JOURNAL OF FORENSIC SCIENCES

J Forensic S
doi: 10.1111/1556-4029.14298
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Kelly E. Carter,[1] *B.A.; Macgregor D. Vogelsang,*[1] *B.S.; John Vanderkolk,*[2] *B.A.; and Thomas Busey,*[1] *Ph.D.*

The Utility of Expanded Conclusion Scales During Latent Print Examinations

ABSTRACT: During fingerprint comparisons, a latent print examiner visually compares two impressions to determine whether or not they originated from the same source. They consider the amount of perceived detail in agreement or disagreement and accumulate evidence toward same source and different sources propositions. This evidence is then mapped to one of three conclusions: Identification, Inconclusive, or Exclusion. A limitation of this 3-conclusion scale is it can lose information when translating the conclusion from the internal strength-of-evidence value to one of only three possible conclusions. An alternative scale with two additional values, support for different sources and support for common sources, has been proposed by the Friction Ridge Subcommittee of OSAC. The expanded scale could lead to more investigative leads but could produce complex trade-offs in both correct and erroneous identifications. The aim of the present study was to determine the consequences of a shift to expanded conclusion scales in latent print comparisons. Latent print examiners each completed 60 comparisons using one of the two scales, and the resulting data were modeled using signal detection theory to measure whether the expanded scale changed the threshold for an "Identification" conclusion. When using the expanded scale, examiners became more risk-averse when making "Identification" decisions and tended to transition both the weaker Identification and stronger Inconclusive responses to the "Support for Common Source" statement. The results demonstrate the utility of an expanded conclusion scale and also provide guidance for the adoption of these or similar scales.

KEYWORDS: decision making, expanded conclusions, fingerprints, friction ridge, model comparison, identification

Fingerprint comparisons are conducted by human examiners rather than by computers, although computer database searches can provide candidate impressions for comparison. In the United States, there are no fixed standards for how much evidence is sufficient to determine that two impressions came from a common source. Instead, human examiners conduct manual examinations, which compare a latent impression from a crime scene against one or more exemplar impressions taken from a suspect, victim, or a computer database (1). First, the examiner decides whether the latent print has enough detail to make a decision about its origin. If they believe the print is "of value," they compare the latent print to a print from a known origin. They then deliver an opinion about whether the two impressions share a common source, which takes the form of a written statement communicated to a consumer such as a detective or prosecutor.

Fingerprint examiners have neither a statute for sufficiency nor a prescribed method from the courts for how to conduct a comparison. In addition to lacking a fixed standard like a minimum number of minutia, examiners do not have a fixed set of features they must rely on or guidance on how to interpret these features, although feature descriptions do exist (2), actual features may lie below the level of conscious awareness (3). In addition, fingerprint examinations must (if examiners are trained properly) also consider the alternative hypothesis that some other person touched this surface. Because typically not all other persons can be measured, this is an *inductive* process, where we must infer the likelihood of some other person having observed only a subset of all other people.

In our view, this lack of specificity by policymakers is not fatal to the use of fingerprint evidence, although it does place an increased burden on the examiner to correctly communicate the results of their examination. The standards of evidence probably emerge through a consensus process within the community, through proficiency tests, and through conflict resolution/verification during the comparison process. Black box testing has revealed a fairly low rate of erroneous identifications (~0.1%), a moderately high erroneous exclusion rate (7.5%), and a fairly high inconclusive rate for mated pairs (~31%) (4). Based on this error rate study and others, the general consensus is that fingerprint examiners appear to contribute information to the court that rises above the level of junk science (5). However, the fingerprint comparison task is much more opaque than the gun barrel measurement task despite recent efforts at quantification (6) and much rests on the nature of the communication with the consumer such as detective or jury.

Central to the comparison process is the fact that much of the evidence accumulates within the mind of the examiner and must be accurately conveyed to a consumer such as a detective, prosecutor, or jury. This communication currently takes the form of one of three conclusions: Identification, Exclusion, or Inconclusive. 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 validated, detectives, defense attorneys,

[1]Indiana University, 107 S Indiana Ave, Bloomington, IN 47405.
[2]Indiana State Police, 5811 Ellison Road, Fort Wayne, IN 46804.
Corresponding author: Thomas Busey, Ph.D. E-mail: busey@indiana.edu

Received 18 Sept. 2019; and in revised form 2 Dec. 2019, 21 Jan. 2020; accepted 24 Jan. 2020.

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. This mapping of evidence to decisions and the subsequent understanding by the consumer has been addressed using a variety of approaches, including measuring the utility of different outcomes using proposed gambling paradigms (7) and direct comparison between different statements (8–10).

To illustrate how errors in calibration can occur when reporting evidence, Fig. 1 depicts the flow of information during a forensic comparison. In Fig. 1, the evidence analyzed in the pattern disciplines accumulates in an examiner's working memory during the comparison (top row). This internal evidence may be thought of as on a *strength-of-evidence* scale and ranges between two propositions: The two impressions came from a common source, and the two impressions came from different sources. This final strength-of-evidence value is then translated to a conclusion through the translation function Θ (middle row). In the friction ridge discipline, this involves Exclusion, Inconclusive, and Identification conclusions.

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 (11), 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 such a decision.

An additional source of miscalibration between the evidence and its use by the justice system can occur during the mapping Ψ 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. However, recent work by Swofford and Cino (12) 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 say and how their conclusions are interpreted. In this case, jurors interpret the evidence as stronger than was originally intended by the examiner.

The use of the phrases “Identification” and “Exclusion” represents an important difference from using likelihood ratios as means of expressing the strength of evidence. Strength-of-evidence approaches have been advocated by a wide scientific consensus (see [13] for a broad treatment of the integration of likelihood ratios into the forensic workflow). There are actually two important differences between categorical conclusions and strength-of-support statements such as likelihood ratios. First, likelihood ratios are typically on a continuous scale, whereas categorical conclusions typically rely on a small number of statements. Second, categorical conclusions are *conclusions* and therefore represent a posterior in Bayesian terminology. As such, they implicitly include a prior in the calculation, although this

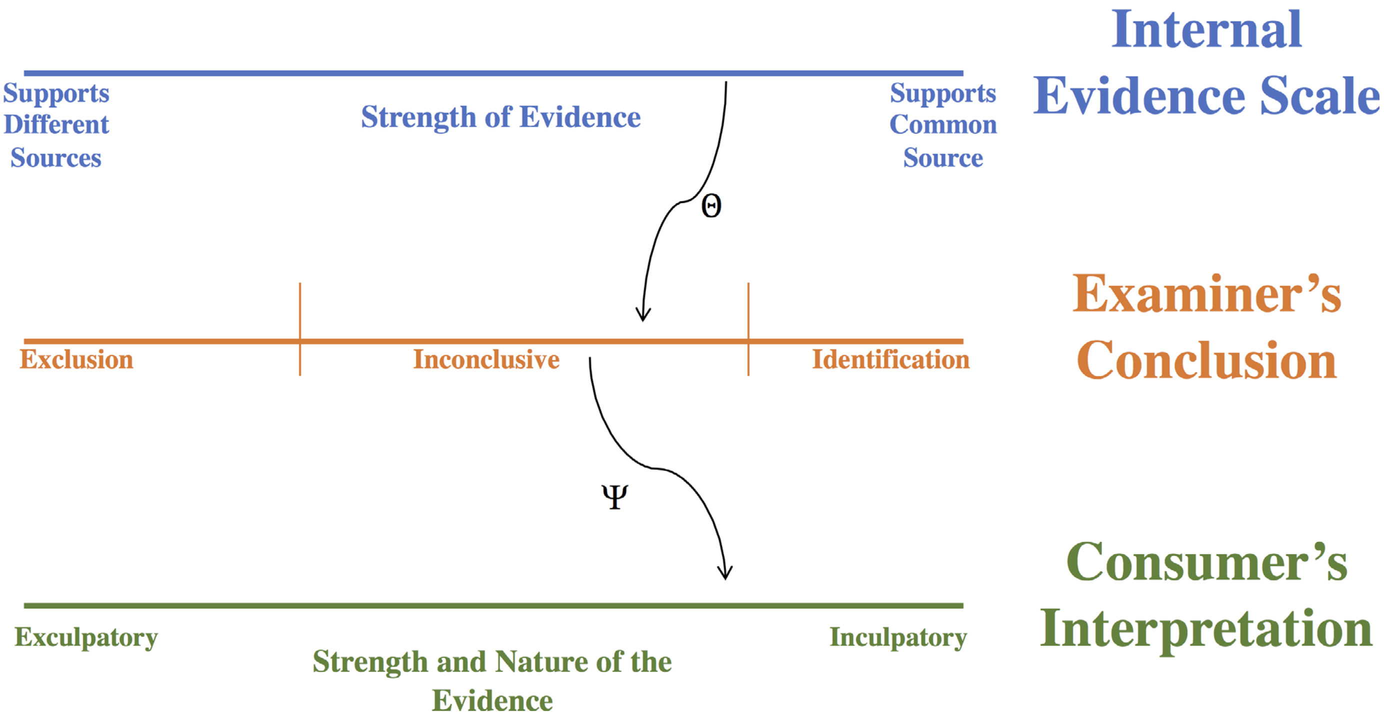

FIG. 1—*Evidence from a pattern comparison is accumulated internally by an examiner, which they then map to a conclusion scale using function Θ. This conclusion is then communicated to the consumer using articulation language, usually in the form of a set of verbal conclusions that may in some cases be supported by likelihood ratio models where available. The consumer (i.e., detective, prosecutor, defense attorney, 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.*

prior is rarely stated or even acknowledged by practitioners in the United States. The threshold placement implicitly includes a judgment of the values of different outcomes, which can be estimated but is rarely acknowledged by examiners (7). In Europe, the European Network of Forensic Science Institutes (ENFSI) has supported the use of likelihood ratio statements of the strength of the evidence even in instances where numerical values are not available (14,15), and some US-based laboratories have moved in the direction of calculating likelihood ratios (6). We return to the differences between the two approaches in the Discussion, although the recent PCAST report (16) expressed support for the continued use of categorical conclusions supported by error rate studies as an alternative to likelihood ratio approaches.

