
Alyssa Nuzzi, Indiana University
John Vanderkolk, Indiana State Police Fort Wayne Regional Laboratory, Retired

Corresponding Author: Tom Busey
Department of Psychological and Brain Sciences
1101 E 10th St.
Indiana University, Bloomington, IN 47408
busey@indiana.edu

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

Abstract

In the pattern comparison disciplines such as fingerprints, footwear, and toolmarks, 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 requiring knowledge of prior probabilities by the examiners and being overinterpreted by laypersons. Alternative statements based on strength of support language have been proposed as replacements. The current study compares traditional conclusion scales against strength of support 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 support conclusion scale. The scale used on each trial was randomly assigned, and participants knew the scale for that trial as they began the comparison. 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 support scales similarly, but toolmark examiners were much less likely to use Extremely Strong Support for Common Source than Identification, similar to fingerprint examiners. The study also compared the traditional conclusion scaled to an expanded scale in fingerprint examiners, and found that fingerprint examiners used Identification less often when an expanded scale was available. This final result demonstrates that the meaning of a term depends in part on the other items available in the scale.

Keywords: strength of support; categorical conclusions; articulation language; friction ridge impressions; footwear; toolmarks

Practitioners working in the pattern comparison disciplines in the US traditionally express their conclusions using an articulation statement drawn from a small set of categorical conclusions. Quantitative tools for assessing the strength of support 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. Footwear and tool mark 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).

Categorical conclusions have been criticized on several fronts, most notably because they represent a statement about a posterior probability rather than a statement about the strength of the support for different propositions. Jackson et al. (Jackson, Kaye, Neumann, Ranadive, & Reyna, 2015) argued for strength of support approaches such as a likelihood ratio in all forensic disciplines, because such an approach does not presume a prior probability and allows the trier of fact to update their beliefs in a transparent way. In addition, categorical statements such as Identification are prone to overinterpretation by laypersons. 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, 2013). 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.

Although pattern disciplines are just beginning to add quantitative tools that might provide a numerical estimate of the strength of the evidence, policymakers have begun to explore strength-of-support type language to avoid posterior conclusions. The present work explores the consequences of adopting this new language, and a companion work (Busey & Klutzke, submitted) measures how laypersons and forensic examiners interpret different articulation statements to determine whether the intent of examiners is understood by laypersons.

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, Kemp, Sayle, & Newell, 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.

In the fingerprint discipline, one potential solution to the issues raised above is to keep the traditional conclusion language, but expand the scale. Prior work on this question by our group revealed a surprising result: the interpretation of the term Identification by fingerprint examiners depended on the scale in which it was embedded (Carter, Vogelsang, Vanderkolk, & Busey, 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 an extension and generalization of this approach, including both a direct replication of the prior work with more realistic time deadlines, as well as comparisons between a traditional scale and a pure strength of support conclusion scale in fingerprint, footwear, and toolmark disciplines. Strength of support 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, Hicklin, Buscaglia, & Roberts, 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 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 support scales, to measure how fingerprint, footwear, and tool mark 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). Together

the combined work allows for a complete calibration of the language such that it accurately reflects the intent of the examiner and is interpreted properly by the layperson.

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 support 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 support 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 1), 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 1 and Table 2 for fingerprint examiners, Table 3 for Footwear examiners, and Table 4 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 2). 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 2. 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. Participants indicated their decision using buttons at the bottom of the screen. If the timer ended before they made a decision, then the prints were hidden, and they were asked to make a decision before continuing to the next trial. 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 1 or Table 2 depending on the comparison they were randomly assigned, and did not have extensive training on the new categories in the expanded or strength of support 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 1, which examiners have had experience with and presumably should not change, although this is an empirical question. Participants in the traditional and trength of support comparison condition were provided the definitions of the terms shown in Table 2.

