Final accepted manuscript, subsequently published as
Austin Hicklin, R., McVicker, B. C., Parks, C., LeMay, J., Richetelli, N., Smith, M., Buscaglia, J., Perlman, R. S., Peters, E. M., & Eckenrode, B. A. (2022). Accuracy, reproducibility, and repeatability of forensic footwear examiner decisions. *Forensic Science International 339*, 111418. https://doi.org/10.1016/j.forsciint.2022.111418

Accuracy, Reproducibility, and Repeatability of Forensic Footwear Examiner Decisions

R. Austin Hicklin[1], Brian C. McVicker[2], Connie Parks[1], Jan LeMay[3], Nicole Richetelli[1], Michael Smith[3], JoAnn Buscaglia[4], Rebecca Schwartz Perlman[5], Eugene M. Peters[4]*, Brian A. Eckenrode[4]

[1] Noblis
[2] Federal Bureau of Investigation, Laboratory Division, Questioned Documents Unit
[3] Denver Police Department Crime Lab, Denver, CO
[3] Federal Bureau of Investigation, Laboratory Division, Biometrics Analysis Section
[4] Federal Bureau of Investigation, Laboratory Division, Research and Support Unit
[5] Ideal Innovations, Inc.

* Corresponding author: empeters@fbi.gov

Abstract

The interpretation of footwear evidence relies on the expertise of forensic footwear examiners. Here we report on the largest study to date of the accuracy, reproducibility (inter-examiner variation), and repeatability (intra-examiner variation) of footwear examiners' decisions. In this study, 84 practicing footwear examiners each conducted up to 100 comparisons between questioned footwear impressions (provided as photographs and digital images) and known footwear (provided as photographs, transparent test impressions, and digital images), resulting in a total of 6,610 comparisons. The quality and characteristics of the impressions were selected to be broadly representative of those encountered in casework. A multilevel conclusion scale was used: 40% of responses were definitive conclusions (identification or exclusion), 14% probable conclusions (high degree of association or indications of non-association), 40% class associations (association of class characteristics or limited association of class characteristics), and 6% neutral conclusions (inconclusive or not suitable). On nonmated comparisons, 0.2% of conclusions were erroneous identifications (false positives), and 1.4% were incorrect responses of "high degree of association." The majority of erroneous identifications were made by a single participant. On mated comparisons, 6.0% of conclusions were erroneous exclusions (false negatives), and 1.8% were incorrect responses of "indications of non-association." Erroneous conclusions were sometimes reproduced by different examiners, but rarely repeated by the same examiner—1.1% of erroneous identifications were reproduced (none were repeated) and 19.9% of erroneous exclusions were reproduced (just one was repeated). Examiners' assessments of whether a questioned impression was suitable for comparison were notably inconsistent and may benefit from standardization. Rates of correct definitive conclusions are directly associated with the quality of the questioned impression and the extent of class similarities/differences between the questioned impression and known footwear.

1 Introduction

The interpretation of footwear evidence is conducted by forensic footwear examiners (FFEs) and relies upon their expertise, which is accumulated through training and operational casework. Owing to the subjective nature of interpretations in pattern evidence disciplines, including footwear examination, the need to assess the accuracy and reliability of forensic examiners' decisions has been identified in two well-known reports. In 2009, the National Academy of Sciences (NAS) published their report *Strengthening Forensic Science in the United States: A Path Forward* [1], recommending "the development and establishment of quantifiable measures of the reliability and accuracy of forensic analyses." That report asserted that such measures should be acquired through appropriately designed studies that present realistic casework data and scenarios to a variety of forensic scientists and laboratories, and also evaluate the limitations of accuracy and reliability as evidence conditions vary. These recommendations were reinforced in 2016 by the President's Council of Advisors on Science and Technology (PCAST) in their report *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* [2,3], which focused specifically on the need for improvement of pattern evidence methods and interpretation (including footwear impression evidence), advising that targeted research efforts should be dedicated to characterizing the accuracy and reliability of source attribution determinations.

Here we report on the largest study to date of the accuracy and reliability of forensic footwear examiners' comparison decisions. This black box study was conducted to rigorously measure the accuracy, reproducibility, and repeatability of FFE decisions

reported for footwear impressions selected to be broadly representative of casework; relate these results with the quality and characteristics of the footwear impressions; and evaluate the extent of association (if any) of performance with the education, training, and experience of participants. This study is intended to provide essential information about the forensic footwear discipline to practitioners, laboratory managers, the legal system, and researchers.

2 Background

Comparisons between questioned impressions left at crime scenes and the features on the bottom (outsole) of known footwear—for the purposes of determining whether the known footwear can be included or excluded as the source of the evidence—are the cornerstone of the forensic footwear discipline. These examinations are conducted using a sequential procedure outlined in a 2006 standard published by the Scientific Working Group for Shoeprint and Tire Tread Evidence (SWGTREAD) [4], which has been widely adopted in the discipline (see *Appendix B* for a detailed explanation of this process). This procedure begins with the FFE determining whether the questioned impression is suitable for a meaningful comparison with known footwear based upon its quality. Any impression deemed suitable is then compared to the known item of footwear using a sequential process. FFEs use their knowledge and experience to make comparison decisions based on their observations of the correspondence or non-correspondence of class characteristics and randomly acquired characteristics (RACs) between the questioned impression and known footwear. Class characteristics are features shared by two or more footwear items that can be used as the basis for inclusion or exclusion decisions. Class characteristics include outsole design and physical size features resulting from the manufacturing process, and the position and degree of wear on the outsole. RACs are outsole features acquired through use of the footwear that can be used to differentiate outsoles with similar class characteristics, and can be used as the basis for determining that a specific outsole is the source of a given impression. (See *Appendix A* for a glossary.) Examiners use a multi-level scale to report their conclusions, thereby providing a means to convey varying degrees of support for or against the known shoe being the source of the questioned impression in the context of the evidence quality, their observations and findings, and any limitations encountered. Currently, the 2013 conclusion scale published by SWGTREAD remains the prevailing standard in the United States: familiarity with this scale is advocated by the International Association for Identification (IAI) for training and certification [5]; some proficiency tests [6–8] and previous research studies [9–12] required participants to respond using this scale; some forensic service providers use a modified version of this range of conclusions (e.g., the Department of Justice (DOJ) has developed a standard for their forensic practitioners [13]).

Several studies have evaluated the performance of FFEs when making conclusions in footwear comparisons. Prior to the current study, the largest and most comprehensive footwear black box study evaluated the accuracy and reproducibility of 840 comparison conclusions from 70 FFEs [10–12], as well as the reproducibility of several examination determinations, including the assessment of class characteristics and RACs [10]. Other previous studies were more limited in size and scope, with the total number of comparisons ranging from 40 to 240 [9,14–16]. The results of each generally focused on reproducibility of examiner decisions, sometimes under ideal conditions (e.g., no evaluation of class characteristics required and outsole features of interest highlighted). Raymond and Sheldon [9] and Speir et al. [10] both assessed the accuracy of FFE decisions with respect to "expected" responses (as determined by the study designers), participant consensus, and ground truth source attribution. None of the aforementioned studies assessed repeatability of examiner conclusions. Table 1 summarizes the experimental design of these previous studies as well as the current study.

Study	*Participants*	*Comparisons per Participant*	*Questioned-Known Pairings*	*Total Comparisons*	*Additional Study Design Notes*
Majamma & Ytti [14]	33	6	6 distinct, 0 repeated	198	Indicated correspondence of class characteristics; RACs of interest highlighted for evaluation
Shor & Weisner [15]	20	2	2 distinct, 0 repeated	40	Provided real casework impressions (ground truth unavailable)
Hammer et al. [16]	40	6	6 distinct, 0 repeated	240	Indicated correspondence of class characteristics; RACs of interest highlighted for evaluation
Raymond & Sheldon [9]	28	6	12 distinct, 0 repeated	168	Conducted in two trials (6 comparisons each), with some different participants in each trial
Speir et al. ("WVU Study") [10–12]	70	12	12 distinct, 0 repeated	840	One comparison replaced after first cohort of participants (n=5)
Current Study	84	100	269 distinct, 30 repeated	6610	Details provided in this document

Table 1. Experimental design summary for previous work and current study.

3 Study Description

This study was conducted to characterize the accuracy and reliability of conclusions reported by FFEs when comparing questioned impressions with known items of footwear. We assessed reliability in terms of reproducibility (inter-examiner variability) and repeatability (intra-examiner variability). In addition, this research evaluated the relationships between participants' performance and their training and experience, as well as those between the conclusions reported by FFEs and the attributes of the questioned impressions.

3.1 Participation

Participation in this study was open to U.S. FFEs who had performed forensic footwear evidence comparisons in operational casework within the last five years, along with non-U.S. FFEs if they met the above requirements and used either the 2006 or 2013 SWGTREAD conclusion scale. A total of 84 FFEs participated in this study, with 55 (65%) of those completing all 100 comparisons. A background questionnaire was required to be completed by all participants prior to starting the study. Of the participants, 40% had more than ten years of experience, whereas 30% had less than five years. Footwear examination is generally not the only professional focus of the participants: none of the participants conducted footwear examinations daily, only eight participants conducted footwear examinations a few times a week, and most (57%) conducted footwear examinations only a few times a year; only one participant indicated spending more than half of a typical work week conducting footwear examinations. Eighty of the 84 participants also perform fingerprint examination, crime scene processing, and/or tire impression examination. Of the participants, 31% completed a formal program of forensic footwear examination instruction for 1 year or more, and an additional 37% completed formal instruction for at least 6 months. Just under a third (31%) of participants are or have been certified as footwear examiners. Almost all participants were from U.S. local agencies (31%), U.S. state agencies (35%), and international governmental agencies (30%); 71% of participants were from accredited agencies, almost all of whom were accredited under ISO 17025. A more detailed description of participants is included in *Appendix C3*, and their survey responses are included in *Appendix P*.

3.2 Study Overview

Each participant was assigned 100 footwear comparisons, to be completed over a period of one year. As shown in Figure 1 and Figure 2, each comparison set (QKset) included up to three reproductions of a single questioned impression (Q), to be compared against two test impressions and up to five outsole images from a single known footwear item (K). Each participant was assigned 40 *mated* QKsets (in which the K is the source of the Q), and 60 *nonmated* QKsets (in which a footwear item other than the K is the source of the Q); participants were not told of these proportions. Each QKset was provided to the participants in an envelope containing high-quality photographs of the questioned impression and the known footwear item, and transparencies of the test impressions. All images depicting outsoles and lifts were reversed (i.e., the images were flipped horizontally) so that they orient with the impression on the ground; such images were labelled "Reversed." A custom online interface was created to enable participants to register for this study, complete both pre- and post-test surveys, respond to pre-defined comparison questions, mark the positions of corresponding RACs (if applicable), and indicate the orientation of the questioned impressions. This online interface also allowed participants to download high-resolution (600 pixels per inch (ppi)) digital images, in both JPEG and TIFF formats. The footwear comparisons were assigned in five "packets," each containing 20 QKsets; responses to all QKsets in a packet needed to be submitted before receiving the next packet. To evaluate repeatability (intra-examiner variability), each participant was assigned ten QKsets that contained the same imagery as an earlier comparison (i.e., each participant was assigned 100 total QKsets, but only 90 distinct QKsets). These repeated sets were assigned different QKset numbers and were assigned to participants in different packets, to reduce the likelihood that participants would recognize the questioned impressions and/or known footwear items (outsole photos, test impressions).