The goal of the present study was to determine whether the fingerprint comparison process would benefit from an expanded conclusion scale. This might reduce the information lost near the boundaries between conclusions (e.g., a detective might want to know whether a comparison was inconclusive, but the two impressions had strong similarities). A possible scale is shown in Fig. 2, which illustrates a strength of support continuum, which is subdivided by the traditional 3-conclusion scale on the top and then expanded to include the conclusions of Support for Different Sources and Support for Common Source as shown on the lower set of conclusions.

In principle, scales with more conclusions preserve more information from the underlying psychological dimension upon which they are based. Cicchetti, Showalter (17) simulated how the number of conclusions on a rating scale might affect inter-rater reliability. They demonstrated that inter-rater reliability was lowest for scales with only two conclusions and highest for 100 conclusions, but reached an asymptote at about seven conclusions. Later work by Lozano, García-Cueto (18) addressed both the reliability and validity of rating scales of different response category sizes. They varied the number of response conclusions from 2 to 9. Their simulations demonstrated that as the number of response conclusions increased, both the reliability and validity increased, with an optimal number ranging from 4 to 7. Below 4 and the validity and reliability suffer, and above 7 produces no measurable improvement, much like that found by Cicchetti, Showalter (17). Thus, expanded scales appear to have a statistical advantage, and even larger scales have been used: Within the domain of memory research, Ref. (19) used a 99-point scale as means to directly assess the assumptions of signal detection theory (20).

However, there are several factors that suggest that expanded scales are not always preferred:

First, on a theoretical level, additional conclusions require the maintenance of more category boundaries (or criteria in signal detection models). If these drift over time, this has the consequence of introducing noise into the measurement system, ultimately reducing performance. Benjamin, Tullis (21) tested this proposition in memory work and found small but significant drops in the area under the receiver operating characteristic (ROC) curve. Thus, one goal of the present study was to test for drops in performance such as reduced sensitivity values (as measured by d’ using the models in signal detection theory (20)) when using expanded scales.

Second, and most importantly, the expanded scale may *change the definition of identification*. This would have serious implications for the criminal justice system, because a suspect might be prosecuted using a 3-conclusion scale but might not be prosecuted if a 5-conclusion scale is used. Although such inter-jurisdictional differences likely already exist, it is important to anticipate the effects that changes in policy may have for an individual agency. An expanded conclusion scale may also affect the exclusion threshold, although that decision threshold may have fewer consequences because the exculpatory value of an exclusion decision may depend more on the facts of the case and may not have much inculpatory value.

The current study addresses both of these concerns, using casework-like comparisons and working latent print examiners. We explored an expanded conclusions scale that has been

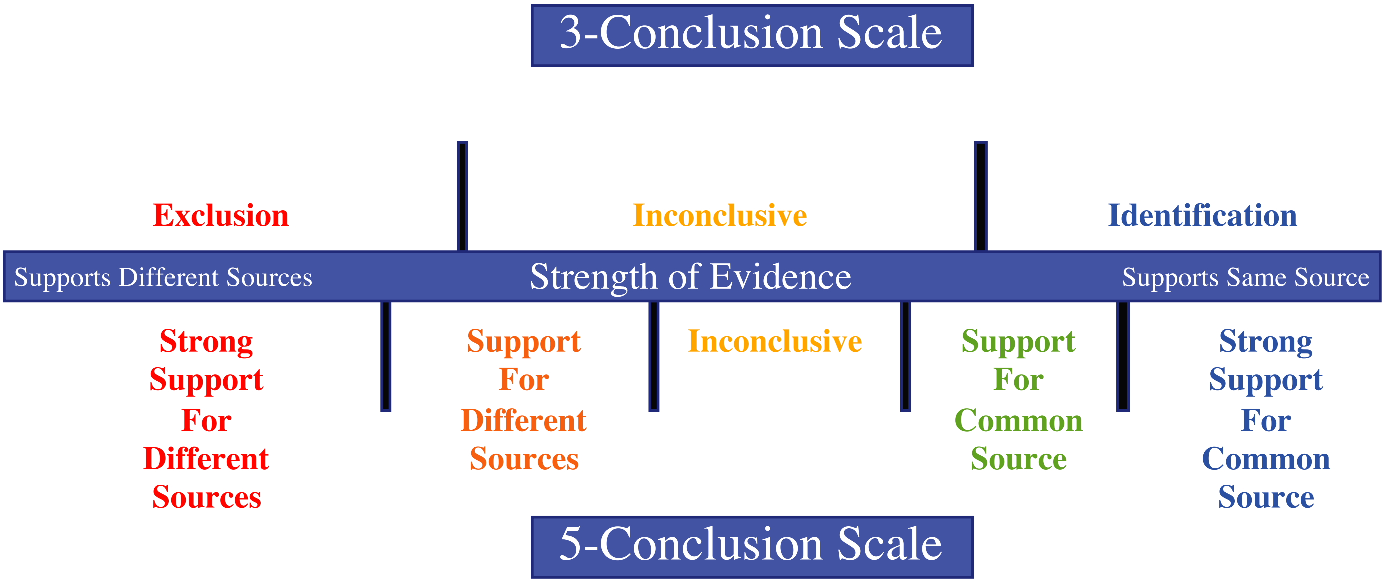
3-Conclusion Scale
Exclusion
Inconclusive
Identification
Supports Different Sources
Strength of Evidence
Supports Same Source
Strong Support For Different Sources
Support For Different Sources
Inconclusive
Support For Common Source
Strong Support For Common Source
5-Conclusion Scale

FIG. 2—*Conceptual comparison of a 3-conclusion scale with a 5-conclusion scale, and how the different thresholds might be distributed across the internal evidence axis (amount of perceived detail in agreement). In this hypothetical case, the qualified conclusions of “Support for Different Sources” and “Support for Common Source” are not simply a subdivision of the Inconclusive category as in Figure 4 nor do they simply qualify the definitive conclusions as in Figure 5, but instead capture some of the weaker definitive conclusions (Exclusion and Identification) while also capturing some of the previous Inconclusive responses.*

proposed as a draft standard by the Friction Ridge Subcommittee of the Organization of Scientific Area Committees (OSAC)(1). To assess how examiners might use an expanded scale, we used stimuli and displays as shown in Fig. 3. Examiners knew in advance on each trial whether they would use the traditional 3-conclusion scale or the expanded 5-conclusion scale, illustrated in Table 1. Each participant completed 60 trials, which included a mixture of mated (i.e., same source) and nonmated (i.e., different sources) trials. Because we collected the impressions ourselves, we knew the ground truth for each comparison. No feedback was given at the end of each trial.

The expanded conclusion scale contained in Ref. (1) and shown in Table 1 is technically a mixture of strength-of-support statements and categorical conclusions. This is perhaps inelegant, because to make the strongest form of conclusion the examiner must consider the prior, which is arguably in the domain of the jury (13), while the "support for" statements merely express the strength of the evidence. This may lead to strange situations: consider a smudged fingerprint on the Space Station. Without considering the priors, the strength of evidence may only merit a "support for common source" statement. However, when considering the priors (perhaps only a dozen astronauts could have touched that surface), an examiner may feel comfortable making an "identification" conclusion (see [22] for a more complete argument). When constructing the expanded scale, the support for statements were likely seen by the draft standard authors as qualified statements whose inclusion into an expanded scale may have the ultimate consequence of moving the entire US forensic community to a strength-of-support approach, as opposed to phrases such as "Almost Identification" or "Tending Identification." Thus, we chose to test the proposed draft standard language as one way to assess the consequences of adopting this approach rather than using a scale that was entirely categorical or entirely strength-of-support, and we return to full strength-of-support scales in the Discussion.

Possible outcomes for the experiment are shown in Fig. 4, which illustrates an outcome where the "support for" conclusions subdivide the inconclusive category; in Fig. 5 where the "support for" conclusions provide more information about weak definitive conclusions; and previously in Fig. 2 where the data support a mixture of these two outcomes. We use variants of signal detection theory to test these alternative accounts of how an expanded set of conclusions might affect the definition of the definitive conclusions. In addition, the signal detection theory models will test whether the 5-conclusion scale reduces d', which would suggest that expanded scales might reduce overall performance in a way that might not be desirable. One advantage of using the expanded conclusion scale in Table 1 is that by including the "Identification" and "Exclusion" conclusions as part of the expanded scale, we can directly assess how the inclusion of additional categories in the scale affects the interpretation of the definitive conclusions.