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 1. 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 8 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 5. 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 6. The interface included a split screen controlled by a slider. The questioned image (on the left in Figure 4) could be dragged around, as could the known impressions on the right in Figure 4. The two known impressions were clearly separated with a red line to differentiate each. 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 7 through Table 10 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 7 illustrates the response distribution for participants who compared the traditional scale with the expanded traditional scale. Table 8 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 7), and more pronounced in the Traditional/Strength of Support comparison (250 to 180 in Table 8). These results are again consistent with the finding that examiners redefine the definition of the term Identification when the scale is expanded (see Table 7) or are less likely to use Extremely Strong Support for Common Source than the term Identification (see Table 8).

The number of Inconclusive responses to mated pairs also dropped as the traditional scale was expanded. These dropped from 220 to 131 in Table 7, and from 244 to 179 in Table 8. 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 7 and 265 to 195 in Table 8. This suggests that examiners become risk averse for the expanded scales on the exclusion side.

Footwear Examiners

Table 9 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 9 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 10 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 7 through Table 10 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 5 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 7, 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 6. 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 7 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 8. 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 7.

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 9 demonstrated that there were no large shifts in responses across the two scales. Consistent with this result, Figure 9 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 10 illustrated that toolmark examiners grew more risk-averse when using the Strength of Support scale. As shown in Figure 10, 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 10). 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 1. 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 2. 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 3. Definitive conclusion and strength-of-support 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 4. Definitive conclusions and strength of support 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 5. 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 6. Instructions given to Toolmark examiners.

	Traditional Scale				
Ground Truth	Exclusion		Inconclusive		Identification
Nonmated	272	*NA*	207	*NA*	2
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 7. 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 8. 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	Indetermi-nate 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 9. 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 10. Response distribution for toolmark examiners.

Figure 1. Custom 3D printed Jig for making toolmark impressions.

Figure 2. Interface used by fingerprint examiners to conduct casework-like comparisons.