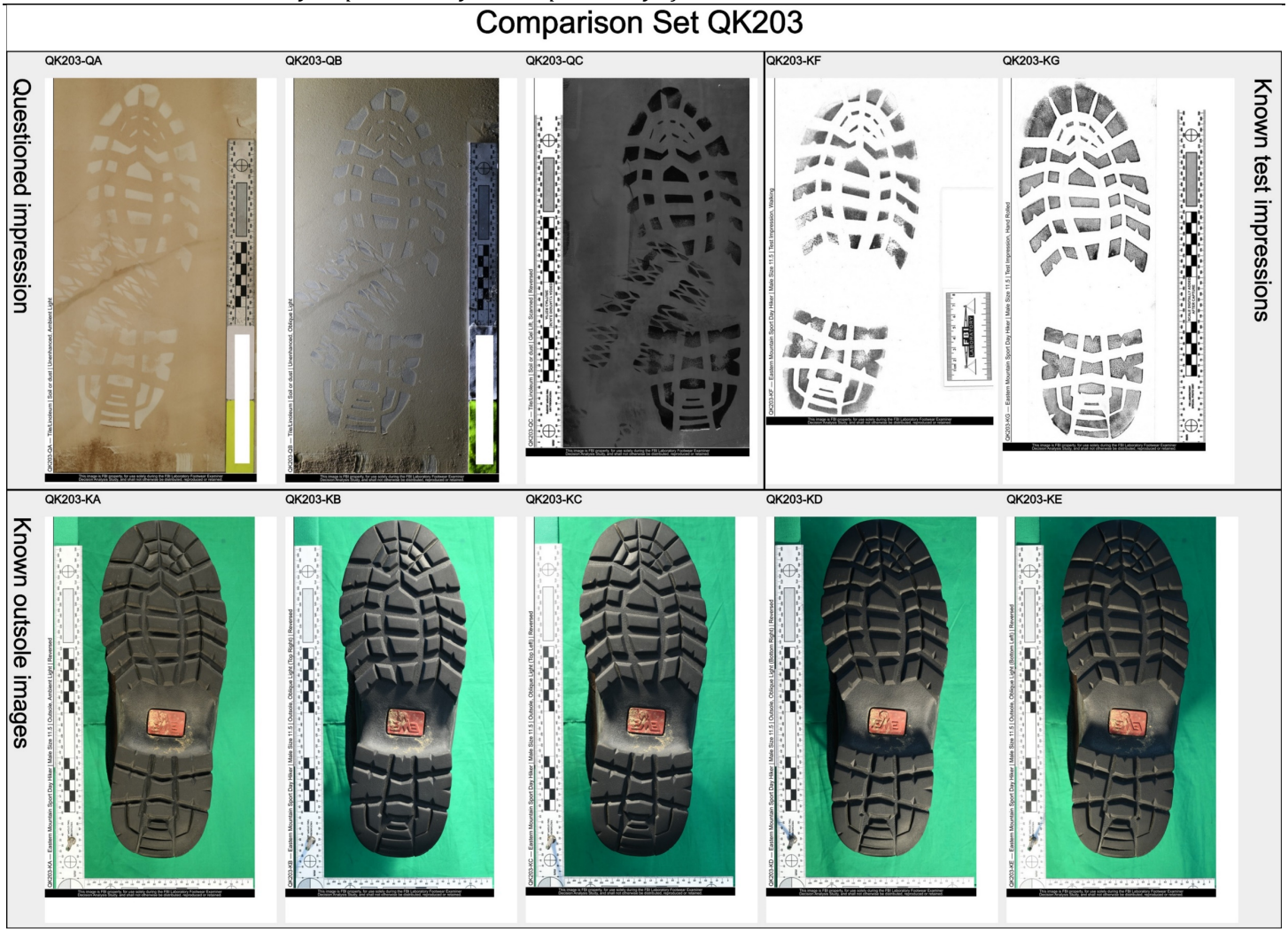

Figure 1. Example of a nonmated comparison set. Eastern Mountain Sports Day Hiker boot (male size 11.5, right foot). Questioned impression is a negative impression in which a thin layer of soil or dust was removed from the substrate (tile); note that the same questioned impression is captured in three images (ambient light (QA), oblique light (QB), and black gel lift (QC)). Known boot was reproduced as two test impressions (walking (KF) and hand rolled (KG)), and five outsole images (ambient light (KA) and four directions of oblique light (KB-KE)). This was the only comparison set that had more than one erroneous ID. (19 Assignments: 2 *ID*, 0 *HighAssn*, 4 *Assn*, 0 *LimitedAssn*, 0 *Inc*, 0 *NotSuitable*, 1 *NonAssn*, 12 *Excl*).

Figure 2. Example of a mated comparison set. Sperry Top-Sider shoe (male size 10.0, left foot). Questioned impression is residue on wood. This QKset had a high rate of erroneous exclusions. (26 assignments: 0 *ID*, 0 *HighAssn*, 0 *Assn*, 11 *LimitedAssn*, 3 *Inc*, 1 *NotSuitable*, 3 *NonAssn*, 8 *Excl*).

All study samples (i.e., questioned impressions, outsole images, and test impressions) were collected under controlled laboratory conditions with quality assurance measures designed to guarantee the ground-truth source attribution (mating) of the QKsets. The study team sought to create study samples that were similar to those encountered in casework, spanned the spectrum of quality, incorporated a variety of footwear models, and provided the opportunity for participants to utilize the full range of conclusions; see *Appendix C4* for details. A novel method of assessing the quality of questioned impressions was developed for this study (published separately in [17], summarized in *Appendix F1*), and was used to ensure an appropriate distribution of a variety of attributes in the selection of questioned impressions and assignment of QKsets. The study samples included 25 different footwear makes or models, but the Eastern Mountain Sports Day Hiker boot (shown in Figure 1) was used in about half of the QKsets (133/269). Of the 67 participants who completed the post-test survey (*Appendix P2*), 49 (73%) assessed the quality of the outsole images as exceptional, 46 (69%) assessed the test impressions as exceptional, and none assessed the quality as unacceptable. The inability to examine the original (physical) known footwear items was a limitation for some participants: 7 (10%) said this prevented them from making more definitive conclusions, and 38 (57%) said it sometimes did. A third of the participants (22) said the physical footwear was not required. Their responses regarding the comparability of the QKsets with casework further validated the creation and selection of study samples: 52 (78%) said the comparisons were comparable with casework; 7 (10%) said they were harder than casework, and 8 (12%) said they were easier (or much easier) than casework.

Participants were asked to complete their analyses and comparisons in a manner similar to casework, reporting responses for each phase, as illustrated in Figure 3. For each of the assigned QKsets, the participants evaluated suitability; assessed the correspondence of outsole design, mold variation, physical size, and wear; marked corresponding RACs (if applicable); selected a conclusion; rated difficulty; detailed any limitations encountered; and oriented the questioned impression with the toe pointing up. (See *Appendix C5* for details.)

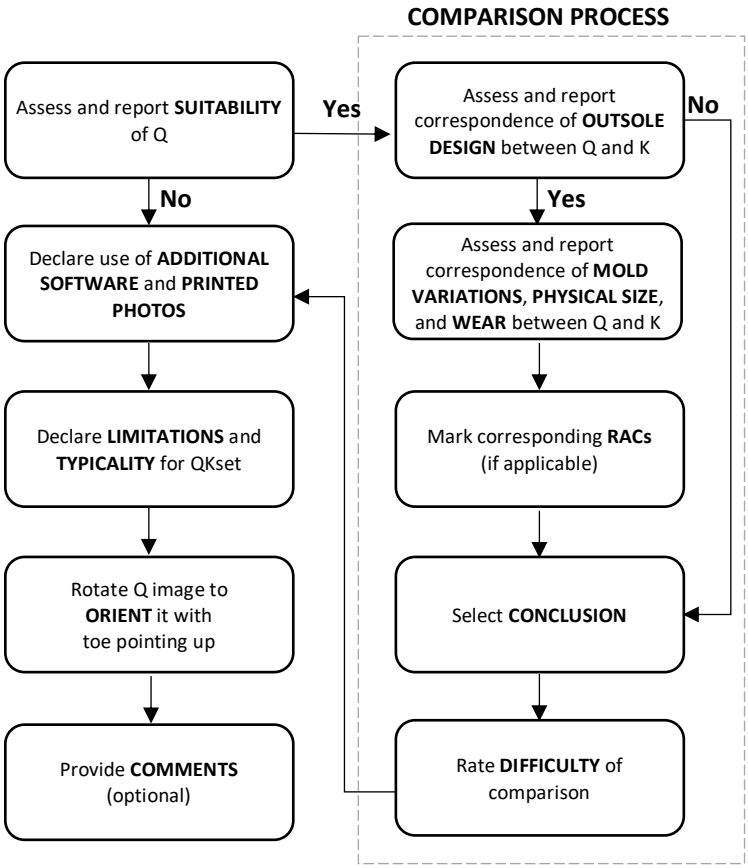

Figure 3. Flowchart detailing the comparison and reporting process presented in this study.

The study used the following conclusion scale (summarized here; see *Appendix C5.4* for the complete definitions used in the instructions):

- Not suitable (***NotSuitable***) — The questioned impression lacks sufficient detail to enable a meaningful comparison.
- Suitable
  - Identification (***ID***) — The particular known footwear was the source of, and made, the questioned impression. Another item of footwear being the source of the impression is considered a practical impossibility.
  - High degree of association (***HighAssn***) — The observed characteristics exhibit strong association between the questioned impression and the known footwear item; however, the quality and/or quantity were insufficient for an identification.
  - Association of class characteristics (***Assn***) — The class characteristics of design and physical size (and possibly wear) correspond between the questioned impression and the known item of footwear; the known item of footwear is a possible source of the questioned impression and therefore could have produced the impression.
  - Limited association of class characteristics (***LimitedAssn***) — Some similar class characteristics were present; however, there were significant limiting factors in the questioned impression that did not permit a stronger association between the questioned impression and the known item of footwear.
  - Inconclusive (***Inc***) — Significant limitations in the evidence prevented any specific association or non-association; it could not be determined whether the known item of footwear is or is not the source of the questioned impression.
  - Indications of non-association (***NonAssn***)— Dissimilarities between the questioned impression and the known footwear indicated non-association; however, the details or features were not sufficient to permit an exclusion.
  - Exclusion (***Excl***)— The known item of footwear was not the source of, and did not make, the questioned impression.

Note that this conclusion scale is a modification of the SWGTREAD 2013 conclusion scale [18], which does not provide an option for examiners to report a completely neutral opinion: therefore, "inconclusive" (from the SWGTREAD 2006 scale [19]) was added by the research team to accommodate that situation. In the background questionnaire, 67% of participants indicated that they use the SWGTREAD 2013 range of conclusions [18] in casework, and an additional 13% use the SWGTREAD 2006 range of conclusions [19]; 25% use "inconclusive" as a category in casework.

For some analyses we group the conclusion categories, referring to "definitive conclusions" (*ID* and *Excl*), "probable conclusions" (*HighAssn* and *NonAssn*), "class associations" (*Assn* and *LimitedAssn*), and "neutral responses" (*NotSuitable* and *Inc*).

4 Results and Discussion

Analyses were based on a total of 6,610 responses from 84 participants on 269 distinct QKsets. Responses were omitted from two participants who did not complete at least 20 QKsets, and from one mated QKset in which the Q on a glass substrate was inadvertently reversed when photographed. To evaluate repeatability (intra-examiner variability), each participant who completed the study was assigned ten QKsets twice: the responses on second assignments were not included in most analyses because a subset of the QKsets would have a disproportionate effect on the results. For this reason, second responses were not used to calculate overall conclusion rates .

For most analyses we use the *Baseline Dataset* (6,032 responses from 84 participants on 269 distinct QKsets), which omits the 578 second assignments. For analyses of repeatability, we use the *Repeat Dataset*, which includes 578 pairs of first and second assignments (1,156 total responses) by 64 participants on 30 distinct QKsets. For the analyses that computed and compared rates for individual participants, analysis is limited to only those participants who completed at least 40 QKsets (*Examiner Comparison Dataset*—5,749 responses by 71 participants on 269 distinct QKsets). Deidentified participant responses are included as supplemental information to this publication.

Each QKset was assigned to one third of the participants. The *Baseline Dataset* includes responses from 16-30 participants per QKset (mean 22.4, median 23) — overall (including repeats), responses were received from 16-54 participants per QKset (mean 24.6, median 23).

4.1 Conclusion Rates, Accuracy, and Errors

Figure 4 illustrates the distribution of the participants' conclusions for the *Baseline dataset*, with the corresponding rates shown in Table 3 (see *Appendix D5* for discussion, detailed results, and confidence intervals). We use "accuracy" as a general term referring to the extent to which conclusions are consistent with ground truth. As we will show, accuracy can be measured using a variety of methods, which are affected by the use of a multilevel conclusion scale.

Accuracy can be described as the extent to which errors and incorrect responses are avoided. When discussing responses, we only use the term "error" to refer to definitive conclusions that contradict ground truth: we explicitly differentiate errors from probable conclusions and class associations that contradict ground truth, which we describe as "incorrect;" we make this distinction because the participants explicitly differentiate between definitive and probable conclusions, and it would therefore be inappropriate to lump them together into a single category of error. In a few cases, we describe responses as "debatable" when it is arguable whether a response is or is not consistent with ground truth, such as *Assn* on a nonmated QKset in which the Q and K are the same make and model but differ by ½ size. We are cautious in referring to responses that are consistent with ground truth as "correct" because in a multi-level conclusion scale more than one response may be consistent with ground truth—please see the discussion in Section 4.3 regarding whether a given response may be considered "appropriate." Assessing the accuracy of class associations requires a novel approach, because these are not evaluated with respect to mating, but to the ground truth class characteristics of the questioned impressions and known footwear, as shown in Table 2: *Assn* and *LimitedAssn* are considered consistent with ground truth if the Q and K are the same make and model, and size (whether they are mated or nonmated); they are only considered incorrect if the Q and K are of different makes, models, or opposing feet (e.g., the source of Q is a left shoe while the K is from a right shoe).

		Mated	*Nonmated*			
			Same make/ model/size	*Same make/model, ± ½-1 size*	*Different make or model*	*Different foot*
Conclusion	*Excl*	False negative (FN)	True negative (TN)			
	NonAssn	Incorrect non-association (IN)	"Correct" non-association (CN)			
	NotSuitable	Neutral				
	Inc					
	LimitedAssn	"Correct" classification		Debatable	Incorrect classification	
	Assn					
	HighAssn	"Correct" association (CA)	Incorrect association (IA)			
	ID	True positive (TP)	False positive (FP)			

Table 2. Categories of conclusions. Errors and incorrect conclusions are highlighted; neutral and debatable conclusions are shown in gray. "Correct" is used here only to indicate a conclusion is consistent with ground truth, not that a given conclusion is necessarily appropriate. Note that some of these categories were developed for this study: the use of a multilevel conclusion scale requires expanding on existing approaches (such as used in [20,21]). (See *Appendix D5* for definitions and more detailed discussion.)

On mated QKsets, 6.0% of trials yielded erroneous *Excl* conclusions (False Negative Rate, FNR)* and 1.8% of trials resulted in incorrect *NonAssn* conclusions (Incorrect Non-association Rate, INR). On nonmated QKsets, 0.2% of trials yielded erroneous *ID* conclusions (False Positive Rate, FPR), and 1.4% of trials resulted in incorrect *HighAssn* conclusions (Incorrect Association Rate, IAR). These errors were disproportionately caused by a subset of participants, which will be discussed in detail in Section 4.2.