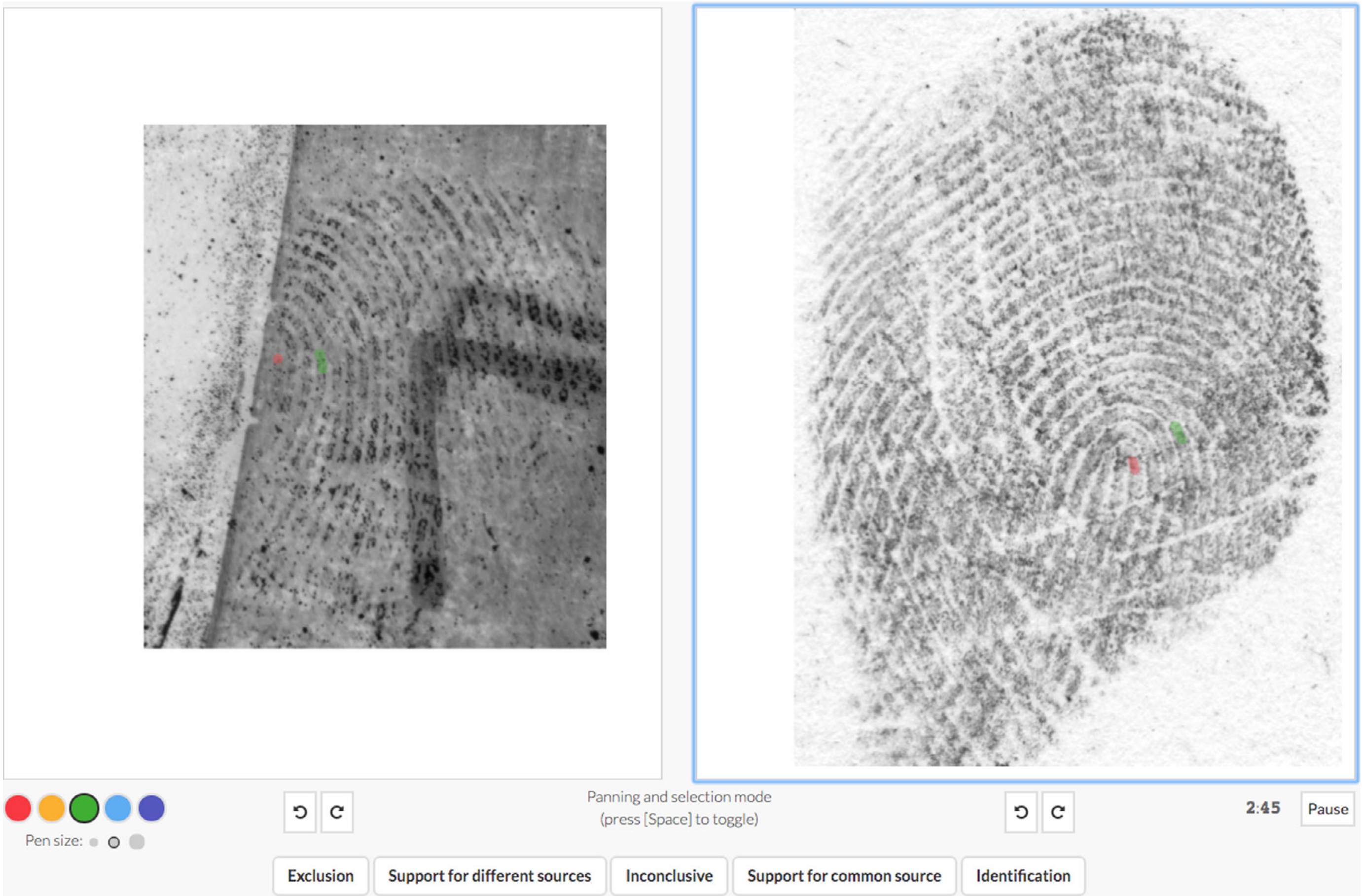
Panning and selection mode
(press [Space] to toggle)

FIG. 3—*Example of the online interface that allows an examiner to compare fingerprints, along with the scale used on this trial. Participants conducted a latent print comparison and made their response by clicking on one of the buttons at the bottom of the screen. The interface allows participants to zoom on each image, pan, rotate, and add markers to each image to assist with the comparison. The right image is currently highlighted to indicate that panning active on that image.*

TABLE 1—*Instructions for different conclusions.*

Instructions

Within the field of latent print identification, various groups are contemplating changes to the way that conclusions are reported, including additional categories beyond the traditional Identification/Inconclusive/Exclusion conclusions that have historically been used. (We also acknowledge that different labs use variants of these, such as just Identification/Exclusion or Identified/Not Identified, but for this experiment we are using the standard 3 conclusion scale). The goal of this experiment is to understand the consequences of moving to conclusions scales that have more categories

Structure

In this experiment, you will be conducting latent print identifications using both traditional and expanded conclusion scales on separate trials within the same experiment. This will allow us to measure how responses might change if examiners are given more choices. You will know on each trial how many conclusions you have available, which would be true in practice if an expanded conclusions scale were implemented

We would like you to use this scale when making your conclusions:

Exclusion: The two fingerprints originated from different fingers.

Support for different sources: The observed characteristics of the items are insufficient for exclusion but provide more support for exclusion than 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).

Support for common source: The observed characteristics of the items are insufficient for identification but provide more support for identification than exclusion.

Identification: The two fingerprints originated from the same finger.

You will make one of these choices on each trial using buttons at the bottom of the screen. There will either be 3 (exclusion, inconclusive, and identification) or 5 buttons (exclusion, support for different sources, inconclusive, support for common source, and Identification), depending on which scale is assigned to that trial.

Materials and methods

Participants

Twenty-seven latent print examiners from varying state and federal forensic facilities were participated. There were 17 female and 10 male participants, and they were required to have at least two years of experience during which they were qualified to testify in court.

Stimuli

The impressions used were from a 3,000-print database pulled 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 and 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 and then scanned into the same editing software to create the database.

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., [4]), 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, or perhaps slightly more difficult, than casework. To that end, we selected our nonmated images using left–right reversed impressions from the opposite hand of the donor individual. We used a subject matter expert (the third author) to select both mated and nonmated pairs that were similar in difficulty to what they experienced during typical casework. Thus, our exemplar impressions for nonmated pairs were designed to be challenging exclusions that for the most part bore superficial similarity to the latent impression.

For inspection purposes, the images are available from the corresponding author, although they are not publicly available because they are in use as stimuli in ongoing research projects.

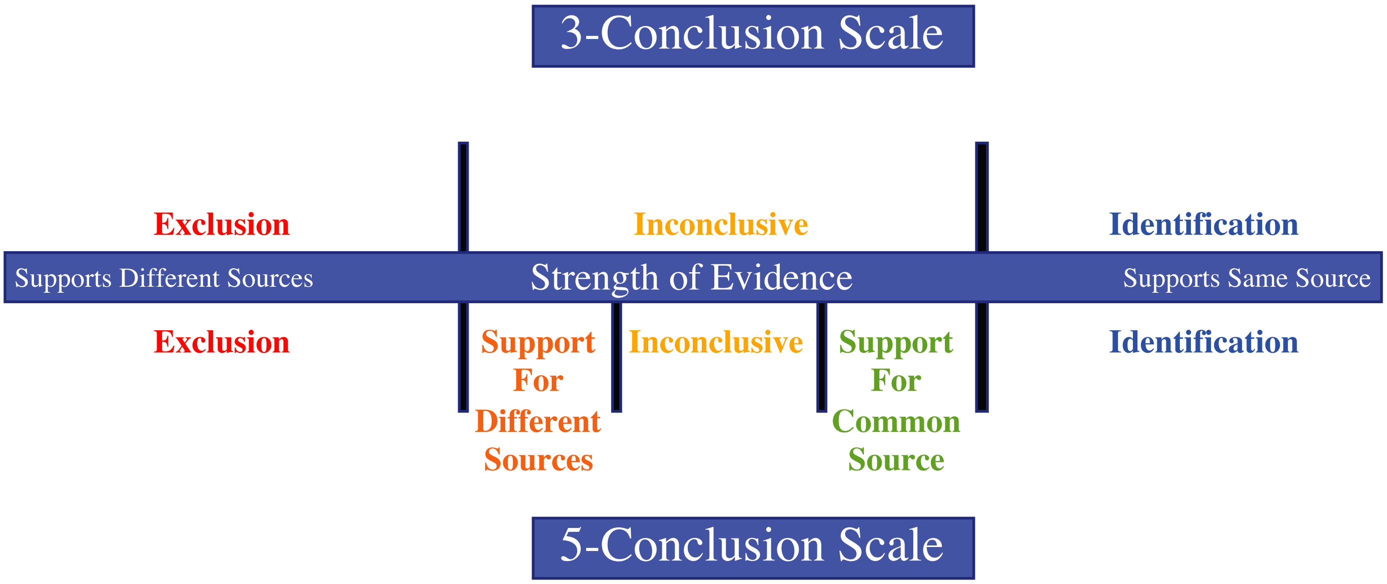

FIG. 4—*Conceptual comparison of a 3-conclusion scale with a 5-conclusion scale, and how the different thresholds might be distributed across the internal evidence axis (amount of perceived detail in agreement). In this hypothetical case, the qualified conclusions of "Support for Different Sources" and "Support for Common Source" simply subdivide what was previously the Inconclusive category.*

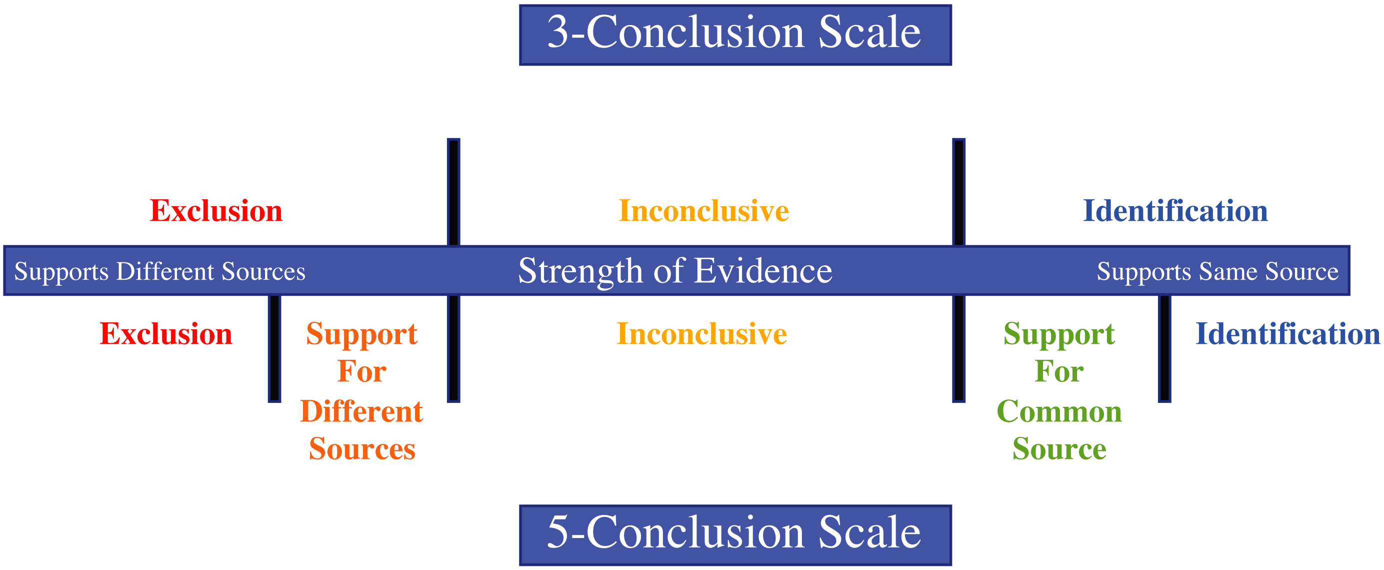

FIG. 5—*Conceptual comparison of a 3-conclusion scale with a 5-conclusion scale, and how the different thresholds might be distributed across the internal evidence axis (amount of perceived detail in agreement). In this hypothetical case, the qualified conclusions of "Support for Different Sources" and "Support for Common Source" act as qualifications on what were previously the definitive conclusions of Exclusion and Identification.*