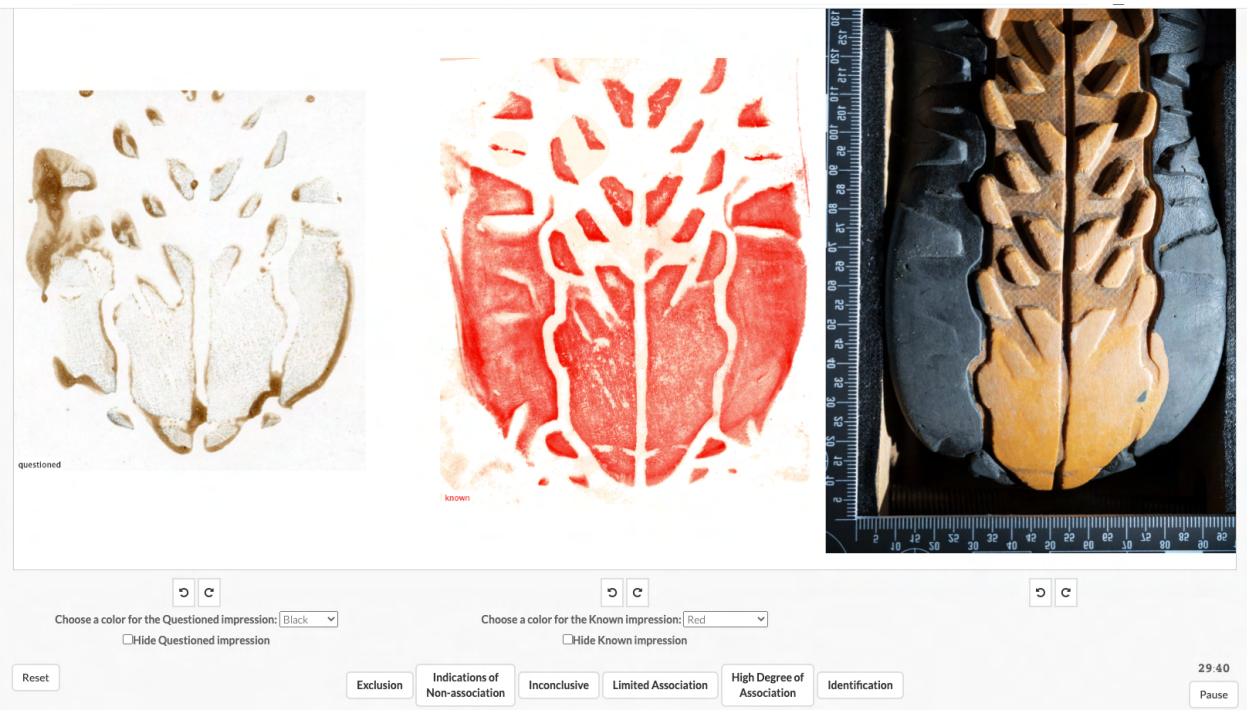

Figure 3. Interface used by footwear examiners to conduct casework-like comparisons.

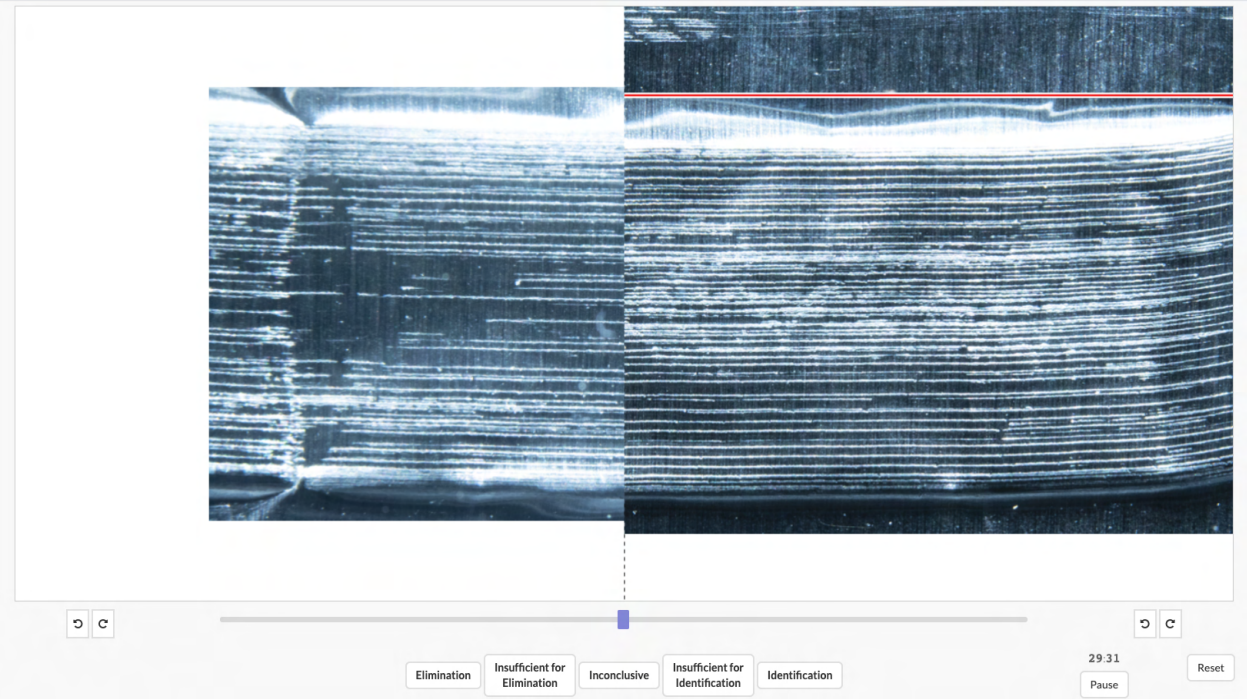

Figure 4. Interface used by toolmark examiners to conduct casework-like comparisons.