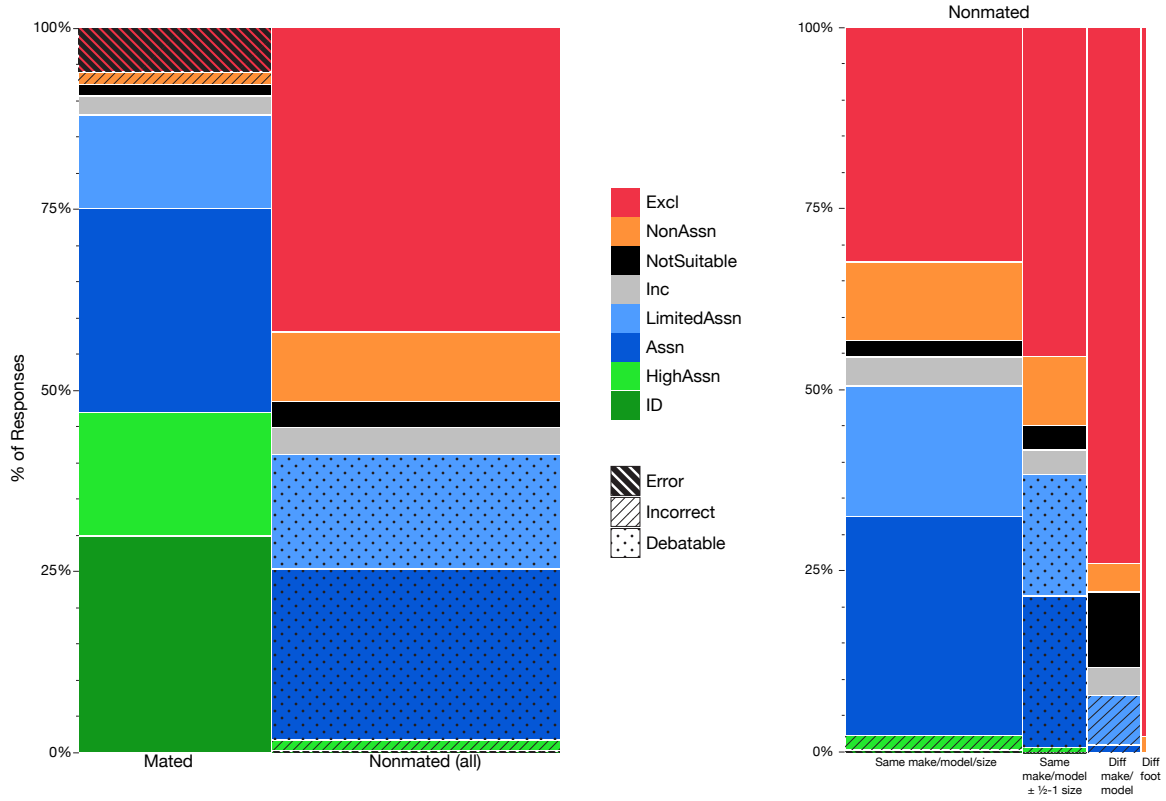

Figure 4. Distribution of conclusions: graphical representation of data in Table 3. The right chart subdivides nonmated data by class characteristic category. (*Baseline Dataset*: n=6,032 responses on 107 mated and 162 nonmated distinct QKsets; 2,417 mated and 3,615 nonmated responses.)

The accuracy of responses on nonmated trials varied substantially with respect to the extent that the Q and K shared class characteristics. As part of the study design, nonmated QKsets were divided into categories, shown in the right columns of Figure 4 and Table 3 (details in *Appendix D3*). The proportions of *Excl* responses (true negatives) dramatically increased as the differences in class characteristics increased. Note that when limited to comparisons with the same make, model, and size, the

* *We calculate conclusion rates based on all presentations ("PRES") of QKsets; see Appendix D5 for discussion and detailed results.*

mated and nonmated distributions of conclusions are relatively comparable, with about 30% definitive conclusions (*ID* and *Excl*) and about 30% *Assn* responses.

		Mated		*Nonmated*									
				Nonmated (all)		*Same make/ model/ size*		*Same make/model, ± ½-1 size*		*Different make or model*		*Different foot*	
Conclusion	*Excl*	146	6.0%	1,515	41.9%	724	32.3%	367	45.3%	382	73.7%	42	97.7%
	NonAssn	43	1.8%	346	9.6%	246	11.0%	78	9.6%	21	4.1%	1	2.3%
	NotSuitable	39	1.6%	131	3.6%	51	2.3%	27	3.3%	53	10.2%	0	0.0%
	Inc	60	2.5%	137	3.8%	88	3.9%	28	3.5%	21	4.1%	0	0.0%
	LimitedAssn	312	12.9%	575	15.9%	405	18.0%	135	16.7%	35	6.8%	0	0.0%
	Assn	682	28.2%	852	23.6%	676	30.1%	170	21.0%	6	1.2%	0	0.0%
	HighAssn	410	17.0%	50	1.4%	46	2.0%	4	0.5%	0	0.0%	0	0.0%
	ID	725	30.0%	9	0.2%	8	0.4%	1	0.1%	0	0.0%	0	0.0%
Total responses		*2,417*		*3,615*		*2,244*		*810*		*518*		*43*	
Distinct QKsets		*107*		*162*		*100*		*36*		*24*		*2*	

Table 3. Distribution of conclusions: supporting data for Figure 4. Errors and incorrect conclusions are highlighted; neutral and debatable conclusions are shown in gray. (*Baseline Dataset*)

An alternative method of describing accuracy is through the use of posterior probabilities, which assess the proportions of responses for a given conclusion that are consistent with ground truth. Using this method, 98.8% of *ID* responses were consistent with ground truth (positive predictive value, PPV), as were 91.2% of *Excl* responses (negative predictive value, NPV). Further, 89.1% of *HighAssn* responses and 88.9% of *NonAssn* responses were consistent with ground truth. If we limit posterior probabilities to comparisons in which the Q and K were the same make, model, and size, the proportion of accurate *Excl* responses drops (NPV=83.2%), but PPV is almost unchanged (98.9%). These rates are notably affected by the proportions of mated vs. nonmated data; see *Appendix D5* for details and projections.

Note in Figure 4 and Table 3 that for both mated and nonmated QKsets, less than half of the responses were definitive conclusions. Although identification conclusions often garner the widest attention and scrutiny [1–3], *IDs* generally comprise a small portion of casework: in the participant background survey (*Appendix P1*), only 8% of the participants in this study indicated that they frequently make identifications, while 29% never make identifications. Instead, class associations and exclusions are much more common: 92% of participants indicated that they frequently make class associations, while 62% frequently report limited association of class and exclusion. In this study, there were as many class associations as definitive conclusions: in the *Baseline Dataset*, 39.7% of conclusions were definitive, 14.1% were probable conclusions, 40.1% were class associations, and 6.1% were neutral. For these reasons, any study seeking to assess the accuracy of FFE decisions needs to do so across the entire range of conclusions to fully characterize performance. Table 4 provides a summary of the same results shown in Table 3, but summarized by classification accuracy. For example, on mated QKsets, any conclusion from *ID* through *LimitedAssn* is consistent with ground truth and therefore could be considered a "correct" classification.

	Mated	*Nonmated*			
		Same make/ model/size	*Same make/model ± ½-1 size*	*Different make or model*	*Different foot*
Classification consistent with ground truth ("correct")	88.1%	91.4%	54.9%	77.8%	100.0%
Classification contradicting ground truth (error/incorrect)	7.8%	2.4%	0.6%	7.9%	0.0%
Neutral classification	4.1%	6.2%	6.8%	14.3%	0.0%
Debatable classification			37.7%		

Table 4. Classification accuracy: data from Table 3 summarized by classification accuracy. Highlights correspond to Table 3. (*Baseline Dataset*)

Accuracy and error can be measured using a variety of methods beyond those shown here: for further discussion, additional rates, and confidence intervals, see *Appendix D5*. For a more detailed breakdown of errors and incorrect conclusions, see *Appendix E*. The accuracy rates obtained in this study are consistent with those reported in the WVU Study [10–12] (see *Appendix M* for details).

4.2 Comparing Participants' Performance

Rates of accuracy and error were not evenly distributed among participants. Because of the use of a multilevel conclusion scale, FFE performance cannot be assessed using a single measure: performance is multidimensional and should consider rates of errors and incorrect conclusions (FPR, FNR, IAR, INR, shown in Figure 5) and rates of "correct" conclusions (TPR, TNR, CAR, CNR, shown in Figure 6). Each point in Figure 5 and Figure 6 represents a single participant and each is present in all four quadrants of each figure. The symbols and colors correspond between Figure 5 and Figure 6. Figure 5 shows the rates of conclusions contrary to ground truth, with error rates (FPR and FNR) in the top right quadrant, and the interactions with the

incorrect probable conclusion rates in the other quadrants. Figure 6 shows the corresponding rates of conclusions consistent with ground truth (TPR and TNR) in the top right quadrant, and the interactions with the "correct" probable conclusion rates in the other quadrants. When comparing examiner performance at an individual level, we limit analyses to the *Examiner Comparison Dataset*, a subset of the *Baseline Dataset* limited to the 71 participants who completed at least 40 QKsets (omitting 13 participants who completed 20-34 QKsets each because the number of trials per examiner was deemed insufficient for calculating meaningful individual rates).

For example, in the top right quadrant of Figure 5, the orange diamond represents the participant with the highest FPR (9%) but a FNR of 0%; following that point to the other quadrants, we see that participant also had the highest IAR (19%) but a 0% INR. In Figure 6, we see that participant had the highest TPR (58%) but a TNR slightly below average (37%), while CAR and CNR were both average. In other words, that participant had the highest rate of "correct" *IDs*, but that came at the cost of the highest rate of erroneous *IDs*; on *Excls*, that participant was more cautious, with an average rate of "correct" responses and no errors. Similarly, the participant with the highest rate of erroneous *Excls* (blue asterisk, FNR=39%) also had a rate of correct *Excls* well above average (TNR=65%). The open circles indicate low rates of errors and incorrect conclusions — but note that this should be considered in terms of the rates of correct conclusions: the black open circles had low error rates, but accomplished this by making few "correct" definitive conclusions; the blue open circles were better than average overall. One participant with a particularly strong performance (blue open circle in the top right quadrant of Figure 6) had nearly the highest TPR (50%) and TNR (74%), an FPR of 0%, a FNR below average (6%), and no incorrect conclusions (IAR and INR both 0%); the other participants with high rates of "correct" conclusions all had above average rates of errors or incorrect conclusions.

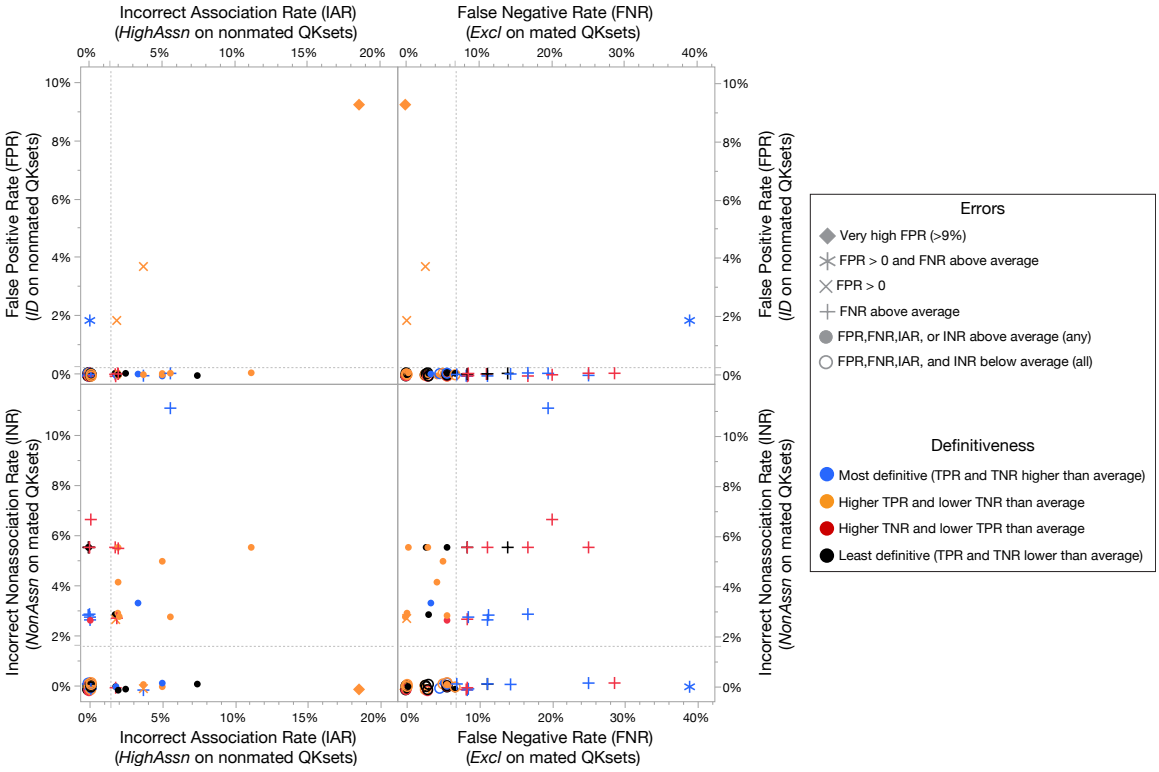

Figure 5. Comparison of participants by rates of errors and incorrect conclusions. Means are shown as dotted lines. Markers are jittered to minimize superimpositions. (*Examiner comparison dataset*. n=71 participants who completed at least 40 QKsets; rates calculated from 5,749 responses. Rates are calculated based on a mean of 32.6 mated and 48.6 nonmated QKsets per participant.)