Procedure

The study was composed of 60 trials, and each trial consisted of one fingerprint comparison. The experiment was administered electronically using a custom JavaScript-based interface designed to mimic the tools available during casework. 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 Fig. 3. The interface allowed the participants to zoom, rotate, and pan the individual images, as well as mark individual features with transparent digital markers. The experiment differed from normal casework in that participants only had 3 min to complete each trial and latent prints and exemplar prints were shown at the same time. This 3-min deadline might lead to more inconclusive conclusions than in casework, but as we will see <10% of the trials actually took the full 3 min, and this did not differ depending on which scale was used.

This experimental design omitted the "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. We made this decision because the interpretation of our results depends 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.

At the beginning of a trial, the two impressions were shown, and the participants had 3 min to make a decision. At the end of the 3-min 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. Half of the trials were designated as mated pairs, and half were nonmated. However, a coding error on two mated and two nonmated images produced as few as 13 mated or 13 nonmated pairs for some participates, or as many as 17 mated or 17 nonmated pairs. This coding error did not change the model fits or the conclusions, other than some conditions had slightly fewer or more trials than others. The overall number of trials is still sufficient to perform individual model fits even for participates affected by this coding issue.

Participants were instructed to never leave the program unless it was during a break screen. This forced participants to stay true to the 3-min time limit per trial, while allowing them to finish the 60 trials in multiple sittings if necessary. If the participant left the program during a trial without pausing the trial, when they opened it again it would auto-advance to the conclusion screen. No participants reported issues with the interface or recording their responses.

Participants received only the instructions and training provided by the text in Table 1 and did not have extensive training on the new categories in the expanded scale. We acknowledge that the behavior of examiners may change as they adapt to the use of the "support for" statements if they were to be included in operational casework. However, both the expanded and traditional scales use the "Identification" and "Exclusion" categories, which examiners have had experience with and presumably should not change, although this is of course an empirical question and discussed next.

Results

We first consider the behavioral results and then turn to the modeling via signal detection theory, upon which we derive our conclusions about the distribution of responses at the individual participant level. The individual responses tables for each participant, as well as for each image, are available in text files found at https://osf.io/kmprw/. See the wiki page on the osf.io site for more information on the individual files.

Prior to a discussion of the overall results, we first confront the issue of whether the artificial 3-min time limit affected our results in a systematic way. Once we alleviate this concern, we will discuss the responding that bears on the central question of the research.

Reaction Times and Full-Time Trials

Less than 10% of the overall trials required the full 3 min allowed, and there was no significant difference between the 3-conclusion and 5-conclusion scales in terms of the proportion of trials that took the full 3 min (M_3 = 0.09, M_5 = 0.079, t (26) = 0.961; p = 0.345; D = 0.006). Table 2 presents the number and proportion of trials for each combination of ground truth and response. With this statistical test and these below, we conducted paired t-tests, with M indicating the mean for each condition, p indicating the exact p-value, and D representing Cohen's D, a measure of effect size.

The two scales also did not differ in the overall amount of time taken to make a response. The 5-conclusion scale median is slightly, but not statistically significantly, longer by 5 sec (M_3 = 73.2, M_5 = 78.8, t(26) = −1.80; p = 0.084; *Cohen's* D = −0.35). Table 3 shows the median reaction times and interquartile range for each ground truth and response category. Responses were significantly faster for exclusion responses than for identification responses, for both the 3-conclusion scale (M_{ex} = 59.7, M_{ID} = 112.8, t(26) = −6.34; p < 0.0001; *Cohen's* D = −1.29) and the 5-conclusion scale including the "support for" conclusions (M_{ex} = 70.5, M_{ID} = 115.3, t(26) = −6.31; p < 0.0001; *Cohen's* D = −1.21). Thus, examiners appear to take approximately 70% longer to make an identification than to make an exclusion, but this does not depend on the whether they used the 3-conclusion or 5-conclusion scale.

Distribution of Responses

The distribution of responses is shown in Table 4 combined across all participants, and the distribution of responses converted to proportions is shown in parentheses. The proportions are similar to those in black box studies with fingerprints. For example, the correct identification rate in Ulery, Hicklin (4) is 0.452, and our correct identification rate for the 3-conclusion scale condition is 0.377. Likewise, our erroneous exclusion rate of 0.054 is similar to the 0.075 reported by Ref. (4). Finally, both results imply a larger inconclusive rate than in Ref. (4) (0.56 vs 0.31). There are three important reasons why our response rates might be somewhat lower than Ref. (4): First, we did not allow our examiners to say that a print was not of value, which likely increased the rate of inconclusive responses. Second, we limited the time that examiners could take on each impression to 3 min. Finally, our images may have differed in quality from the Ref. (4), and as those images are part of the criminal justice system and therefore not publicly available, we are unable to assess their difficulty. When translated to sensitivity (d') via signal detection theory modeling, the effective d' for Ref. (4) is about 2.6 while for our data are around 1.6. However, it is important to note that our results need only to be approximately similar to casework, because the goal of this study was not to measure error rates on an absolute scale, but to consider what changes might occur if an expanded conclusion scale is adopted. In addition, we are likely to observe the largest changes for the most difficult cases (13,23–25), and therefore, having more borderline comparisons may improve the efficiency of data collection.

Changes in Responses Across Scales

Inspection of Table 4 suggests several conclusions that we discuss here and support with modeling in a subsequent section.

First, the proportion of Identification responses to mated pairs drops from 0.377 in the 3-conclusions scale to 0.266 in the 5-conclusion scale. This suggests that examiners were redefining the term Identification to represent only the trials with the strongest evidence for same source.

Second, note that the Inconclusive rate drops from 0.569 in the 3-conclusion scale to 0.351 in the 5-conclusion scale. Some of these Inconclusive responses likely distributed to the support for same source response, because not all of the support for same source responses could have come from the weak Identification trials (0.377-0.266 is only 0.111, whereas the proportion of support for same source is 0.241).

Third, we observe two potentially concerning outcomes. We observe one erroneous identification outcome (saying "Identification" to a nonmated pair) in the 5-conclusion scale. This proportion is in line with the erroneous identification rate observed in Ref. (4) and therefore is not unexpected. Of potential more concern is the 17 outcomes that represent what might be thought of as "erroneous support for common source" outcomes (see Table 4). If a detective or prosecutor interprets these qualified conclusions as something more definitive, this has the potential to lead to a miscarriage of justice. However, note that there are 97 correct investigative leads from the Support for Common Source from mated pairs in Table 4, and perhaps slightly more than half of these would be purely inconclusive responses with a 3-conclusion scale (0.377–0.266 = 0.111, which is smaller than 0.241). Thus, there is a trade-off between a large number of investigative leads from mated pairs and what may be viewed as a concerningly large number of erroneous investigative leads from nonmated pairs. An evaluation of this trade-off will depend in part on how the Support for Common Source is interpreted by consumers and the value to society of a guilty individual in jail vs. an innocent person in jail. However, these "erroneous support for common source" outcomes are also an inevitable

TABLE 2—*Number of trials in each condition and response that ended at the 3-min time limit.*

Trial type	Exc	SFDS	Inc	SFCS	ID
3-Conclusion scale mated	8 (0.05)		15 (0.07)		0 (N/A)
3-Conclusion scale nonmated	1 (0.05)		29 (0.12)		21 (0.14)
5-Conclusion scale mated	3 (0.02)	13 (0.10)	10 (0.07)	1 (0.06)	0 (0.00)
5-Conclusion scale nonmated	1 (0.08)	1 (0.02)	14 (0.10)	12 (0.12)	9 (0.08)

Note that the cell counts here are only for trials that terminated at the full time; Table 4 has counts for all trials. The proportion of overall trials in that condition and response that reach 3 min is given in parentheses. These proportions will differ because participants are free to choose any one of the three (or five) conclusions and hover around 10% which is the overall rate at which participants used the full 3 min. N/A is not analyzed due to no responses in that condition and response.