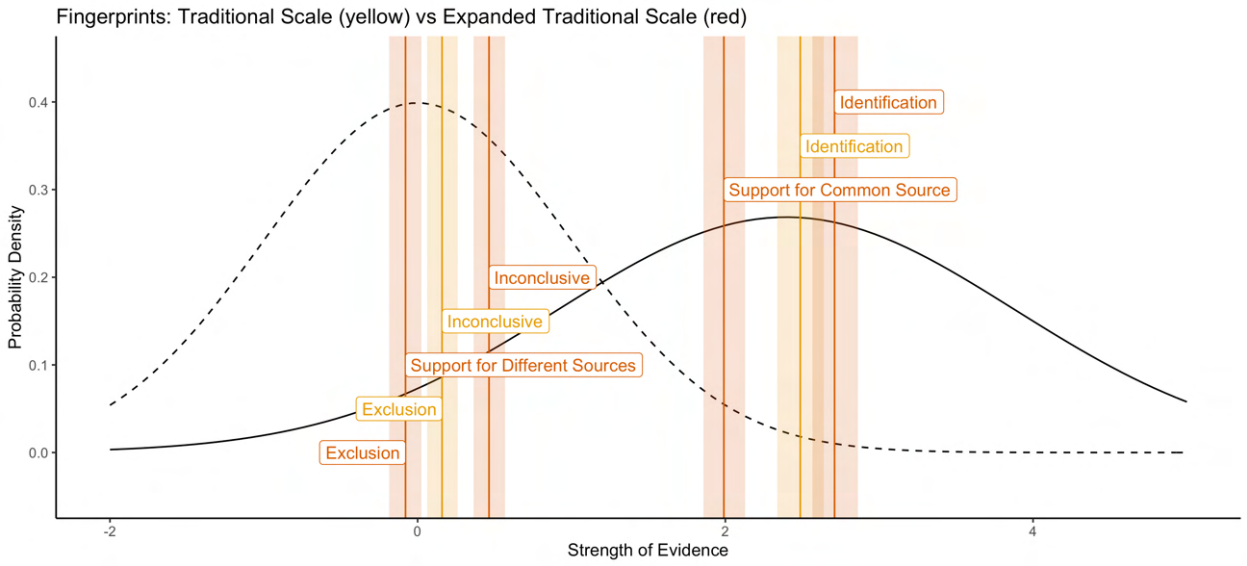

Figure 5. Estimates of the decision criteria for the comparison for Fingerprint examiners between the Traditional 3-item scale (Exclusion/Inconclusive/Identification) with the Expanded Traditional 5-item scale that included the Support For Different Sources and Support for Common Source categories. Color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