Only four (out of 84) of the participants made any erroneous *ID* conclusions. A total of 11 erroneous IDs occurred in this study (nine in the *Baseline Dataset*, and an additional two occurred in the repeatability data, discussed below). One participant made six erroneous *IDs* (orange diamond mentioned previously in Figure 5), and another made three (orange X, 4% FPR). Both of these participants have at least five years of experience, conduct footwear examinations less than weekly, completed a formal training program lasting 6-12 months, are not IAI certified, work for a non-US government agency, have not completed a proficiency test in the past year, and infrequently or never report *IDs* in casework; a total of six participants meet these criteria, but none of the other four reported any erroneous *IDs*.

Erroneous *Excls* were not limited to a few individuals: 56 of the 84 participants made at least one erroneous *Excl* in the *Baseline Dataset*. A total of 151 erroneous *Excl*s occurred in this study (146 in the *Baseline Dataset*, and an additional five in the

repeatability data). One participant (blue asterisk mentioned previously in Figure 5) made 14 false negative errors. This participant has more than 10 years of experience, conducts a few footwear examinations yearly, has testified as a footwear expert, did not complete a formal training program, is not IAI certified, and infrequently or never reports *Excl*s in casework; a total of five participants meet these criteria, but none of the other exhibited high individual false negative rates (one reported just a single erroneous *Excl*, and the remaining three did not report any).

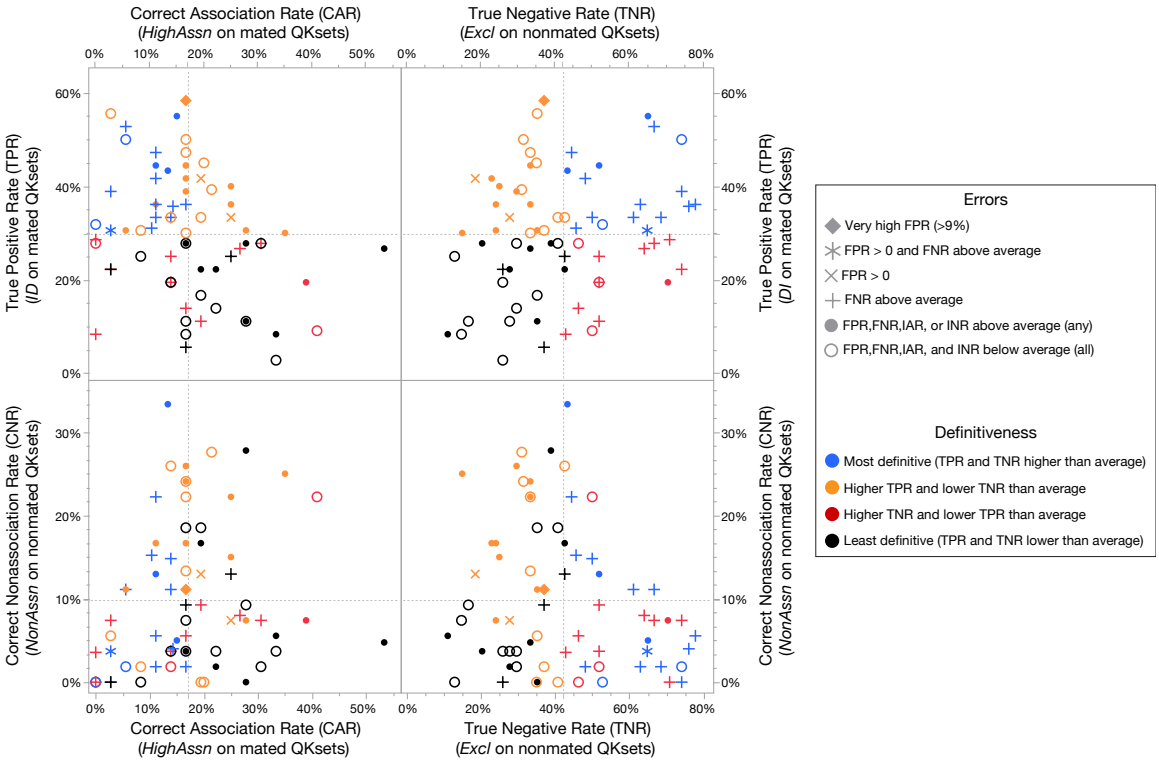

Figure 6. Comparison of participants by rates of correct conclusions. Means are shown as dotted lines. (Same 71 participants as Figure 5, with the same symbols and color-coding.)

In order to determine whether examiner performance was associated with the education, training, and experience of participants, we assessed 16 attributes from the background questionnaire with respect to the performance of the 71 participants to determine whether there was a detectable association between background and performance. For this purpose, we reduced the eight dimensions shown in Figure 5 and Figure 6 into four weighted performance ratios (described in detail in *Appendix I*). We evaluated the significance of associations using two complementary approaches: variable importance analysis (VIA) and attribute-specific significance testing, as described in [22]. VIA considers all variables simultaneously by leveraging both linear regression and random forest techniques, yielding importance scores; significance testing via the Kruskal-Wallis test was conducted for each attribute individually to assess for differences between groups, resulting in p-values and Benjamini-Hochberg q-statistics. Effect thresholds were set for each of these significance measures (importance scores, p-values, and q-statistics) and reporting criteria were developed to determine which (if any) of the background characteristics exhibited sufficient support to indicate an association with performance: an attribute that met the criteria for all three significance measures is considered a *notable association*; an attribute that met criteria for two of three significance measures is considered a *limited association*. See *Appendix J* for additional details.

For the vast majority of background attributes—including education, years of experience, examination frequency, and certification status—we did not detect support for an association with any of the four performance measures. We found a single notable association between an attribute and performance: participants employed by US local agencies generally reported a higher proportion of "correct" *Excls* and *NonAssns* than those employed by international government agencies; however, this can be attributed to differences in reporting tendencies, in that participants employed by US local agencies were more likely than those who work for international government agencies to report definitive conclusions ($p = 0.0012$) and less likely to report class associations ($p = 0.0427$), based upon a Kruskal-Wallis analysis. The only other association we found was a limited association between training and performance: participants with at least one year of formal training were generally less likely to make erroneous *Excl* and incorrect *NonAssn* conclusions than others, but this result should be interpreted with caution given that this attribute only meets the criteria for two of the three significance measures and we thus cannot preclude the possibility that this is a spurious effect. See *Appendix J* for additional analysis details and full results.

Prior to data collection, the team hypothesized that there could be performance impacts associated with participants' use of the QKset materials (printed vs. digital) and their use of software to aid them in their comparisons. According to the weighted TC-

CA ratio (as shown in *Appendix J2*), the rates of correct *IDs* and *HighAssns* were higher for those participants who reported using software when conducting their comparisons. Additionally, based upon Bonferroni-adjusted post-hoc analysis, participants who reported never using software during the study had lower rates of correct *Assns* than those who used software a majority of the time (p = 0.0023).

4.3 QKset-Specific Effects and Consensus

Figure 7 shows the same data as Figure 4, but delineates the distribution of conclusions by QKsets (each column represents a single QKset) as well as multiple methods of assessing consensus for each QKset — showing that the distribution of conclusions is strongly affected not just by mating, but by specific QKsets. With respect to conclusions consistent with ground truth, the proportion of *ID* conclusions on the 107 mated QKsets ranged from 0% to 100% (TPR), but nearly a third of mated QKsets yielded no *ID* conclusions; on 27% of the mated QKsets the majority of responses were *ID*. For the 162 nonmated QKsets, the proportion of *Excl* conclusions per QKset also ranged from 0% to 100% (TNR); only 6% of nonmated QKsets yielded no *Excl* conclusions, and the majority of responses were *Excl* on 35% of nonmated QKsets. (See *Appendix G* for additional details.)

Errors or incorrect responses were present on most mated QKsets: in the *Baseline Dataset*, 64 of the 107 mated QKsets had at least one erroneous *Excl* conclusion, and an additional seven QKsets had at least one incorrect *NonAssn*. Most nonmated QKsets had no errors or incorrect responses: in the *Baseline Dataset*, eight of the 162 nonmated QKsets had at least one erroneous *ID* conclusion, and an additional 31 QKsets had at least one incorrect *HighAssn*.

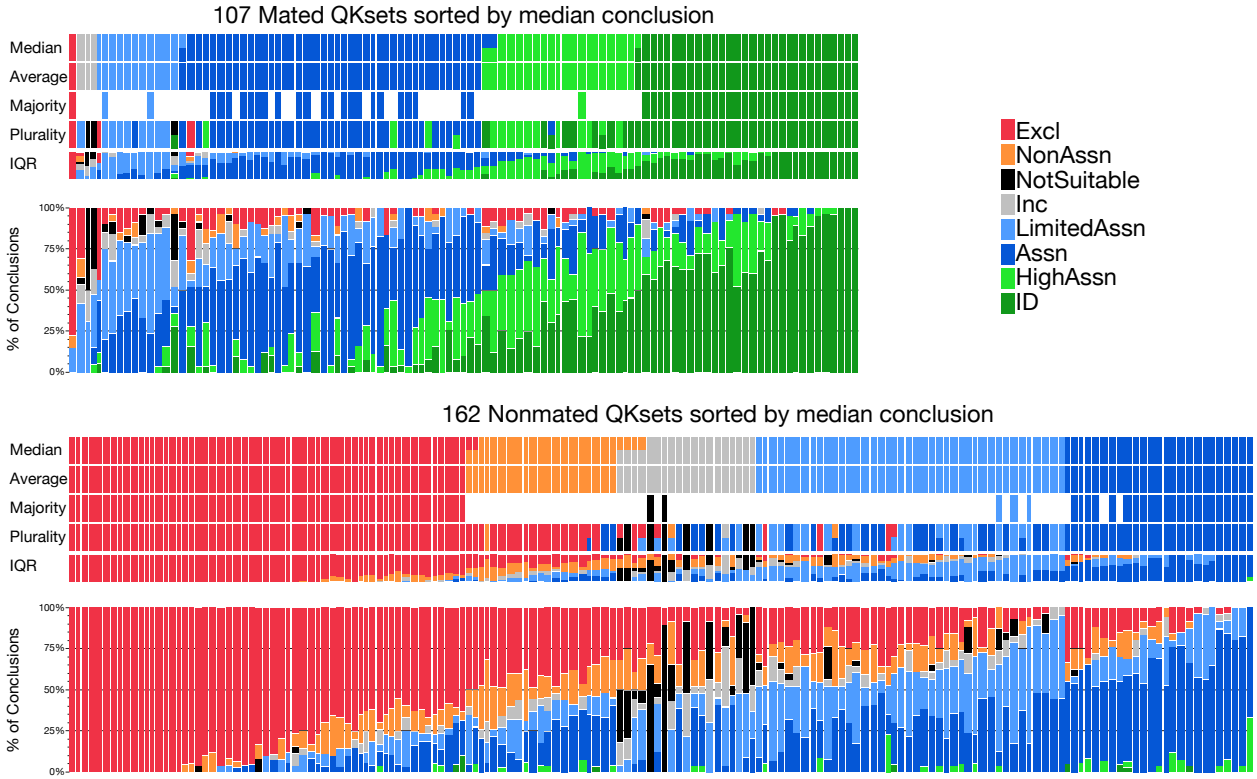

Figure 7. Decision rates and consensus conclusions (via median, average, majority, plurality, and interquartile range) for each QKset. QKsets (x axes) are sorted by median, then average conclusion. Median and average do not distinguish between *NotSuitable* and *Inc*. Conclusions were unanimous on three mated QKsets (shown as solid green columns at the right of the upper charts) and 17 nonmated QKsets (shown as solid red columns at the left of the lower charts). Conclusions had at least a 75% supermajority on 19 mated and 41 nonmated QKsets. (*Baseline Dataset*: n=6,032 responses; mean of 22.4 participants per QKset)

As we see in Figure 7, FFEs often do not agree on the conclusion for a given QKset. Ground truth provides a means of determining that a given conclusion is erroneous (or incorrect), but does not provide a means of determining which of the conclusions consistent with ground truth may be considered "appropriate" (which sometimes is referred to as a "forensically correct" response). There are currently no objective criteria to determine whether a given comparison justifies making a definitive conclusion, probable conclusion, class association, or neutral response. FFEs make this decision based on their training and expertise, considering factors such as the known and questioned impression quality, and the availability and degree of correspondence (or non-correspondence) of class, subclass, and/or randomly acquired characteristics. The multilevel conclusion scale provides a means of reporting conclusions along a continuum that accounts for the FFE's observations as well

as the FFE's perceived reliability of these observations. The various conclusions among FFEs on a given QKset can be considered as votes regarding whether there is sufficient support in the QKset to make a given decision. Among conclusions that are consistent with ground truth, the only practical mechanism we have to evaluate the appropriateness of a conclusion is consensus: the collective judgments of the experts provide a scale to evaluate the extent of support among FFEs that a given conclusion is justifiable.

Figure 7 shows several approaches for defining consensus. Majority provides an intuitive threshold for consensus, but is only defined in about half of both mated (55/107) and nonmated (84/162) QKsets. Median, average, and plurality conclusions all may be considered reasonable means of determining consensus when there is no majority, but note that the results differ among these approaches. Interquartile range (IQR) uses a different approach: rather than a point estimate of a single conclusion, any of the middle 50% of conclusions (between the 25th and 75th percentiles) are considered appropriate conclusions.

Consensus responses are usually but not always consistent with ground truth. The only QKset on which the majority of responses were erroneous was a mated QKset (QK213), which accounted for 20 FNs (77% FNR); there was consensus on this error for all five consensus methods (and met a 75% supermajority threshold as well). These results may indicate the difficulty of the substrate: the questioned impression in this case was on a terry cloth towel (Figure 8). There was not consensus on any other errors using the median or average, nor on any incorrect conclusions. Consensus conclusions assigned using plurality were never incorrect and were rarely erroneous: three mated QKsets had plurality conclusions of *Excl*. Similar to the other methods, the IQR range of conclusions was rarely erroneous or incorrect: IQR consensus ranges for four mated QKsets contained *Excl* errors, and one nonmated QKset had an IQR that contained *HighAssn*. Note that we do not show these various approaches to recommend a specific manner of assessing consensus, but rather to illustrate that it is not always apparent what the most appropriate conclusion is for a given QKset. For a more detailed discussion of these approaches see *Appendix G*.