TABLE 3—*Median reaction time (in sec) and inter-quartile range for the responses in each condition and response category. N/A is not analyzed (no responses in that cell).*

Trial type	Exc	SFDS	Inc	SFCS	ID
3-Conclusion Scale Mated	45.0 (40.7-66.6)		46.7 (22.6-98.1)		N/A
3-Conclusion Scale Nonmated	100.6 (52.4-120.9)		87.9 (51.7-110.6)		108.5 (88.1-145.1)
5-Conclusion Scale Mated	37.0 (21.7-51.4)	90.0 (52.9-109.9)	56.4 (24.4-97.8)	81.3 (22.2-155.8)	98.7 (98.7-98.7)
5-Conclusion Scale Nonmated	67.2 (54.5-125.1)	100.5 (70.1-117.8)	62.4 (30.4-106.1)	130.1 (94.2-161.4)	111.3 (68.5-136.2)

TABLE 4—*Data comparing the 3-conclusion scale (top two rows) with the 5-conclusion scale (bottom two rows).*

Trial type	Exc	SFDS	Inc	SFCS	ID
3-Conclusion Scale Mated	22 (0.054)		232 (0.569)		154 (0.377)
3-Conclusion Scale Nonmated	176 (0.438)		226 (0.562)		0 (0.000)
5-Conclusion Scale Mated	13 (0.032)	44 (0.109)	141 (0.351)	97 (0.241)	107 (0.266)
5-Conclusion Scale Nonmated	127 (0.311)	127 (0.311)	136 (0.333)	17 (0.042)	1 (0.002)

Shaded rows are mated pairs or impressions originating from the same finger. Nonshaded rows are nonmated pairs or impression originating from different fingers but are chosen to be similar in appearance. Numbers in parentheses are counts converted to proportions by summing the rows and diving each cell count by the row total.

consequence of the fact that fingerprint comparisons have less than perfect discrimination, and misleading evidence will tend to be higher for weaker conclusions (e.g., likelihood ratios near 1.0, see section 7.2.3 of (13)).

The observations above suggest that examiners may have redefined what is meant by Identification, but it is difficult to tell from the behavioral data along whether examiners get objectively worse when moving to a 5-conclusion scale. This requires modeling via signal detection theory, which is discussed next.

Modeling Results

Full code to reproduce the modeling results below is found in the Supplementary Information, located at https://osf.io/kmprw/. We modeled each individual subject's data using extensions of signal detection theory (20), and Fig. 6 illustrates a graphical representation of this formulation. The signal detection model requires the following assumptions:

- There exists a latent (unobservable) unidimensional evidence axis along which examiners accumulate evidence. This might be thought of as representing the balance of the strength of evidence between two propositions: one in which the two impressions share a common source, and one in which the two impressions have different sources. Evidence used to exclude two impressions as having a common source might be different than evidence used to conclude they share a common source, but the examiner should be able to combine these two sources of evidence into a single value that represents the relative strength of support for one position over another. This is similar to a likelihood ratio (or posterior odds ratio, since examiners often conflate the two or tend to ignore priors).
- The mated and nonmated distributions are distributed along the evidence axis according to a Gaussian distribution and have equal variances. Some models assume that the signal distribution has greater variance (the mated distribution in our case), but for simplicity and tractability when modeling individual subject data, we have made this simplifying assumption. Note that the observed distribution need not be Gaussian; we only need assume that the values are sampled from an underlying Gaussian distribution.
- The evidence axis is partitioned using either two (for the 3-conclusion scale) or 4 (for the 5-conclusion scale) decision criteria. Latent print examiners often refer to these as "thresholds," but for our purposes they are simply the criteria placed along the evidence axis such that any trial that produces an evidence value to the right of the upper decision criterion will elicit an Identification response, and a value to the left of the lower decision criterion will elicit an Exclusion response, and otherwise an Inconclusive response will result. The 5-conclusion scale model fits included two additional decision criteria placed in between the upper and lower decision criterion that represent the identification and exclusion criterion, respectively.
- The nonmated distribution is fixed at zero, and both distributions have a standard deviation of 1.0. This establishes the scale for the underlying evidence axis.

These four assumptions allow us to fit the data from each participant. We use a maximum-likelihood criterion and used custom Matlab (26) code (see supplementary material) and the function fminsearch to iteratively adjust the locations of the mated distribution for the 3- and 5-conclusion scales, along with the locations of all six decision criterion.

We report the results of three different models, each of which has a different set of constraints. These models allow us to test two specific hypotheses listed at the end of the Introduction that describe how the two conclusion scales relate to each other.

Does the Expanded Scale Produce Worse Sensitivity?

The first model is a full model, which allows the mated distributions to differ between the 3- and 5-conclusion scales and allows all six decision criterion (two from the 3-conclusion scale and four from the 5-conclusion scale) to freely vary. We will compare this model against a reduced model where the two conclusion scales are constrained to have equal mated distribution means. This model comparison specifically tests the hypothesis that examiners become objectively worse when moving from a 3-conclusion to a 5-conclusion scale. If the full model fits only slightly better than this reduced model, then we can conclude that sensitivity is not reduced by the addition of two extra conclusions to the scale. However, if the full model fits much better

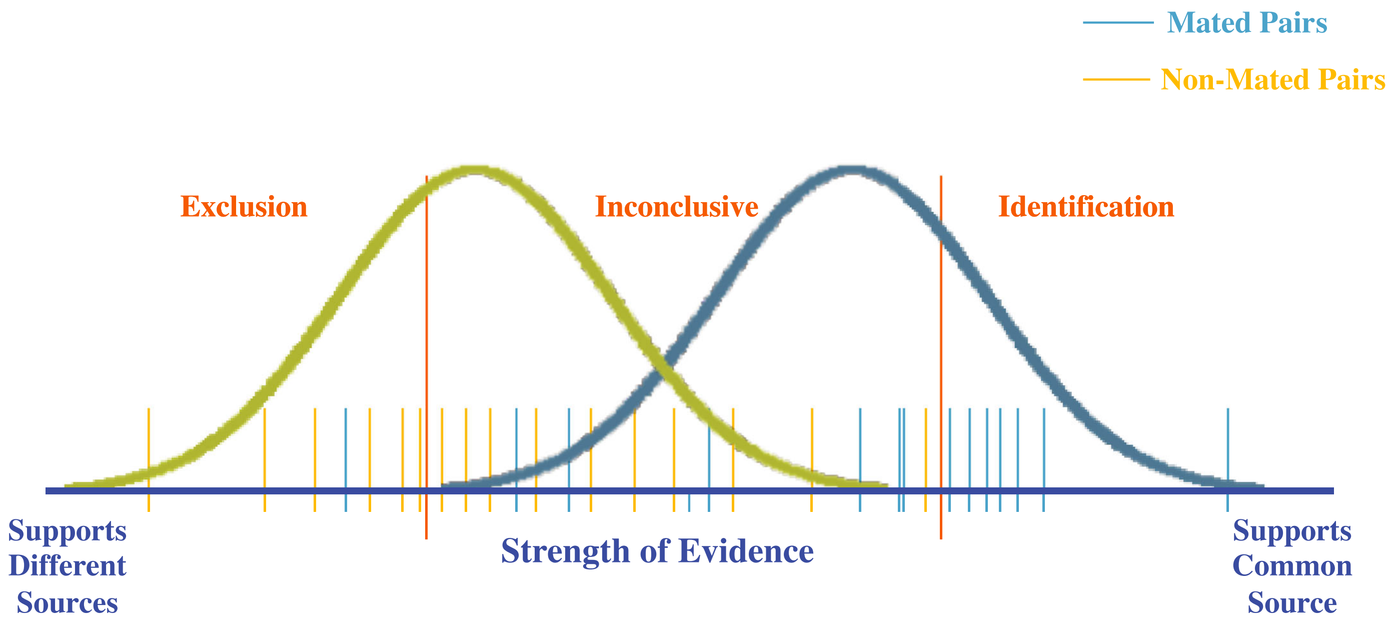

FIG. 6—*Example graphical representation of signal detection theory modeling of the response data from the 3-conclusion scale condition. Each vertical mark represents the hypothetical strength of evidence on a particular trial, where blue marks come from mated trials and yellow come from nonmated trials. The vertical red lines correspond to the two decision criteria that partition the evidence axis into Exclusion, Inconclusive, and Identification conclusions. We summarize the mated and nonmated distributions using Gaussian curves. The locations of the mated curve and the two decision criteria are iteratively adjusted such that the area under the curves between the decision criteria most closely corresponds to the observed proportions from an individual subject.*

than the reduced model, we will then conclude that one scale produces overall better sensitivity, and we can look at the values to determine which condition produces better sensitivity.

One possible outcome is graphically illustrated in Fig. 7, which shows a greater overlap between the mated and nonmated normal distributions in the lower graph. This reduction in sensitivity is not typically expected when the only change is the number of response conclusions. However, an expanded conclusion scale requires the maintenance in memory of the locations of each of the four decision criteria, and if these tend to shift over time, this will produce noise that computationally gets shifted into the variance of the nonmated and mated distributions. This effectively moves the two curves closer together. These effects were empirically observed by Benjamin, Tullis (21), and we would not want to recommend an expanded conclusion scale that made examiners objectively worse at separating mated from nonmated impression pairs.

We fit the data to the overall dataset and also examined fits to individual participants. Typically, signal detection theory models are not fit to group data, but instead to individual participants. However, there are two reasons to fit group data here. First, we are answering a *systemic* question: How does the aggregate behavior of all examiners change when the conclusion scale is expanded? In this case, the group data are the most appropriate to answer this question. Second, the group data are of course less noisy and will tend to give more stable outcomes. We tested this by doing both group analyses as well as individual model fits, and both the group analyses and the fits to individual participants give the similar results.

We fit a full model to the overall data with eight parameters (the mean of the mated distribution for each conclusion scale, 2 decision thresholds for the 3-conclusion scale, and 4 decision thresholds for the 5-conclusion scale) and a reduced model that fixed the mean for the mated distribution for both scales. Neither model is saturated because there are 12 degrees of freedom in the data, and no model has more than 8 free parameters (a saturated model essentially cannot be rejected under most circumstances and can prove complicated to interpret). The reduced model fits almost as well as the full model (D = 0.89, difference in parameters = 1, p = 0.34), which demonstrates that the 5-conclusion scale does not reduced examiner sensitivity.