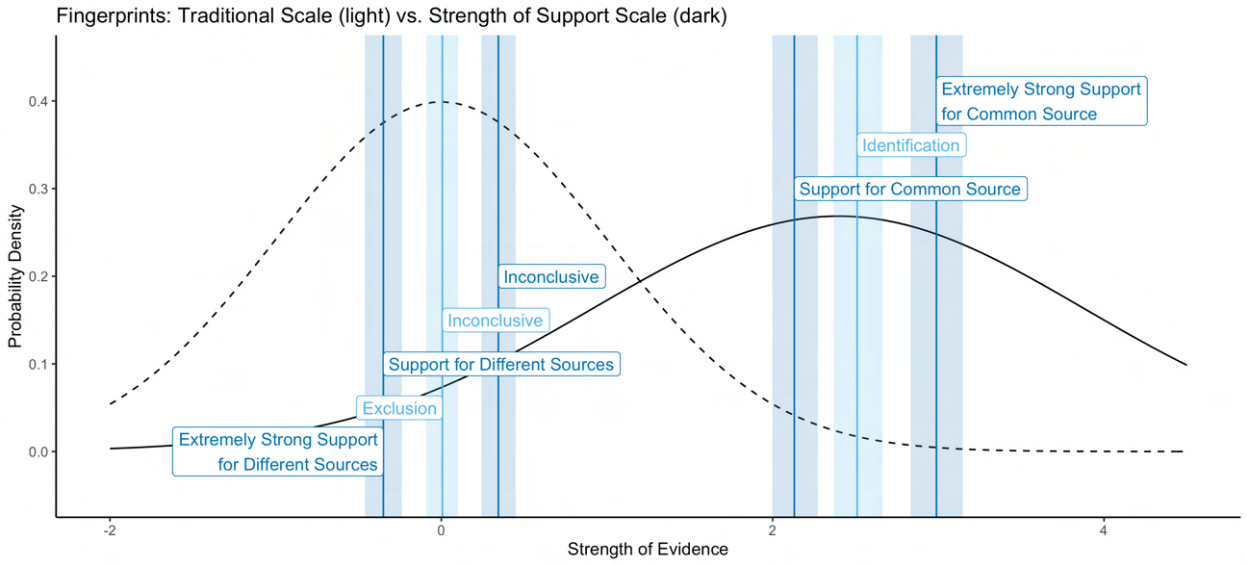

Figure 6. Estimates of the decision criteria for the comparison between the Traditional 3-item scale (Exclusion/Inconclusive/Identification) with the Strength of Support 5-item scale for Fingerprint examiners. Examiners become more risk-averse when using the expanded strength of support scale (see text for details).

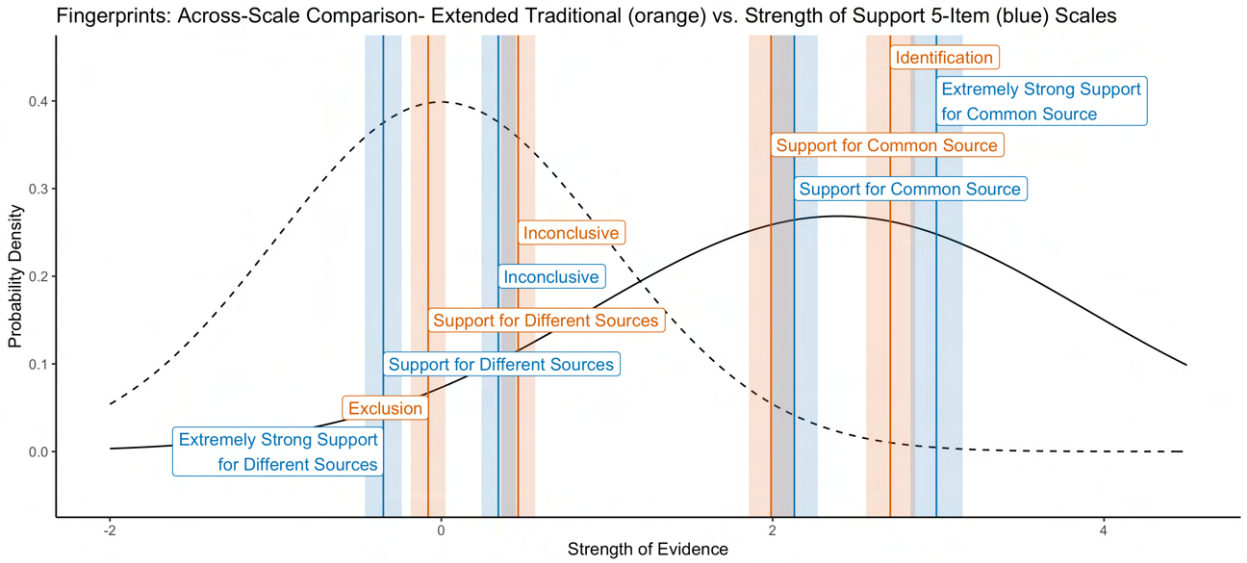

Figure 7. Across-scale comparison between the two 5-item scales for Fingerprint examiners. This comparison combines the data from both participant groups to estimate how each scale would be used if adopted for casework. Color bands represent 95% confidence intervals. The strength of support scale tends 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.