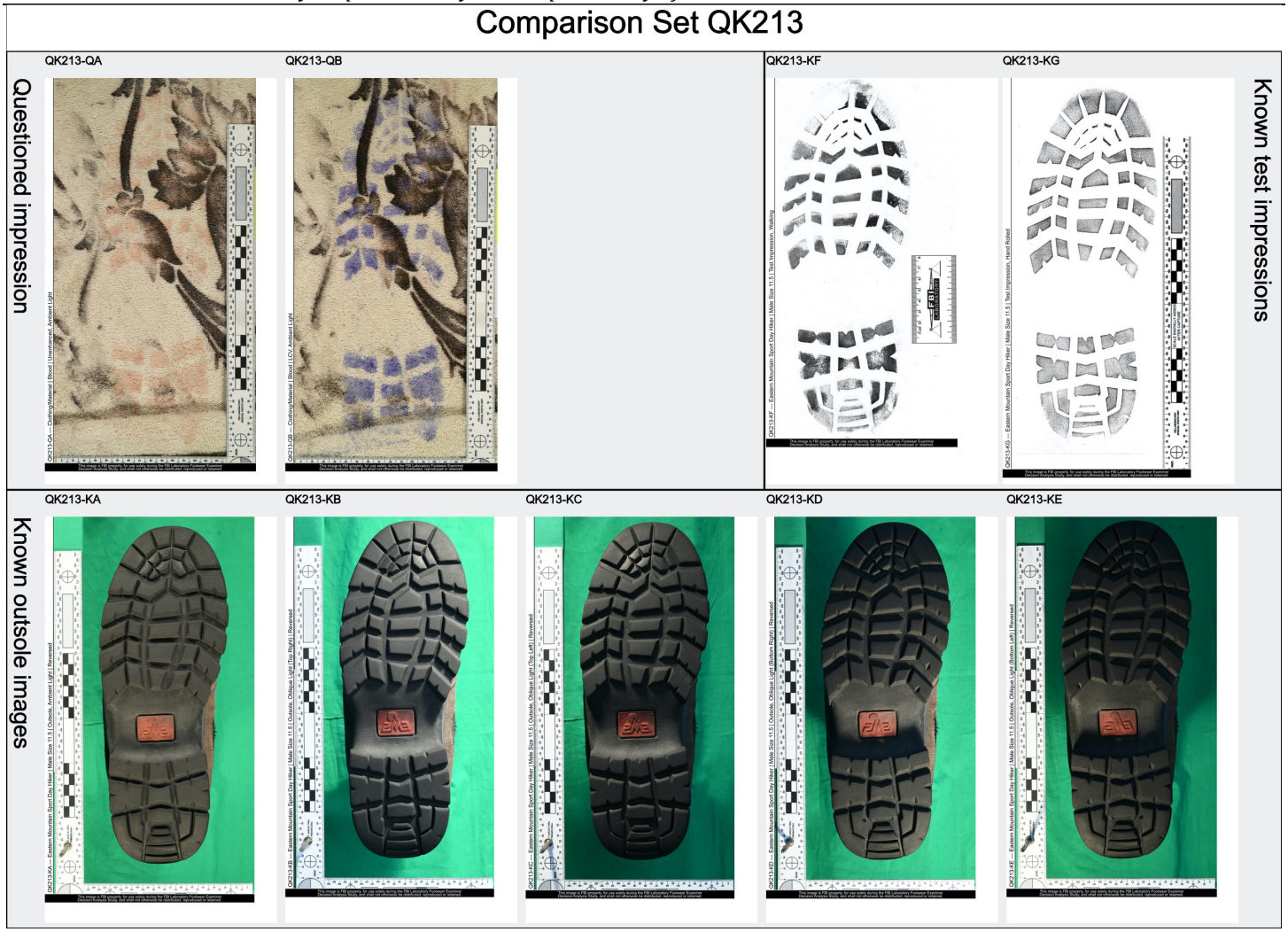

Figure 8. QK213: Mated QKset that resulted in 20 FNs. The questioned impression is synthetic blood on cloth (pre and post chemical enhancement). Conclusion rates for this QKset: 0 *ID*, 0 *HighAssn*, 0 *Assn*, 4 *LimitedAssn*, 0 *Inc*, 0 *NotSuitable*, 2 *NonAssn*, 20 *Excl*. Of the participants who concluded Excl, 2 reported that they observed differences in outsole design, and the remaining 18 reported differences in size between the questioned impression and known item of footwear. The majority of participants assessed the difficulty of this QKset as "easy," which may be reflective of this perceived lack of class correspondence. See *Appendix E3* for more detailed images.

4.4 Effects of Questioned Impression Quality

For the purposes of this study, we developed a method for assessing the quality of each questioned impression using a rubric that evaluates ten discrete impression attributes. This approach is detailed in [17], and summarized here in *Appendix F1*. These attributes assess the quantity of information reproduced, the clarity of that information, and any interferences such as distortion or overlapping impressions, among others. This approach results in a quality score of 0-20, which here we summarize by quintiles to an A-F quality grade (see Appendix F2 for additional information). Figure 9 details the distribution of conclusions reported as a function of quality grades.

Figure 9 shows that higher quality was associated with more definitive conclusions, fewer class associations, and fewer neutral responses. This association is particularly pronounced for mated trials (left), where as we move from F to A we see the proportions of *IDs* increase and the proportions of *Assns*, *LimitedAssns*, *Inc*, and *NotSuitable* all decrease. Nonmated trials on QKsets of the same make/model/size show a similar effect as we move from C to A, but for lower quality (F,D,C) the proportions of *Excls* (TNR) remain flat; however, increased quality reduced neutral responses and *LimitedAssns* and increased *Assns*. For nonmates with class differences, TNR increases notably with quality: for QKsets in which the Q and K differed by one size, almost every trial with a quality of A or B resulted in *Excl*; when make or model differed, almost every trial with a quality of C or better resulted in *Excl.* Note that trials on QKsets from different feet were almost always *Excl* even though those QKsets were of F quality. In short, rates of correct definitive conclusions are directly associated with the quality of the questioned impression and

the extent of class similarities/differences between the Q and K. These results indicate that the quality required for a comparison of class characteristics differs from the quality required for a comparison of source. For a more detailed discussion of the effects of questioned impression quality, see *Appendix F*.

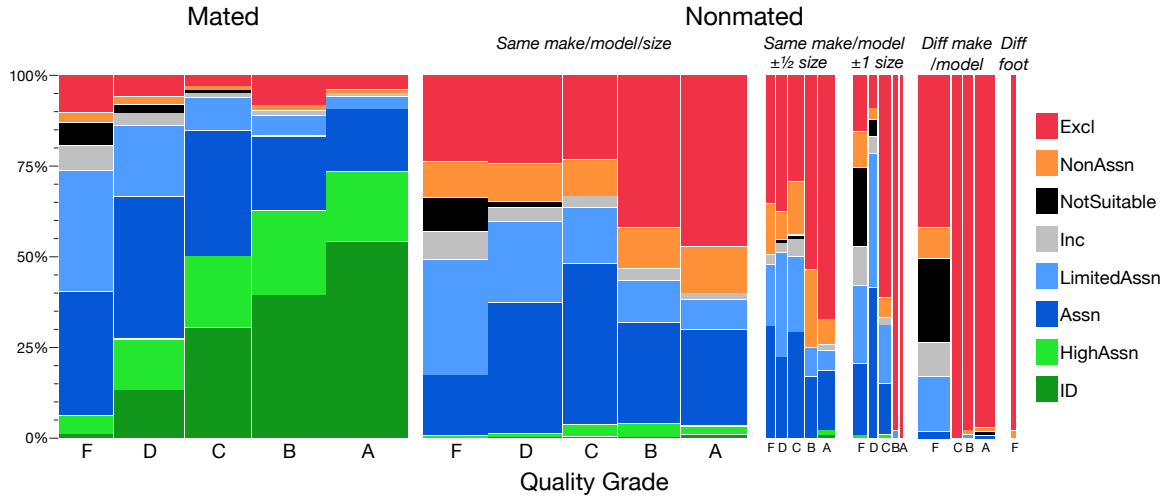

Figure 9. Association between quality grade and conclusions. (*Baseline Dataset*. Mated: 2,417 trials on 107 QKsets; Nonmated same make/model/size: 2,244 trials on 100 QKsets; same make/model, ±½ size: 483 trials on 21 QKsets; same make/model, ±1 size: 327 trials on 15 QKsets; diff make/model: 518 trials on 24 QKsets; diff foot: 43 trials on 2 QKsets)

4.5 Reproducibility

Reproducibility refers to inter-examiner consistency: the extent to which responses from different participants agree when given the same QKset. To assess reproducibility we use the *Reproducibility Dataset*, which is a self join of the *Baseline Dataset*: each individual response is paired with every other response on the same QKsets, resulting in 132,074 inter-examiner decision pairs derived from the 6,032 responses in the *Baseline Dataset*.

Figure 10 summarizes the reproducibility of conclusions, based on all pair-wise combinations of responses from different participants on the same QKsets. The y-axis is associated with responses by any single participant, whereas the x-axis is associated with responses by all other participants (on the same QKsets). The height of each row in this figure is proportionate to the number of responses in each conclusion category. For example (in the bottom row of the top chart in Figure 10), for every participant who responded *ID* on a mated QKset, 62% of the other participants also responded *ID*, 19% responded *HighAssn*, and 10% responded *Assn*.

Overall, 44% of conclusions were reproduced exactly, and 70% of conclusions were reproduced within one conclusion level. Note, however, that the only conclusions reproduced by a majority of participants were *IDs* in mated trials and *Excls* in nonmated trials. Other than *Assn*, no other conclusions were reproduced by even a plurality of other participants: for example, for each *LimitedAssn* response, the other participants were more likely to conclude *Assn* than *LimitedAssn*. Some amount of variability in reproducibility may be expected due to the multilevel conclusion scale: the psychology literature has shown that the use of categorical conclusion scales, and the number of categories in those scales, have notable effects on the measurement of reproducibility (e.g., [23,24]); see [25] for discussion specific to forensic conclusions. In general, a higher number of categories in a decision scale reduces the probability that responses will be identical, but also provides a finer level of granularity in assessing those differences. Because the conclusion scale includes multiple levels and is intended to be used as a continuum by examiners, it is expected that FFEs may vary in reporting conclusions. In other words, even if two examiners observe the same features in correspondence/non-correspondence, they may assign different strengths to these observations based upon factors such as their training and experience.

Some errors and incorrect conclusions were reproduced. With respect to mated QKsets, 19.9% of *FNs* and 3.3% of *INs* were reproduced; overall, 22.4% of *FNs* or *INs* were reproduced as either an *Excl* or *NonAssn* decision by a second examiner. Over half of the reproduced *FNs* or *INs* occurred on a single QKset (QK213, depicted in Figure 8 and detailed in *Appendix E3*, Fig S5), and most of the remainder occurred on four other QKsets. For nonmated QKsets, 1.1% of *FPs* and 7.9% of *IAs* were reproduced; overall, 7.1% of *FPs* or *IAs* were reproduced as either an *ID* or *HighAssn* decision by a second examiner.

The reproducibility of conclusions is associated with the participants' assessments of difficulty: overall, the reproducibility of conclusions tends to decrease as difficulty increases. In particular, for correct definitive conclusions (*IDs* in mated trials and *Excls* in nonmated trials), there is a substantial decrease in reproducibility as difficulty increases. (See *Appendix H3* for details)

Suitability assessments showed notably low reproducibility: most assessments of *NotSuitable* were not reproduced. Overall, 3% of trials in the *Baseline Dataset* resulted in a decision of *NotSuitable* — only 31% of *NotSuitable* decisions were reproduced by a second examiner. If we do not distinguish between *NotSuitable* and *Inc*, only 28% of neutral responses were reproduced by a second examiner. For additional discussion and detailed reproducibility results, see *Appendix H*.