Individual model fits were conducted on the data from each participant in which the two conclusion scale conditions were allowed to have different mated means (the full model above). For these fits, we compared the fitted values for the mated means for the two conclusion scales. Out of 27 participants, only 11 had smaller mated mean values in the 5-conclusion scale, which is an exact probability of 0.22. Based on these results, examiners do not perform worse when given a 5-conclusion scale relative to the 3-conclusion scale. This is different from the results of Benjamin, Tullis (21), but reassuring to the latent print community should an expanded scale be adopted. We should also note that the effects seen in Benjamin, Tullis (21) were quite modest with respect to changes in sensitivity with expanded scales.

Does the Expanded Scale Cause the Identification Conclusion to be Redefined?

A second and potentially more concerning question is whether the expanded conclusion scale results in a change to the identification threshold, as illustrated by Fig. 8. Our previous summary of the results suggested that the answer might be affirmative: Table 4 illustrates that the correct Identification rate drops from 0.377 in the 3-conclusion scale to 0.266 in the 5-conclusion scale. We approached the answer to this question in two ways. First, we can compare full and reduced models in which the full model allows separate identification thresholds for both scales, and the reduced model is constrained to have the same identification threshold for both 3-conclusion and 5-conclusion scales. If the full model fits

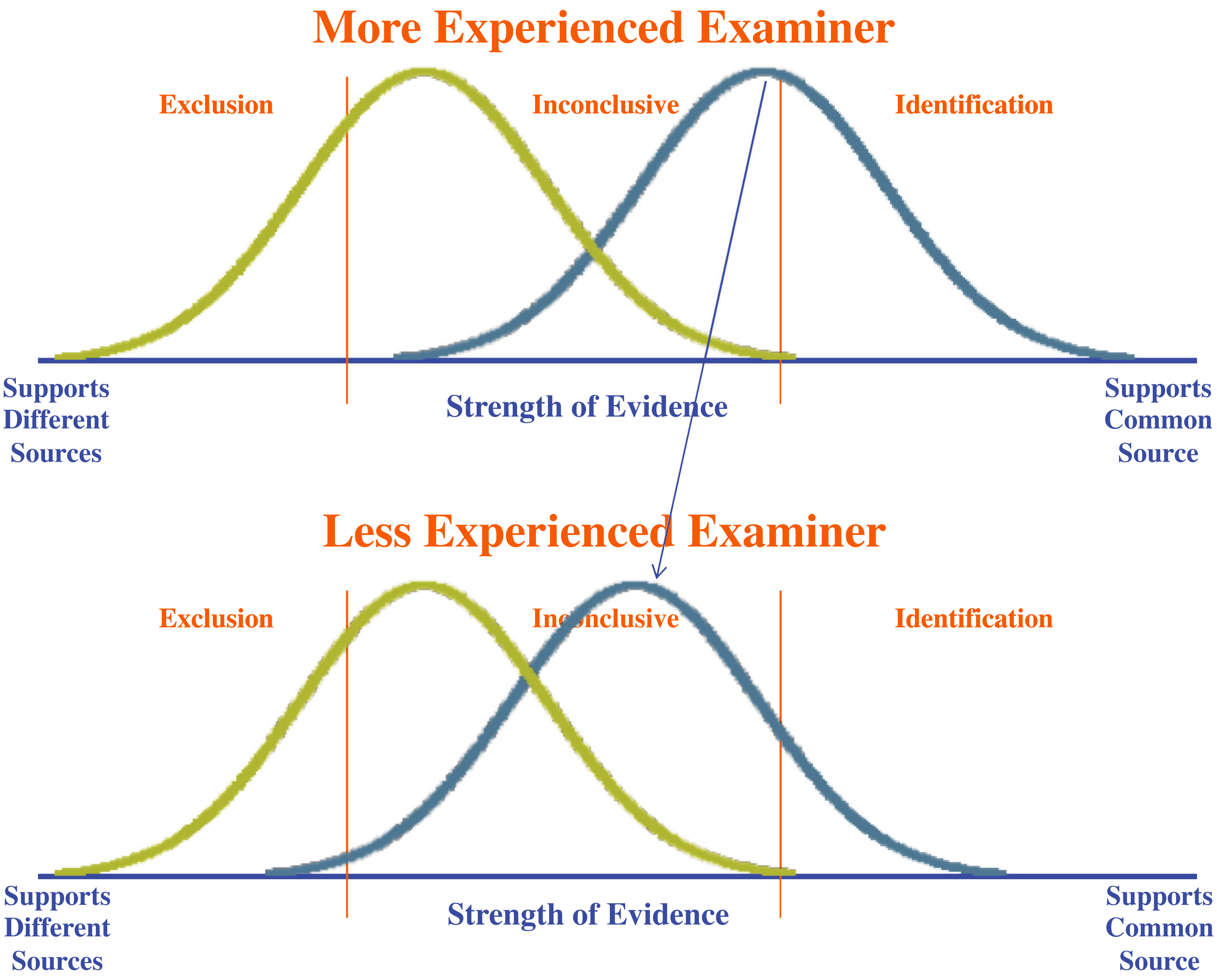

FIG. 7—*Example graphical representation of a reduction of overall sensitivity when examiners shift from a 3-conclusion scale to a 5-conclusion scale. In this case, they lose some of their ability to separate mated from nonmated pairs, as illustrated by the smaller separation between the yellow and blue curves in the lower figure and indicated by the arrow. Note that this can be determined independently of where they place their decision criteria in each condition, which in this example are arbitrarily placed.*

significantly better, we have evidence to support a redefinition of the Identification threshold with the 5-conclusion scale. Second, we can consider the full model fits for individual participants and determine whether the identification threshold for the 5-conclusion scale is systematically shifted relative to the 3-conclusion scale. Note that for these analyses, both conclusion scales were constrained to have identical mated means, as demonstrated by the results of the previous section, although relaxing this assumption does not affect the results in meaningful ways.

Our model comparisons take the difference in the fit statistic (the likelihood of observing the measured data given a particular model) between the two compared models. This value is D and is distributed as chi-squared with the difference in parameters between the two models as the degrees of freedom. From that, we can compute the associated p-value. When using this approach to compare the full model to the reduced model, we find that the full model performs significantly better than the reduced model, even considering its extra free parameter ($D = 4.18$, difference in parameters = 1, $p = 0.041$). This result demonstrates that participants systematically shifted their Identification decision criteria to more conservative values in the 5-conclusion scale (2.00 in the 3-conclusion scale and 2.26 in the 5-conclusion scale). The best-fitting parameters for this model fit are illustrated graphically in Fig. 9, which assumes equal sensitivity (d’) in the two conditions but allows all decision thresholds to vary. Note that, consistent with other representations of signal detection theory results, the abscissa is on a scale of units of the nonmated distribution standard deviation. Likelihood values can be computed at each location along this axis by dividing the height of the mated distribution by the height of the nonmated distribution, although standard admonitions about the behavior of ratios when the denominator is estimated by the tail of a distribution apply here.

We fit a model in which sensitivity was constrained to be equal for both conclusion scales, but all six decision thresholds were free to vary. When examining these model fits for individual participants, we find strong evidence for a shift in the Identification threshold for the 5-conclusion scale relative to the

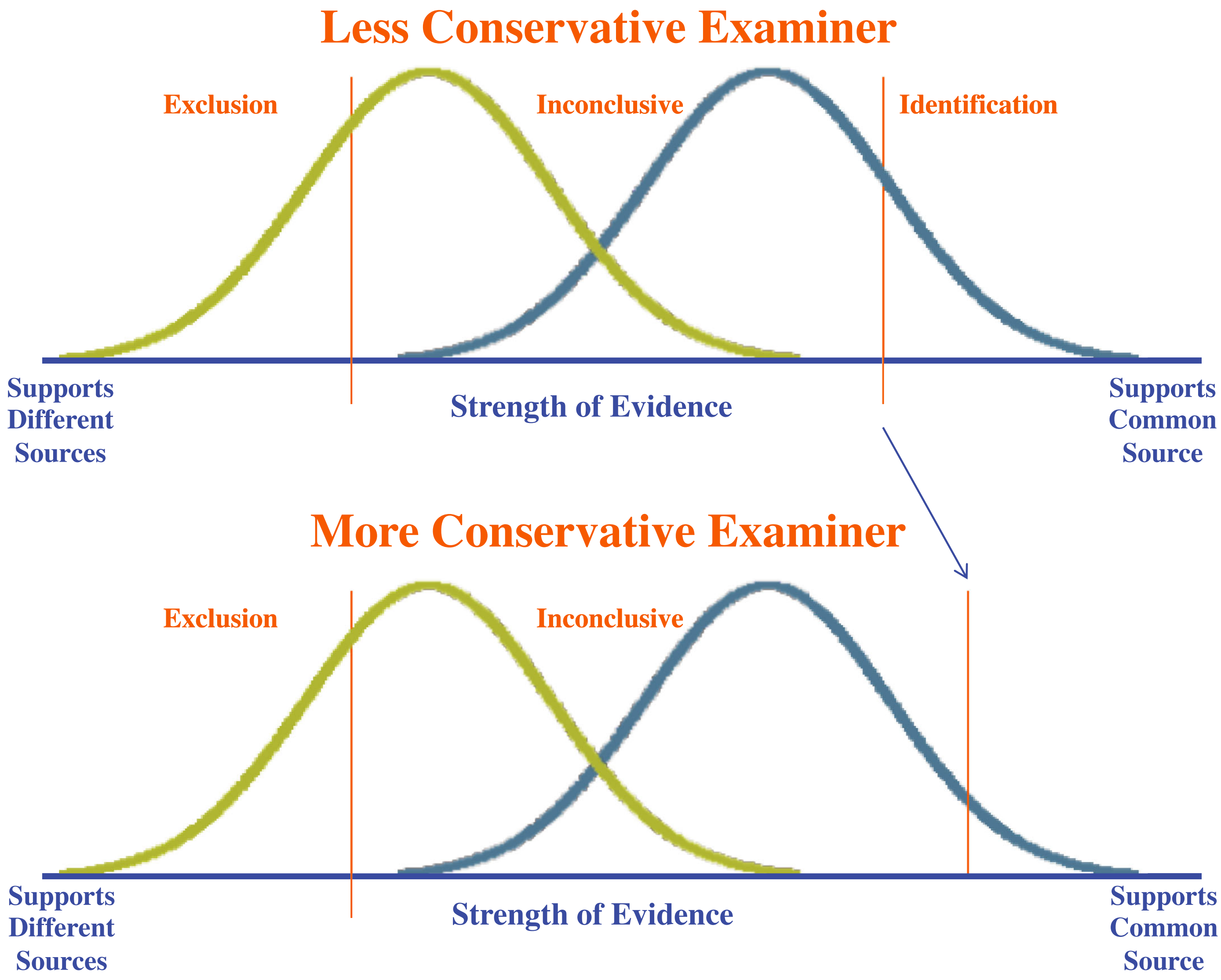

FIG. 8—*Example graphical representation of a shift in the decision threshold for the Identification when examiners shift from a 3-conclusion scale to a 5-conclusion scale. In this example, examiners shift their decision threshold to the right in the 5-conclusion scale condition (lower graph), as indicated by the blue arrow.*