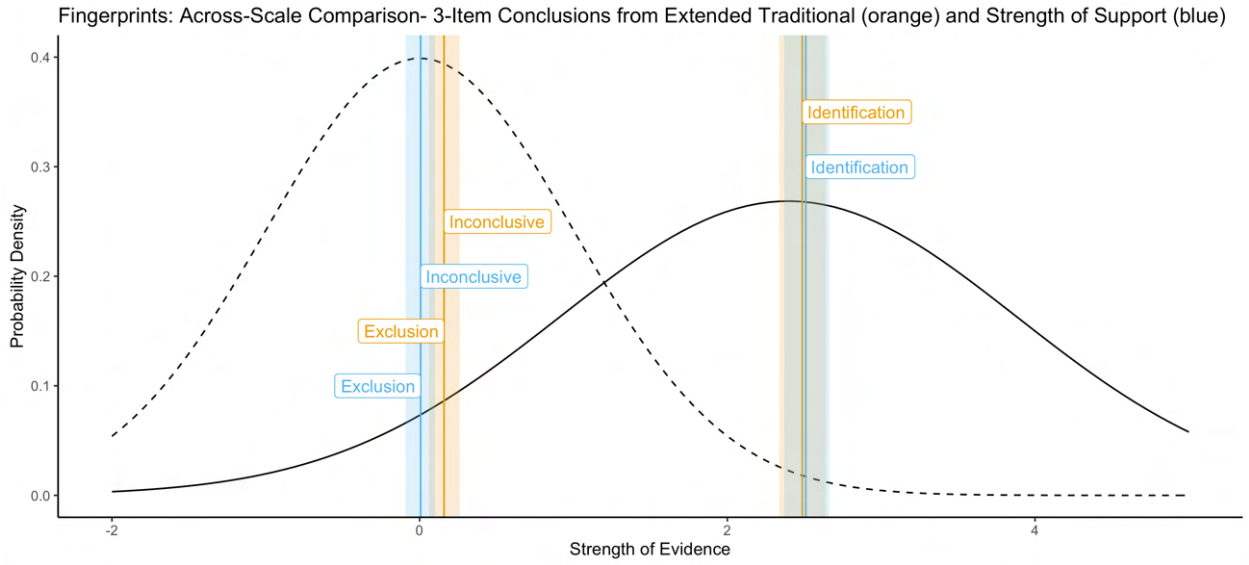

Figure 8. Comparison of estimates for decision criteria for the two 3-item scales for Fingerprint examiners. Color bands represent 95% confidence intervals. The two sets of participants had slightly different instructions for the 3-item scales (one provided explicit definitions, while the other asked 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.