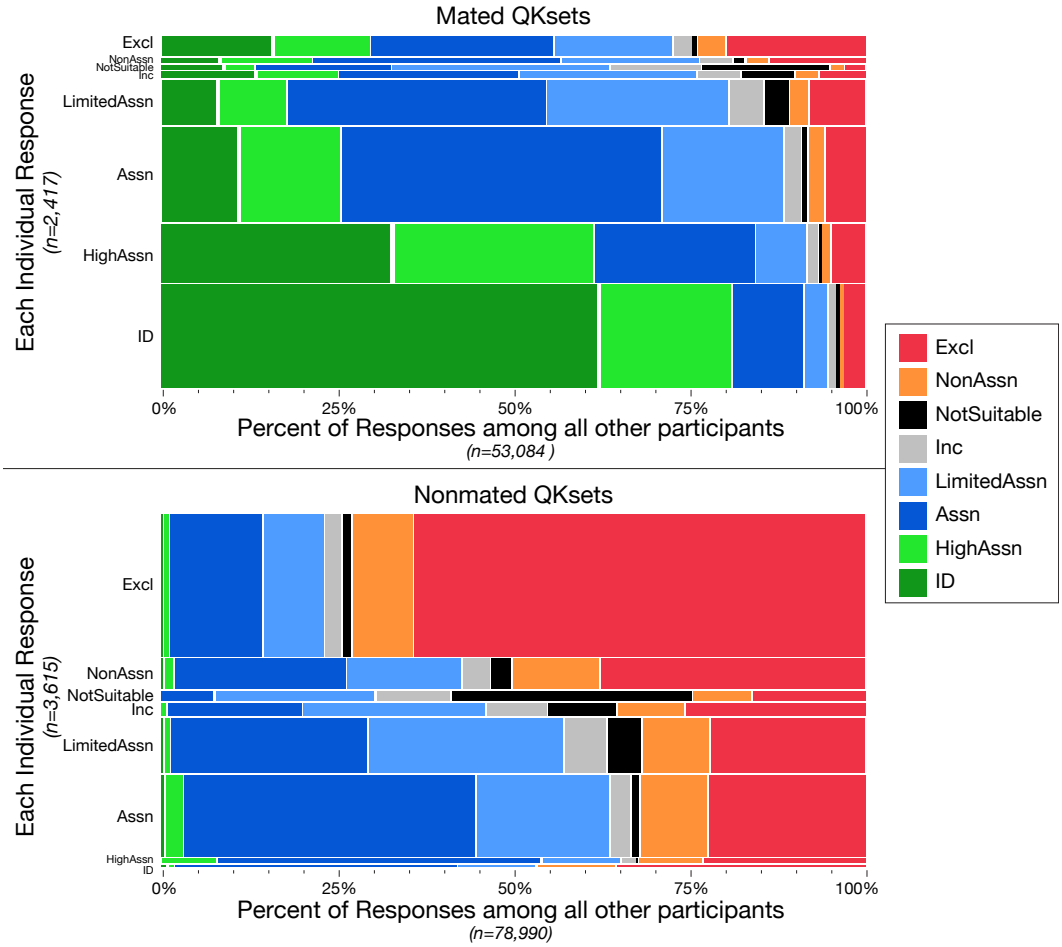

Figure 10. Reproducibility of participants' conclusions illustrated using mosaic displays of contingency tables. All responses in the *Baseline Dataset* are shown as rows; the x axis and color-coding show the proportions of each type of response among all other participants (132,074 inter-examiner decision pairs derived from the 6,032 responses in the *Baseline Dataset*.)

4.6 Repeatability

Repeatability refers to intra-examiner consistency: the extent to which responses from the same participant agree when given the same QKset. To evaluate repeatability, each participant was assigned ten QKsets that contained the same imagery as QKsets that were assigned earlier. The first and second assignments used different QKset numbers, but were otherwise identical. These repeats were assigned with at least 20 intervening QKsets, so that the participants encountered the repeated set weeks or months after completing the first assignment. To lessen the chance of recognizing the repeated assignments, the repeats only included comparisons of the same make, model, and size.

Repeatability was assessed by comparing the first and second responses reached on a given QKset, as shown in Figure 11. Overall, 60% of conclusions were repeated exactly, and 79% of conclusions were repeated within one conclusion level. In mated trials *IDs*, *HighAssns*, and *Assns* were repeated in a majority of trials; in nonmated trials *Excls* and *Assns* were repeated in a majority of trials. Only 0.7% of all repeated QKsets (four repeated trials) were contradictions (*ID* vs. *Excl*), all of which occurred on mated QKsets. A single *FN* was repeated, but there were no repeated *INs*. No *FPs* were repeated although one decision of *HighAssn* (*IA*) resulted in an *ID* (*FP*) on the second assignment. Only two *NotSuitable* decisions were reported in the repeatability data, and neither was repeated. Examiners often changed their assessments of Design, Mold, Size, or Wear (49% of repeated trials), especially when they changed conclusions (of the trials in which conclusions were not repeated, 66% also had changed

assessments of Design, Mold, Size, or Wear). Examiners also often changed their assessments of difficulty (48%). For additional discussion and detailed repeatability results, see *Appendix H*.

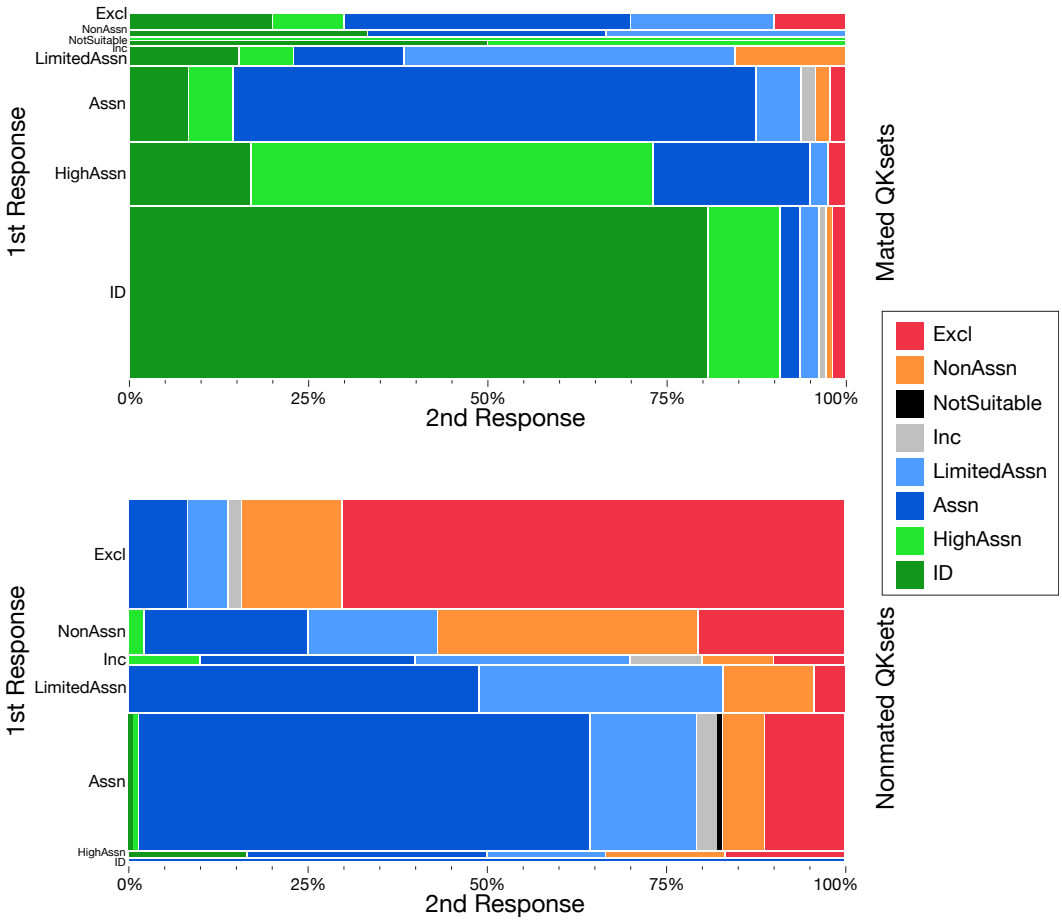

Figure 11. Repeatability of conclusions illustrated using mosaic displays of contingency tables. (*Repeat Dataset*: 1,156 responses (578 responses to 1st assignments; 578 responses to 2nd assignments) from 64 examiners on 30 QKsets)

5 Additional Results

The study focused on assessing accuracy, reproducibility, and repeatability, but various other assessments were also performed, the results of which are summarized here and reported in detail in the appendices.

The study included some QKsets from unused footwear items, to assess how results were affected by items without RACs or wear. Three mated QKsets ("new-used") included a Q from an unused footwear item and the K (from the same footwear item) after it was worn for up to two weeks. Three nonmated QKsets ("new-new") included a Q from an unused footwear item and a K was from a different unused footwear item (of the same make, model, and size). The "new-used" mated QKsets resulted in a 27% TPR and 27% CAR: although these responses are consistent with ground truth, these conclusions appear to have been inappropriately based on the correspondence of manufacturing artifacts. The "new-new" nonmated QKsets resulted in a 14% IAR and one FP, which likewise appear to have been inappropriately based on the correspondence of sub-class characteristics imparted during manufacture. See *Appendix K* for additional details.

The study included some QKsets in which the footwear items were worn for up to two weeks between the Q and K. Wear between the Q and K increased the relative proportion of class associations and decreased the proportion of definitive and probable conclusions (as would be expected). See *Appendix L* for additional details.

As part of the comparison process, participants were asked to assess whether the Q and K correspond in design, size, mold, and wear (indicating "same," "different," or "unsure" for each). These assessments can be evaluated as incorrect if they contradict the actual ground truth similarities and differences of class characteristics between the Qs and Ks in the QKsets. On nonmated QKsets, assessments of whether the Q and K were of the same design or size were often incorrect. When the Q and K were the same make, model, size, and foot, 15% of responses incorrectly indicated that the design, mold, size, or foot differed; when the Q and K were of different makes or models, 23% of responses incorrectly indicated that the design and mold were the same (see

Appendix N1, Table S31). Out of the erroneous *Excls* on mated QKsets (FNs), 49% incorrectly indicated differences in design, size, or mold; an additional 29% indicated incorrect differences in wear (on QKsets where the Q and K were collected without intervening wear), for a total of 78%. Out of the *NonAssns* on mated QKsets (INs), 16% incorrectly indicated differences in design, size, or mold; an additional 23% indicated incorrect differences in wear, for a total of 40%. These assessments can also be used to help illuminate the basis for some errors and incorrect conclusions. See *Appendix N* for additional details.

A variety of additional factors were assessed that had had relatively minor results. See *Appendix O* for results with respect to limitations indicated by participants during comparison, effects of collection attributes for questioned impressions, participants' use of printed materials, and the typicality of study samples and overall difficulty of the study.

6 Discussion and Conclusions

The accuracy and reliability of FFEs are of interest to both the legal and forensic science communities. Here we evaluated decisions made by practicing FFEs on items and tasks designed to resemble operational casework. The results reported here are in close alignment with other research targeted at evaluating the accuracy and reliability of FFE conclusions (see *Appendix M*).

When participants reported definitive conclusions, 98.8% for *IDs* were consistent with ground truth (PPV) and 91.2% of *Excls* were consistent with ground truth (NPV). Erroneous *IDs* (*FPs*) were rare and limited to a few participants: only four participants made erroneous *IDs,* for an overall FPR of 0.2%. One examiner made over half of all *FPs* in the study. Erroneous *Excls* (*FNs*) were more prevalent and more broadly distributed among FFEs: 56 of the 84 participants made at least one erroneous *Excl*, for an overall FNR of 6.0%. In considering the accuracy of definitive conclusions, note that less than half of the responses in the study were definitive conclusions. This is not a defect but reflects the purposes and capabilities of the footwear examination process, in which probable conclusions and class associations may be appropriate responses to many comparisons. Source attribution (i.e., a conclusion of *ID*) should not be considered the primary goal of comparisons: an *ID* decision is an examiner's determination that 1) the compared items (impressions and footwear) correspond in class characteristics, 2) contain sufficient identifying characteristics (features specific to the items being compared) to differentiate them from other items with the same class characteristics, and 3) those identifying characteristics are in sufficient correspondence to conclude that the known footwear is the source of the questioned impression. Footwear or impressions without sufficient observable identifying characteristics (e.g., unworn shoes without RACs, or low quality impressions without sufficient detail to reliably observe RACs) do not provide a basis for an *ID* decision. The sequential comparison method employed by FFEs during their comparisons (see *Appendix B*) is designed to elicit discriminating features between the Q and the K that can lead the examiner to an *Excl*. Only when an FFE is unable to discriminate between the items being compared based on their class characteristics, and they observe sufficient corresponding RACs, is it appropriate for the examiner to report an *ID*. These conditions present themselves infrequently during operational casework: according to the participants, only 7 (8%) of them reported that they frequently provided *IDs*, and the remaining 92% reported they do so either infrequently or never. (*Appendix P1*, question 12).

Given that a few FFEs have notably high error rates, agencies may wish to consider additional quality assurance measures to assess their FFEs' performance. This may be addressed through a combination of training, competency testing, proficiency testing, and technical review and/or blind verification. Addressing errors may require a two-pronged approach. Based on this study, erroneous *IDs* are rare and were disproportionately caused by a few FFEs: these may be addressed by detecting the individuals with high error rates and having processes in place to detect the (rare) *FPs* that occur. Erroneous *Excls* are much more broadly distributed among FFEs, indicating a need for further research to understand the causes of erroneous *Excls* in detail, and a need to address the issue in training.

Although this study did not evaluate verification or technical review by a second examiner (which should be conducted operationally for quality assurance), the reproducibility results serve to model the extent to which ***blind*** verification would result in the same or different conclusions. In the background questionnaire (*Appendix P1*, questions 22 and 23), 98% of participants indicated that their employers require review by a second examiner (which could be described as technical review or verification), but only 31% require blind verification (which is performed by a second FFE who does not know the primary FFE's conclusions). The reproducibility rates indicate that two FFEs would be expected to report conclusions that agree exactly about 43% of the time, or within a single conclusion category approximately 70% of the time. Disagreements between examiner conclusions (if conducted as part of a documented, transparent conflict resolution process) would aid in detecting errors or incorrect conclusions and may allow them to be properly rectified prior to final reporting. However, 20% of erroneous *Excls* and 1% of erroneous *IDs* were reproduced in this study, which suggests that blind verification may not be expected to detect all errors. The FFE community may wish to use the low level of reproducibility of conclusions as a basis for reevaluating its conclusion scale.

Suitability assessments showed strikingly low reproducibility: only 31% of *NotSuitable* decisions were reproduced by a second examiner. The forensic footwear discipline might benefit from standardization of suitability assessments. The method for characterizing questioned footwear impression quality that was developed for this study ([17], summarized in *Appendix F1*)

could provide a basis for more consistency in assessing the suitability of questioned impressions. The results show that the quality required for a comparison of class characteristics differs from the quality required for a comparison of source.