3-conclusion scale. Out of 27 participants, 21 had an identification threshold shifted to the right (i.e., more conservative) in the 5-conclusion scale relative to the 3-conclusion scale (exact probability is 0.0029). This demonstrates that examiners *redefine* what they mean by an Identification when given more conclusions in the scale and become more conservative. This result is not an artifact of increasing the number of response items; see Appendix A for simulations verifying this conclusion.

Our fitted value for the Identification threshold as shown in Fig. 9 illustrates that examiners adopt an extremely rise-averse decision criterion, at or above two standard deviations away from the center of the nonmated distribution. This is consistent with other reported values: For Ref. (4), the Identification threshold is estimated at 2.97 standard deviation units away from the center of the nonmated distribution.

Discussion

The distribution of proportions shown in Table 4 suggests that our comparisons were of similar difficulty to those from black box studies, which are designed to emulate the difficulty of impressions encountered in casework. Thus, we believe that our choice of latent impressions and comparison exemplars produced an environment that is similar to actual casework. We expect that distribution of responses for the 3- and 5-category scales would be similar to what might happen in actual casework if the examiners were to adopt an expanded scale.

We did not find major reaction time differences across scales despite the added complexity of the 5-conclusion scale. We did find that examiners took longer to make an identification rather than an exclusion, which also aligns with conversations with examiners.

We did not find changes in overall sensitivity (as measured by the model fits of the mated and nonmated distribution in the signal detection theory model) between the 3-conclusion and 5-conclusion scales. This suggests that examiners are equally adept at mapping an internal strength-of-evidence value onto either of the two scales and that the additional burden of keeping two additional decision thresholds in memory was not large enough to reduce overall sensitivity.

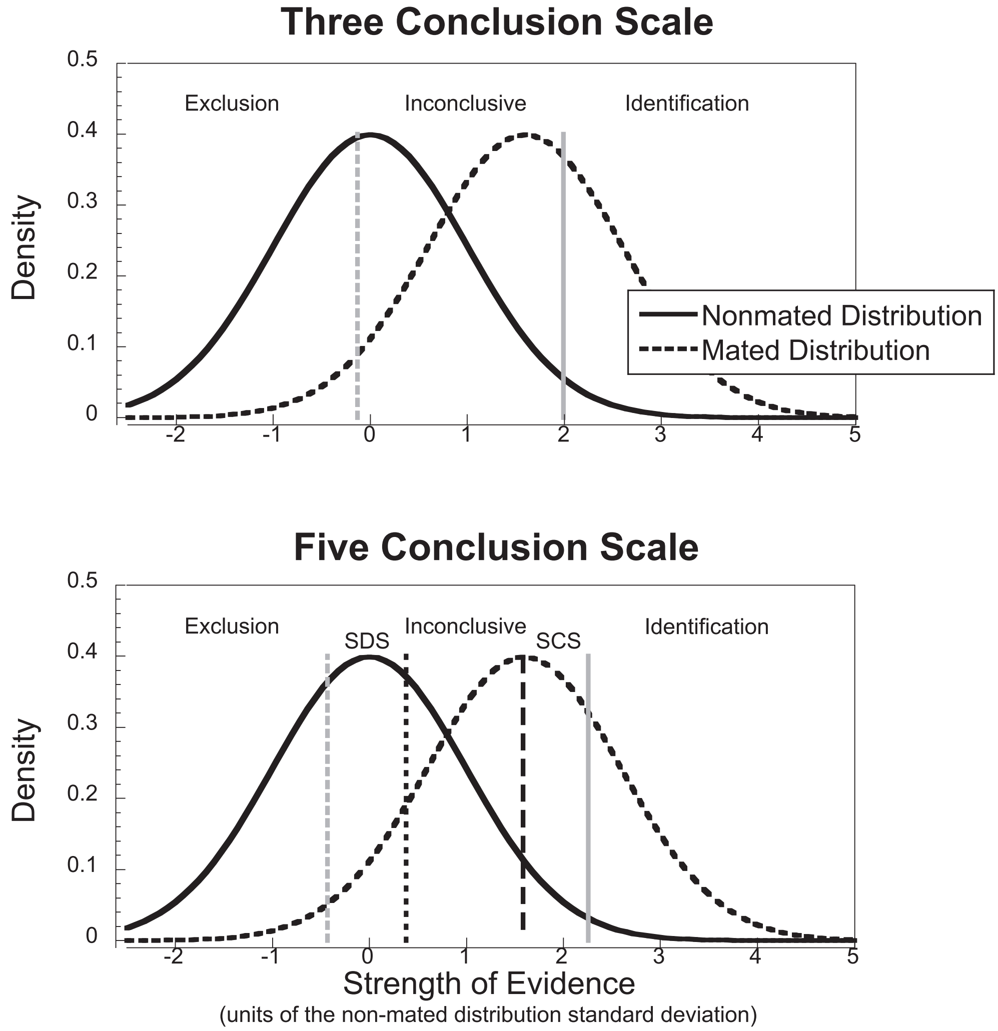

FIG. 9—*Model fit of the final signal detection model for the overall data for the 3-conclusion scale and 5-conclusion scale. The Identification threshold has shifted to the right in the 5-conclusion scale model fit, which demonstrates that examiners redefine the strength of evidence that is required to merit an "Identification" conclusion when given more categories in the scale. Note that both the Identification criteria moves to the right, and the Exclusion criteria moves to the left for the expanded 5-conclusion scale.*

The counts and proportions in Table 4 illustrate that the number of Identifications for mated pairs decreased when the two additional items were added to the 3-conclusion scale to create the 5-conclusion scale. These results were statistically significant, as demonstrated by the robust shifts in the Identification threshold across examiners. We found that 21 out of 27 examiners shifted their Identification decision threshold to the right in the 5-conclusion scale condition, which is graphically illustrated in Fig. 9.

Would our results change if examiners had more training on our revised scale? This is a difficult question to answer without additional data collection, especially given that the current thresholds are set through a vague combination of mentorship-based training, proficiency testing, verification within laboratories, and site-based training. We also see variations among laboratories, with some labs adopting a workflow that includes likelihood ratios (6). The present data were collected prior to these types of changes, but other inter-laboratory differences also existed, such as reporting "identified" and "not identified" as the only two conclusions. Our within-subject manipulation controls for some of these issues, but may have advantaged some participants over others. However, we return to two central issues: First, both scales included the term "identification," and second, examiners used the "support for" conclusions, when they were available, almost 35% of the time as illustrated by Table 4. Thus, there appears to be a need for these qualified conclusions and examiners readily make use of them.

Implications for policy

Inspection of Table 4 reveals both positive and potentially negative outcomes for casework. First, the Support for Common Source conclusion produced 97 investigative leads, which represents 25% of the mated pairs. Some of these came from the weaker Identification conclusions for mated pairs in the 3-conclusion scale, but many of them represent conclusions that would have been labeled as Inconclusive in the 3-conclusion scale. If these are viewed as pure investigative leads, we view these as a possible improvement to the criminal justice system.

However, the flip side of these 97 correct investigative leads is the 17 "erroneous" support for common source conclusions. These represent about 4% of the nonmated cases, and of course, the absolute number of these outcomes will depend on the base rates of mated and nonmated pairs in the laboratory. If nonmated pairs are more common, this could lead to many more erroneous support for common source conclusions than the 17 observed here. Thus, we view it as important that *consumers of investigative leads understand that these are not firm conclusions*.

Finally, it is worth speculating about whether a shift to an expanded scale is beneficial or harmful to the forensic community. To answer this question, one must consider not only the outcomes, but also the value to society of the various outcomes for more on the trade-offs that occur with changing decision threshold and how the outcomes are valued by society. Critically, the consumers must not over-interpret the Support for Common Source conclusion. This might be done by explaining that the examiner *could have* in principle reached an Identification conclusion, but in this case, did not, and has done so in the past with other cases. This may place the interpretation of this conclusion in the appropriate context. This illustrates the practical considerations when implementing an expanded scale: The additional (or revised) language must be interpreted correctly by consumers. In addition, more conclusions will require more conflict resolution, and proficiency testing will have to accommodate more options.

Is the shift in the identification threshold seen with an expanded conclusion scale good or bad? We believe that this in part depends on the value to society of a correct identification conclusion, as well as how the "support for" conclusions are interpreted by consumers. A more conservative identification threshold might help reduce (almost nonexistent) erroneous identifications but could also potentially lead to fewer correct identifications. These erroneous exclusions (or inconclusive decisions to mated pairs) receive much less attention despite the fact that they reduce the effectiveness of the criminal justice system. Thus, a conservative decision threshold may not be preferred. We view this as an important point of discussion as the field considers adopting expanded conclusion scales. In our view, the strengths of expanded conclusion scales outweigh the limitations, but care must be taken with their implementation.