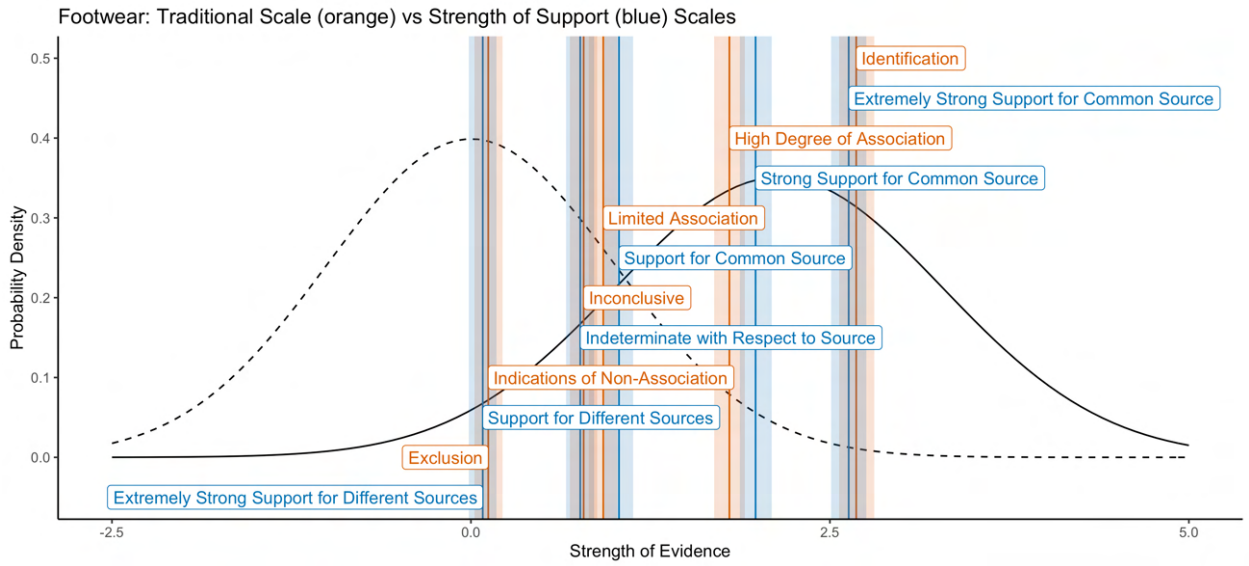

Figure 9. Estimates of the decision criteria for the comparison for the Traditional Scale and the Strength of Support Scale for Footwear examiners. Color bands represent 95% confidence intervals. The two scales appear to be used similarly, with perhaps a slight difference between High Degree of Association and Strong Support for Common Source.

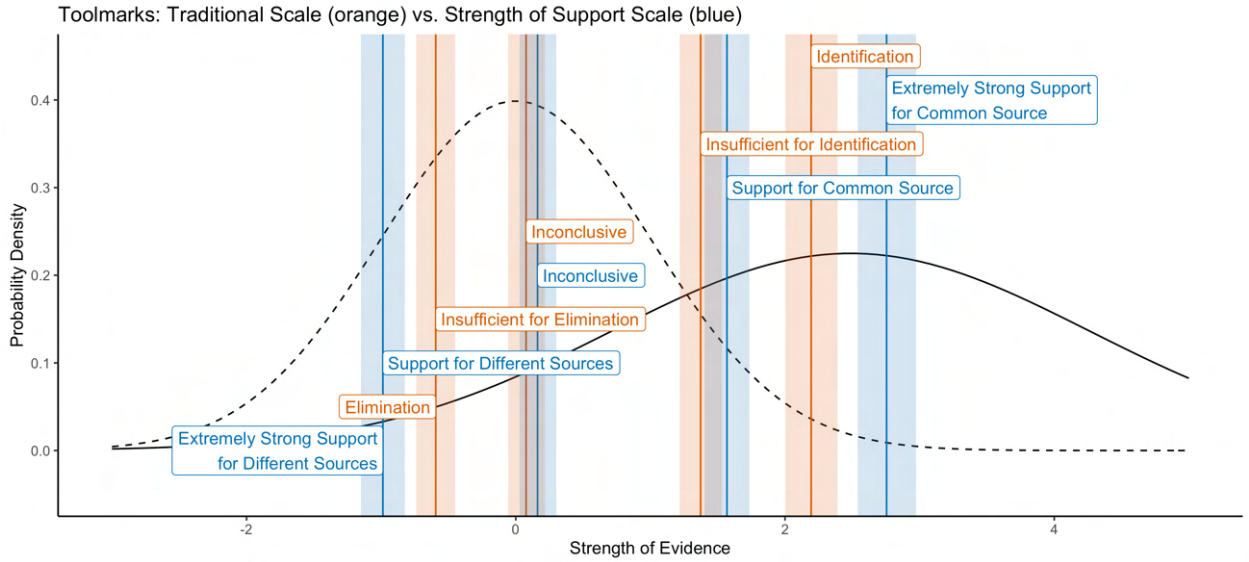

Figure 10. Estimates of the decision criteria for the Traditional Scale (orange) and the Strength of Support Scale (blue) for Toolmark examiners. Color bands represent 95% confidence intervals. Examiners become more risk-averse when using the expanded strength of support scale (see text for details).

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

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.

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

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.

Jackson, G., Kaye, D. H., Neumann, C., Ranadive, A., & Reyna, V. F. (2015). Communicating the results of forensic science examinations.

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

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

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

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

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., & 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