The rates measured in this study are intended to serve as overall assessments to inform decision making and guide future research. They should not be taken to be precise measures of operational accuracy and error rates. We show here that accuracy varies among FFEs and is affected by the specific items considered: overall rates therefore cannot be assumed to apply precisely to a given examiner on a given comparison. The results may not be representative of all FFEs or casework, and do not account for operational quality assurance measures such as verification or technical review. The examination procedure and conclusion scale used for this study may depart from what is used by FFEs. Participants were not provided the physical footwear items for evaluation.

Acknowledgments

We thank the footwear examiners who participated in this study and the volunteers who loaned or donated shoes and boots, as well as Scott Roth, Matt Eichler, Andrew Isett, and Gene Ariani for software/website development; Madeline Ausdemore for developing the performance-attribute model used in assessing participants (detailed in [22]); Michelle Luby, Katherine Ky, Kim Dang, Paige Riley, Colbey Ryman, Tom Kopczynski, and Jocelyn Abonamah for data collection and curation; Matt Marvin of Ron Smith & Associates for providing shoes and outsole images; Mike Gorn and Eric Gilkerson for FFE expertise and reviewing of study samples; and Jason Carvalho and Derrick Spearman for coordinating the mailing of packets. This is publication number 22.11 of the FBI Laboratory Division. Names of commercial manufacturers are provided for identification purposes only and inclusion does not imply endorsement of the manufacturer or its products or services by the FBI. This work was funded by the FBI Laboratory Division; Ideal Innovations and Noblis were funded under a contract award to Ideal Innovations Inc. from the FBI Laboratory. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the U.S. Government.

Contributor Roles (CRediT)—Conceptualization: RAH,BM,CP,JL,MS,JB,RP,BE; Methodology: RAH,BM,CP,JL,NR,MS,JB; Validation: RAH,BM,NR,MS; Formal analysis: RAH,NR,MS; Investigation/Data Curation: RAH,BM,CP,JL; Resources/Sample collection: RAH,BM,CP,JL,JB,RP,BE; Writing - Original Draft: RAH,BM,CP,JL,NR; Writing - Review & Editing: All; Visualization: RAH,NR; Supervision: RAH,BM,CP,RP,GP,BE; Project administration: RAH,BM,RP,GP,BE; Funding acquisition: RP,GP,BE.

References

[1] National Research Council, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington, D.C., 2009.

[2] President's Council of Advisors on Science and Technology (PCAST), Report to the President. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Executive Office of the President, Washington, D.C., 2016.

[3] President's Council of Advisors on Science and Technology, An Addendum to the PCAST Report on Forensic Science in Criminal Courts, Executive Office of the President, Washington, D.C., 2017.

[4] Scientific Working Group for Shoeprint and Tire Tread Evidence (SWGTREAD), Guide for the Examination of Footwear and Tire Impression Evidence, (2006). https://www.nist.gov/system/files/documents/2016/10/26/swgtread_08_guide_for_the_examination_of_footwear_and_tire_impression_evidence_200603.pdf.

[5] International Association for Identification (IAI), Footwear Certification Process, Requirements & Qualifications, (2021).

[6] Collaborative Testing Services Forensic Testing Program, Footwear Imprint Evidence Test No. 18-5331/2/5 Summary Report, (2018).

[7] Collaborative Testing Services Forensic Testing Program, Footwear Imprint Evidence Test No. 19-5331/2/5 Summary Report, (2019). https://cts-forensics.com/reports/19-5331.2.5_Web.pdf.

[8] Collaborative Testing Services Forensic Testing Program, Footwear Imprint Evidence Test No. 20-5331/5 Summary Report, 2020.

[9] J. Raymond, P. Sheldon, Standardizing Shoemark Evidence- An Australian and New Zealand Collaborative Trial, J. Forensic Identif. 65 (2015) 868–883.

[10] J.A. Speir, N. Richetelli, L. Hammer, Forensic Footwear Reliability : Part I — Participant Demographics and Examiner Agreement, J. Forensic Sci. (2020). https://doi.org/10.1111/1556-4029.14553.

[11] N. Richetelli, L. Hammer, J.A. Speir, Forensic Footwear Reliability : Part II — Range of Conclusions , Accuracy , and Consensus, J. Forensic Sci. (2020). https://doi.org/10.1111/1556-4029.14551.

[12] N. Richetelli, L. Hammer, J.A. Speir, Forensic Footwear Reliability : Part III — Positive Predictive Value , Error Rates , and Inter-Rater Reliability, J. Forensic Sci. (2020). https://doi.org/10.1111/1556-4029.14552.

[13] United States Department of Justice, Uniform Language for Testimony and Reports (ULTR) for the Forensic Footwear Discipline, (2020). https://www.justice.gov/olp/page/file/1284771/download (accessed November 6, 2020).

[14] H. Majamaa, A. Ytti, Survey of the conclusions drawn of similar footwear cases in various crime laboratories, Forensic Sci. Int. 82 (1996) 109–120. https://doi.org/10.1016/0379-0738(96)01972-X.

[15] Y. Shor, S. Weisner, A Survey on the Conclusions Drawn on the Same Footwear Marks Obtained in Actual Cases by Several Experts Throughout the World, J. Forensic Sci. 44 (1999) 14468J. https://doi.org/10.1520/JFS14468J.

[16] L. Hammer, K. Duffy, J. Fraser, N. Daeid, A study of the variability in footwear impression comparison conclusions, J. Forensic Identif. 63 (2013) 205–218.

[17] B. McVicker, C.L. Parks, J. LeMay, B. Eckenrode, R.A. Hicklin, A Method for Characterizing Questioned Footwear

Impression Quality, J. Forensic Identif. 71 (2021) 205–217.

[18] Scientific Working Group for Shoeprint and Tire Tread Evidence (SWGTREAD), Range of Conclusions Standard for Footwear and Tire Impression Examinations, (2013). https://treadforensics.com/images/swgtread/standards/current/swgtread_10_conclusions_range_201303.pdf.

[19] Scientific Working Group for Shoeprint and Tire Tread Evidence (SWGTREAD), Standard Terminology for Expressing Conclusions of Forensic Footwear and Tire Impression Examinations, (2006).

[20] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Accuracy and reliability of forensic latent fingerprint decisions, Proc. Natl. Acad. Sci. U. S. A. 108 (2011). https://doi.org/10.1073/pnas.1018707108.

[21] OSAC Human Factors Committee, Human Factors in Validation and Performance Testing of Forensic Science (OSAC Technical Series 0004), 2020. https://doi.org/https://doi.org/10.29325/OSAC.TS.0004.

[22] R.A. Hicklin, K.R. Winer, P.E. Kish, C.L. Parks, W. Chapman, K. Dunagan, N. Richetelli, E.G. Epstein, M.A. Ausdemore, T.A. Busey, Accuracy and Reproducibility of Conclusions by Forensic Bloodstain Pattern Analysts, Forensic Sci. Int. 325 (2021). https://doi.org/https://doi.org/10.1016/j.forsciint.2021.110856.

[23] L.M. Lozano, E. García-Cueto, J. Muñiz, Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales, Methodology. 4 (2008) 73–79. https://doi.org/10.1027/1614-2241.4.2.73.

[24] D. V. Cicchetti, D. Shoinralter, P.J. Tyrer, The Effect of Number of Rating Scale Categories on Levels of Interrater Reliability : A Monte Carlo Investigation, Appl. Psychol. Meas. 9 (1985) 31–36. https://doi.org/10.1177/014662168500900103.

[25] R.A. Hicklin, B.T. Ulery, M. Ausdemore, J. Buscaglia, Why do latent fingerprint examiners differ in their conclusions?, Forensic Sci. Int. 316 (2020). https://doi.org/10.1016/j.forsciint.2020.110542.

[26] AAFS Standards Board, ASB Technical Report 097: Terminology Used for Forensic Footwear and Tire Evidence, (2019).

[27] J.M. Butler, H. Iyer, R. Press, M.K. Taylor, P.M. Vallone, S. Willis, DNA Mixture Interpretation: A NIST Scientific Foundation Review (NISTIR 8351-DRAFT), 2021. https://doi.org/10.6028/NIST.IR.8351-draft.

[28] S.A. Cole, Is Fingerprint Identification Valid? Rhetorics of Reliability in Fingerprint Proponents' Discourse, Law Policy. 28 (2006) 109–135. https://doi.org/10.1111/j.1467-9930.2005.00219.x.

[29] W.J. Bodziak, Footwear impression evidence: Detection, recovery and examination, 2nd Editio, CRC Press, Boca Raton, 2017. https://doi.org/10.1201/9780203755587.

[30] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Repeatability and reproducibility of decisions by latent fingerprint examiners, PLoS One. 7 (2012). https://doi.org/10.1371/journal.pone.0032800.

[31] Scientific Working Group for Shoeprint and Tire Tread Evidence (SWGTREAD), Guide for the Forensic Documentation and Photography of Footwear and Tire Impressions at the Crime Scene, (2006).

[32] M. Marvin, A look at close non-match footwear examinations, in: Int. Assoc. Identif. Conf., 2015.

[33] Scientific Working Group for Shoeprint and Tire Tread Evidence (SWGTREAD), Guide for Lifting Footwear and Tire Impression Evidence, (2007).

[34] Scientific Working Group for Shoeprint and Tire Tread Evidence (SWGTREAD), Guide for the Chemical Enhancement of Bloody Footwear and Tire Impression Evidence, (2008).

[35] B.T. Ulery, R.A. Hicklin, M.A. Roberts, J. Buscaglia, Measuring what latent fingerprint examiners consider sufficient information for individualization determinations, PLoS One. 9 (2014). https://doi.org/10.1371/journal.pone.0110179.

[36] R.A. Hicklin, B.T. Ulery, T.A. Busey, M.A. Roberts, J.A. Buscaglia, Gaze behavior and cognitive states during fingerprint target group localization, Cogn. Res. Princ. Implic. 4 (2019). https://doi.org/10.1186/s41235-019-0160-9.

[37] OSAC Human Factors Committee, Draft Guidance on Testing the Performance of Forensic Examiners, 2018. https://www.nist.gov/system/files/documents/2018/05/21/draft_hfc_guidance_document-may_8.pdf.

[38] H. Tobi, P.B. van den Berg, L.T. de Jong-van den Berg, Small proportions: what to report for confidence intervals?, Pharmacoepidemiol. Drug Saf. 14 (2005). https://doi.org/10.1002/pds.1081.

[39] ISO/IEC, Information Technology - Biometric Sample Quality - Part 1: Framework (ISO/IEC 29794-1:2016), (2016). http://webstore.ansi.org/RecordDetail.aspx?sku=INCITS%2FISO%2FIEC+29794-1-2010.

[40] S. Byrne, A note on the use of empirical AUC for evaluating probabilistic forecasts, Electron. J. Stat. 10 (2016) 380–393. https://doi.org/10.1214/16-EJS1109.

[41] M. Drury, The Probabilistic Interpretation of AUC, Scatt. Smoothers. (2017). http://madrury.github.io/jekyll/update/statistics/2017/06/21/auc-proof.html (accessed July 19, 2022).

[42] A. Gossman, Probabilistic Interpretation of AUC, 0-Fold Cross-Validation. (2018). http://www.alexejgossmann.com/auc/ (accessed July 19, 2022).

[43] D.P. Chakraborty, New developments in observer performance methodology in medical imaging., Semin. Nucl. Med. 41

(2011) 401–418. https://doi.org/10.1053/j.semnuclmed.2011.07.001.

[44] J. Hair, W. Black, B. Babin, R. Anderson, R. Tatham, Multivariate Data Analysis, 7th ed., Prentice Hall, Upper Saddle River, 2014.

[45] U. Grömping, Variable Importance Assessment in Regression: Linear Regression versus Random Forest, Am. Stat. 63 (2009) 308–319. https://doi.org/10.1198/tast.2009.08199.

[46] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, Variable selection using random forests, Pattern Recognit. Lett. 31 (2010) 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014.

[47] S. Schmidt, Encyclopedia of Research Design, in: SAGE Publications, Inc., Thousand Oaks, 2010. https://doi.org/10.4135/9781412961288.

[48] Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, J. R. Stat. Soc. Ser. B. 57 (1995) 289–300.

[49] W.S. Noble, How does multiple testing correction work?, Nat. Biotechnol. 27 (2009) 1135–1137. https://doi.org/10.1038/nbt1209-1135.

[50] K.L. Gwet, Handbook of Inter-Rater Reliability, 4th ed., Advanced Analytics, LLC., Gaithersburg, MD, 2014.