Alternative Scales Based on Strength of Support Statement

As mentioned previously, the expanded conclusion scale illustrated in Table 1 is a mixture of definitive statements and strength-of-support statements. The inclusion of definitive language requires that the examiner consider the prior probability of a mated pair, which leads to a wide range of issues (7,13,22). The major problems include the fact that an examiner may not have all of the relevant information to construct a prior, that it is not their job as part of the legal system to deliver a conclusion, it is difficult to determine the utilities of different outcomes (7), and that once a posterior has been calculated, it becomes difficult to incorporate that evidence with other elements of the case (13). An alternative approach as currently practiced by European examiners relies on subjective likelihood ratio statements (15), which are designed to provide evidence about the outcome of the comparison in a manner that is easy to incorporate with the additional evidence in a case. The data in the present article did not compare pure likelihood ratio statements to the categorical statements, although such comparisons are currently underway in our laboratory (27). However, the extant literature makes a fairly compelling argument for the transition to strength-of-support statements (10,13,28), and we would add our voices to the chorus of scientists in support of this approach. This, however, does not neglect the host of operational difficulties that would arise is such a transition. For example, proficiency tests would have to take the form of measuring the skill of an examiner though specificity and sensitivity rates rather than overall accuracy on categorical scales. Consumers would have to learn to accept the strength-of-support statements into the overall evidence flow, rather than relying on examiners to make a decision for them. In addition, jurors have difficulty interpreting statements that include large probabilities (8), and accommodations involving verbal statements would have to be calibrated (9,10). Perhaps because of these challenges, the PCAST report (16) argued for a continuation of the categorical conclusions, supported by more extensive error rate (black box) studies, rather than suggesting a move to pure strength-of-support statements.

The PCAST report notwithstanding, it is our view that none of these obstacles toward adoption of a strength-of-support approach are intractable, and if the expanded conclusion scale illustrated in Table 1 represents a move toward pure strength-of-support conclusion scale, we encourage its adoption. However, if a transition is made, policymakers and laboratory directors should ask whether it makes sense to simply move directly to strength-of-support reporting.

References

1. OSAC Friction Ridge Subcommittee. Standard for friction ridge examination conclusions [DRAFT DOCUMENT]. Organization of Scientific Area Committees for Forensic Science. (2018). https://www.nist.gov/system/files/documents/2018/07/17/standard_for_friction_ridge_examination_conclusions.pdf (accessed January 22, 2020).
2. Taylor MK, Chapman W, Hicklin A, Kiebuzinski GI, Mayer-Splain J, Wallner R, et al. Extended feature set profile specification. NIST Special Publication (NIST SP). Gaithersburg, MD: National Institute of Standards and Technology, 2013(1134).
3. Ahissar M, Hochstein S. Task difficulty and the specificity of perceptual learning. Nature 1997;387(6631):401–6. https://doi.org/10.1038/387401a0.
4. Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. Accuracy and reliability of forensic latent fingerprint decisions. Proc Natl Acad Sci U S A 2011;108(19):7733–8. https://doi.org/10.1073/Pnas.1018707108.
5. PCAST. Ensuring scientific validity of feature-comparison methods. 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf (accessed January 22, 2020).
6. Swofford HJ, Koertner AJ, Zemp F, Ausdemore M, Liu A, Salyards MJ. A method for the statistical interpretation of friction ridge skin impression evidence: method development and validation. Forensic Sci Int 2018;287:113–26. https://doi.org/10.1016/j.forsciint.2018.03.043.
7. Biedermann A, Bozza S, Taroni F. The decisionalization of individualization. Forensic Sci Int 2016;266:29–38. https://doi.org/10.1016/j.forsciint.2016.04.029.
8. Garrett B, Mitchell G, Scurich N. Comparing categorical and probabilistic fingerprint evidence. J Forensic Sci 2018;63(6):1712–7. https://doi.org/10.1111/1556-4029.13797.
9. Thompson WC, Grady RH, Lai E, Stern HS. Perceived strength of forensic scientists' reporting statements about source conclusions. Law Probability & Risk 2018;17(2):133–55. https://doi.org/10.1093/lpr/mgy012.
10. Thompson WC. How should forensic scientists present source conclusions. Seton Hall L Rev 2017;48:773.
11. IAI. IAI Resolution 2010-18. International Association for Identification. 2010. http://clpex.com/swgfast/Resources/100716_IAI_Resolution_2010-18.pdf (accessed January 22, 2020).
12. Swofford HJ, Cino JG. Lay understanding of "identification": how jurors interpret forensic identification testimony. J Forensic Identif 2017;68(1):29–41.
13. Robertson B, Vignaux GA, Berger CE. Interpreting evidence: evaluating forensic science in the courtroom. Hoboken, NJ: John Wiley & Sons, 2016.

14. Champod C, Biedermann A, Vuille J, Willis S, De Kinder J. ENFSI guideline for evaluative reporting in forensic science: a primer for legal practitioners. Criminal Law and Justice Weekly 2016;180(10):189–93.
15. Willis S, McKenna L, McDermott S, O'Donell G, Barrett A, Rasmusson B, et al.ENFSI guideline for evaluative reporting in forensic science. European Network of Forensic Science Institutes. 2015. http://wp.unil.ch/forensicdecision/files/2016/02/Champod_etal_Primer_2016.pdf (accessed January 22, 2020).
16. Technology PsCoAoSa. Report to the President: forensic science in criminal courts: ensuring scientific validity of feature-comparison methods. (2016) September. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf (accessed January 22, 2020).
17. Cicchetti DV, Showalter D, Tyrer PJ. The effect of number of rating-scale categories on levels of interrater reliability – a Monte-Carlo investigation. Appl Psychol Meas 1985;9(1):31–6. https://doi.org/10.1177/014662168500900103.
18. Lozano LM, García-Cueto E, Muñiz J. Effect of the number of response categories on the reliability and validity of rating scales. Methodology 2008;4(2):73–9. https://doi.org/10.1027/1614-2241.4.2.73.
19. Mickes L, Wixted JT, Wais PE. A direct test of the unequal-variance signal detection model of recognition memory. Psychon Bull Rev 2007;14(5):858–65. https://doi.org/10.3758/Bf03194112.
20. Macmillan NA, Creelman CD. Detection theory: a user's guide, 2nd edn. Mahwah, N.J: Lawrence Erlbaum Associates, 2005.
21. Benjamin AS, Tullis JG, Lee JH. Criterion noise in ratings-based recognition: evidence From the effects of response scale length on recognition accuracy. J Exp Psychol Learn Mem Cogn 2013;39(5):1601–8. https://doi.org/10.1037/a0031849.
22. Biedermann A, Vuille J, Bozza S, Taroni F. Commentary on: Dror IG, Langenburg G. "Cannot decide": the fine line between appropriate inconclusive determinations versus unjustifiably deciding not to decide. J Forensic Sci 2019;64(1):318–21. https://doi.org/10.1111/1556-4029.13944.
23. Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. Repeatability and reproducibility of decisions by latent fingerprint examiners. PLoS ONE 2012;7(3):e32800. https://doi.org/10.1371/journal.pone.0032800.
24. Ulery BT, Hicklin RA, Roberts MA, Buscaglia J. Changes in latent fingerprint examiners' markup between analysis and comparison. Forensic Sci Int 2015;247:54–61. https://doi.org/10.1016/j.forsciint.2014.11.021.
25. Ulery BT, Hicklin RA, Roberts MA, Buscaglia J. Interexaminer variation of minutia markup on latent fingerprints. Forensic Sci Int 2016;264:89–99. https://doi.org/10.1016/j.forsciint.2016.03.014.
26. Mathworks IM. Natick, MA: Mathworks Inc, 2012.
27. Busey T. Validating conclusion scales in the forensic sciences. Bloomington, IN: Indiana University, 2019.
28. Stern HS. Statistical issues in forensic science. Annu Rev Stat Appl 2017;4:225–44. https://doi.org/10.1146/annurev-statistics-041715-033554.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

The Supplementary Information contains all of the Matlab source code for the signal detection theory model fits described in the paper. They can be run by executing the doFitOverallData.m file within Matlab.

Appendix A

Simulations to Test for Small-n Biasing of the Identification Threshold

There is the concern that the result above could be an artifact of the fact that each response bin for the 5-conclusion scale will tend to have fewer values in it (because there are more conclusions but the same number of trials). This might have artificially biased the identification threshold to the right. To address this concern, we conducted the following simulation. First, we fit a model to each participant in which we allowed for separate identification thresholds for each conclusion scale (as well as all of the other thresholds). Next, we set the Identification threshold for the 3-conclusion scale to that from the 5-conclusion scale, essentially forcing the two conclusion scales to have the same Identification threshold. From these parameters, we used Monte Carlo simulation to create simulated data for each participant. This Monte Carlo simulation used the underlying Gaussian distributions and the identification thresholds to create a new simulated dataset with the same number of trials per condition as experienced by the actual subjects. This effectively creates a dataset that is close to the true participant data, with the proviso that the identification thresholds for the two conditions were identical. This new simulated dataset was then fit with the full model that allowed for different Identification thresholds for the two scales. If the result from the previously referenced paragraph was an artifact, we would see a significant shift in the Identification for this simulated data. However, we did not find significant differences in the Identification thresholds, and what differences we did find were in the opposite direction (the median for the 3-conclusion scale = 2.90 and the median for the 5-conclusion scale = 2.75). There were 12 out of 27 simulated participants with larger values for the 5-conclusion scale, which gives an exact probability of 0.35. This can be compared with 21 out of 27 in the actual data. Thus, the larger values for the 5-conclusion scale seen in the individual model fits cannot be attributed to a bias in the model fits or an artifact of the fact that the bins for the 5-conclusion scale contain fewer counts than those from the 3-conclusion scale.