Final accepted manuscript, subsequently published as
Austin Hicklin, R., McVicker, B. C., Parks, C., LeMay, J., Richetelli, N., Smith, M., Buscaglia, J., Perlman, R. S., Peters, E. M., & Eckenrode, B. A. (2022). Accuracy, reproducibility, and repeatability of forensic footwear examiner decisions. *Forensic Science International 339*, 111418. https://doi.org/10.1016/j.forsciint.2022.111418

Accuracy, Reproducibility, and Repeatability of Forensic Footwear Examiner Conclusions

Supplemental Information — Appendices

Contents

Other Supplementary Materials for this manuscript include the following:

Data S1: Response Data (spreadsheet containing comparison responses, deidentified survey responses, and metadata for each QKset)

Appendix A Glossary/Acronyms

This section defines terms and acronyms as they are used in this appendix. Definitions taken verbatim from "ASB Technical Report 097: Terminology Used for Forensic Footwear and Tire Evidence" [26] *are prefixed "ASB." Definitions from "ULTR for the Forensic Footwear Discipline"* [13] *are prefixed "ULTR."*

Accuracy	General term used to refer to the extent to which a conclusion is consistent with (or contradicts) ground truth.
Class Association	A response of *Assn* or *LimitedAssn.*
Class Characteristics	A feature/design element that is shared by two or more footwear items. The footwear outsole design and physical size are two common class characteristics. General wear of the outsole is also a class characteristic. ASB: "A feature shared by two or more items of footwear or tires. The footwear outsole or tire tread design and the physical size features of a footwear outsole or tire tread are two common manufactured class characteristics. General wear of the outsole or tire tread is also a class characteristic. Class characteristics establish membership within a specific group." ULTR: "A feature that is shared by two or more footwear items."
Conclusion Scale	A standardized set of conclusions used when rendering an opinion for a QK comparison. The scale is typically produced by an authoritative body and generally accepted by discipline practitioners.
Correct Association (CA)	A conclusion of *HighAssn* on a mated QKset.
Correct Non-association (CN)	A conclusion of *NonAssn* on a nonmated QKset.
Definitive Conclusion	A response of *ID* or *Excl.*
Error	A conclusion that is verifiably false (e.g., identification of non-mates).
False Negative (FN)	An erroneous conclusion of *Excl* on a mated QKset.
False Negative Rate (FNR)	The proportion of mated responses that resulted in false negatives.
False Positive (FP)	An erroneous conclusion of *ID* on a nonmated QKset.
False Positive Rate (FPR)	The proportion of nonmated responses that resulted in false positives.
Ground Truth	Definitive knowledge that a QK comparison is either mated or nonmated.
IAI	International Association of Identification.
Incorrect Association (IA)	An incorrect conclusion of *HighAssn* on a nonmated QKset.
Incorrect Non-association (IN)	An incorrect conclusion of *NonAssn* on a mated QKset.

Neutral Response	A response of *Inc* or *NotSuitable.*
Matrix	The material acquired by an item of footwear that is subsequently transferred to a substrate upon contact. The matrix can be either wet or dry in origin (e.g., blood or dust, respectively).
Outsole	The bottom portion of the footwear that comes into contact with the ground. ASB: "The bottom portion of the footwear that comes into contact with the substrate."
Outsole Design	ASB: "A specific pattern or arrangement of elements on an outsole typically associated with a manufacturer and having a name and/or style number." ULTR: "The manufactured pattern on the bottom of a footwear item."
Physical Size	The dimensions, shape, spacing and relative positions of the outsole design elements. Physical size is not synonymous with the manufacturer's shoe size. ASB: "The dimensions, shapes, spacing and relative positions of the footwear outsole design components and tire tread blocks (not the same as the manufacturer's footwear or tire size). Physical size is a class characteristic." ULTR: "The size, shape, spacing and relative position of the outsole design components on a footwear item."
Probable Conclusion	A response of *HighAssn* or *NonAssn.*
Randomly Acquired Characteristics (RACs)	A feature on a footwear outsole resulting from random events (e.g., cuts, tears, holes, and or attached debris). The position, orientation, shoe size, and shape of these characteristics contribute to the uniqueness of the shoe outsole. Also described as "identifying characteristics." ASB: "A feature on a footwear outsole or tire tread resulting from interaction with an object(s) including, but not limited to: cuts, scratches, tears, holes, stone holds, abrasions and the acquisition of debris. The position, orientation, size and shape of these characteristics differentiate a footwear outsole or tire tread from other footwear outsoles or tire tread with similar class characteristics. Randomly acquired characteristics are essential for an identification of a particular item of footwear or tire as the source of an impression." ULTR: "A feature (e.g., a cut, a scratch, a tear, a hole, or a stone hold) on the outsole of a footwear item acquired through random events. The position, orientation, size, and shape of these characteristics can be used to differentiate one footwear outsole from another when those outsoles share the same class characteristics. One or more 'randomly acquired characteristics' are required for the 'source identification' of a known footwear item to a questioned impression."
Reliability	The precision or consistency of conclusions (without regard to accuracy) that included repeatability (intra-examiner agreement) and reproducibility (inter examiner agreement). We generally refer explicitly to "repeatability" and "reproducibility" and minimize our use of "reliability" because while many authors use the term as we do, some authors use it to mean "consistently accurate" (e.g. [27]) or others use it as synonymous with "accurate" (see [28] for discussion).
Substrate	The surface upon which a shoe makes contact, such as tile, wood, or carpet. Also includes doors, counters, paper products or any surface on which a shoe may leave a mark.
Superimposition	A comparison method performed by placing one object over the other. ASB: "A visual comparison performed by placing one object over the other."
SWGTREAD	Scientific Working Group for Shoeprint and Tire Tread Evidence.
True Negative (TN)	A conclusion of *Excl* on a nonmated QKset.
True Negative Rate (TNR)	The proportion of nonmated responses that resulted in true negatives.
True Positive (TP)	A conclusion of *ID* on a mated QKset.
True Positive Rate (TPR)	The proportion of mated responses that resulted in true positives.
Wear	Erosion of the surfaces of a footwear outsole during use. ASB: "Erosion of the surfaces of a footwear outsole or tire tread during use." ULTR: "The position and degree of erosion on the outsole of a footwear item."

Appendix B Background: Forensic Footwear Examination

This appendix provides supporting material for Section 2, Background.

Comparisons between questioned impressions and known footwear—for the purposes of determining whether the known footwear can be included or excluded as a possible source of the evidence—are the cornerstone of the forensic footwear discipline. Questioned impressions (Qs) are left by footwear (specifically the bottom of the shoe known as the outsole) unintentionally on substrates (surfaces) found at crime scenes [29]. They are referred to as "questioned" because their source is unknown in casework. (In this study, the Qs were prepared from known sources to simulate Qs in casework). These impressions may be two dimensional (wherein a matrix is transferred from the shoe to a substrate or removed from the substrate by the shoe) or three dimensional (wherein the shoe steps into a deformable substrate, such as soil). Their features can be compared by FFEs to one or more known shoes collected from persons of interest in a particular case or persons with legitimate access to a crime scene. These comparisons are conducted in a sequential manner, and consider a variety of features that become increasingly more discriminating as the process continues. These comparisons can be complex given the variability of the questioned impressions encountered: for example, some questioned impressions are partial, some are distorted from movement by the wearer, and some overlap others.

FFEs are guided by a standard for the examination of footwear impression evidence that was published by the Scientific Working Group for Shoeprint and Tire Tread Evidence (SWGTREAD) in 2006 [4], which remains the prevailing standard for footwear examination today. The procedure therein begins with an evaluation of the questioned impression, focused on determining its suitability for comparison with known shoes. If suitable, the procedure guides the examiner through a stepwise comparison process for assessing both class characteristics and RACs. When conducting these comparisons, FFEs employ both side-by-side comparison and superimposition. Examination begins with an evaluation of class characteristics, which arise as a function of the manufacturing process—outsole design, physical size, and mold variations. (Shoes of the same make, model and manufacturer's size are generally indistinguishable when boxed for shipment to retailers.) If the design corresponds, the FFE next compares the physical size. Test impressions* are collected under controlled conditions to provide a reproduction of the outsole features on the known shoe. Test impressions are often collected on transparent material to enable practitioners to superimpose the outsole features over the questioned impression and compare features that are not easily assessed side-by-side. If both design and size correspond, the examination continues with a comparison of the position and degree of wear. It is only after shoes are worn that the appearance of the outsole changes and damage is acquired. The general degrees and positions of wear are considered class characteristics as the structure of the foot of most individuals wears down the outsole in the same general positions—the ball area in the forefoot and the posterior heel area. Wear can be used as a basis for *Excls*: for example, if a suspect's shoe is less worn (overall or in specific areas) than the source of a questioned impression (collected earlier) evidenced by the appearance of the features in the impression, the suspect's shoe may be eliminated based upon wear alone. With continued use, shoes may acquire RACs, which can be used to differentiate outsoles that share the same class characteristics [29]. As the final step in the examination, the examiner compares any RACs observed, which constitute the basis for source attribution. RACs are evaluated according to their position, size, shape, orientation, clarity, and reproducibility. If an FFE determines that a sufficient number of corresponding RACs are observed between the questioned impression and the known shoe to conclude that the impression was left by the known shoe, an *ID* may be reported. During any step in this comparison sequence, FFEs must evaluate any apparent dissimilarities (potential differences) and determine whether the dissimilarities are sufficient to demonstrate that the known footwear is not the source of the impression ("meaningful differences"); if any meaningful differences are observed, the known shoe can be excluded as the source. Often a definitive conclusion (*ID* or *Excl*) cannot be reached due to limitations associated with the questioned impression, the absence of sufficient corresponding RACs, or the absence of any meaningful differences.

The discipline developed a range of conclusions to enable FFEs to provide varying degrees of support for (*ID*, *HighAssn*, *Assn, LimitedAssn*) or against (*NonAssn, Excl*) the likelihood that the known shoes compared are the source. The range of conclusions was standardized by SWGTREAD in 2006 [19] and revised in 2013 [18]. According to the US DOJ Uniform Language for Testimony and Reporting [13], "a conclusion provided during testimony or in a report is ultimately an examiner's decision [a judgment, an opinion] and is not based on a statistically-derived or verified measurement or comparison to all other footwear item." The results of these comparisons are documented in a formal laboratory report detailing the examiners' observations, findings, and conclusions (and oftentimes accompanied by visual aids); these results can provide valuable information to investigator and prosecutors as they have the potential to identify, include, and exclude a known shoe as the source.

Appendix C Methods and Materials

This appendix provides supporting material for Section 3, Study Description.

Appendix C1 Study Description

This study was designed to evaluate the accuracy and reliability of FFE decisions. This research was modeled after previous black box research studies examining the accuracy and reliability of latent print comparison [20,30].

Participants were each asked to perform 100 footwear comparisons. Test materials were provided as both printed photographs and digital images. No physical items (i.e., items of footwear, lifts, or items bearing Qs) were distributed.

Custom web browser-based software was used to present low-resolution images, allow for download of high-resolution images, present comparison questions, allow markup of randomly-acquired characteristics (RACs), and indicate the orientation of the questioned impressions.

Registration remained open for 29 weeks (February 2019-August 2019), and test access was available for 48 weeks (June 2019-May 2020). Participants received regular study relevant notifications throughout the process. Study administrators were available for assistance for the duration of the test.

The test comprised 100 comparison sets distributed via five physical test packets, each containing 20 comparison sets. Participants were unaware of the overall and per packet mated to nonmated proportions which, although static overall (40% mated and 60% nonmated), varied across individual test packets. To assess intra-examiner repeatability, 10 comparison sets were included twice in each participant's test materials, with at least one packet between repeated comparison sets. Participants

* *See Test Impressions under Appendix C4.1 for details regarding test impression preparation.*

were instructed to conduct the evaluations with the same diligence employed in their operational casework, and not to collaborate. Participants were further instructed not to retain, duplicate or mark study materials. If marks or damage were observed on returned materials, the damaged materials were replaced with duplicates. No time limitation was imposed for the completion of individual packets. The test materials are further characterized in *Appendix C4*.

Participants were permitted to complete the comparisons within a packet in any order and save and revisit worked sets. Participants were also afforded the opportunity to review and edit responses before final submission. Submitted comparison sets could not be revisited.

Participation consent and anonymity were approved by the Federal Bureau of Investigation's Institutional Review Board. Anonymity was maintained through multiple levels of de-identification, data segregation and information flow control. Participant anonymity was provided through the use of randomly assigned ParticipantID numbers. ParticipantIDs were anonymized prior to data analysis, precluding the analysis team's ability to cross associate participants, personally identifying information, questionnaire responses or test results. Destruction of existing cross-reference indices occurred prior to public presentation of results (e.g., indices correlating ParticipantIDs with packet delivery postal addresses). Therefore, participant identities could not be associated with the results at any point during analysis, or subsequently, such as for discovery. Participants were assumed to be volunteers. However, pressure to participate from employers or other entities cannot be precluded, nor the performance effect of such a factor calculated.

Prior to commencement of the study, a Beta test was conducted to assess software functionality, examiner experience and test packet distribution logistics. The Beta test materials were not reused in the formal study, nor were the results included in analysis.

Appendix C2 Experimental Design Considerations

In 2005, the National Research Council (NRC) of the National Academy of Sciences (NAS), was tasked with conducting a study to examine the state of the forensic sciences in the United States. In 2009, the NRC published the results of its multi-year study. The publication, *Strengthening Forensic Science in the United States: A Path Forward*, raised "serious" concerns regarding the lack of demonstrated scientific validity for a number of forensic science disciplines, including forensic footwear examination [1]. The report pointedly indicated an imperative need for rigorous systematic research specifically designed to develop and establish quantifiable measures of the: (1) accuracy and reliability of forensic analyses and (2) uncertainty in the conclusions of those forensic analyses. Research examining issues of human observer bias and error was also strongly encouraged.

Subsequently, in 2016 the President's Council of Advisors on Science and Technology (PCAST) released a report indicating a need within the forensic science community to empirically study the accuracy and reliability of certain feature-based comparison techniques [2]. Footwear impression examination was explicitly discussed (p. 12-13). The PCAST report indicated a need for credible empirical studies that assess the validity and reliability of feature-based comparison methods. This decision analysis study has been specifically designed to address the concerns and recommendations articulated in the NRC and PCAST reports (NRC Recommendation 3a-c and PCAST Recommendation 5a and 5b, p. 17-18 and Finding 7, p. 117).

Much of the design of this study is based on the design of the FBI Laboratory/Noblis latent print black box study [20,30]. Lessons learned from that study that were incorporated here include improved quality assurance procedures in study sample creation, more consideration in selecting substrate and matrix combinations to be representative of distributions found in casework, and verifying that distributions of quality and other attributes are comparable for mated and nonmated sets.

Appendix C3 Participants

Participation was open to all practicing examiners who had conducted footwear evidence casework examinations within the 5 years preceding the study announcement. For this study, a practicing forensic footwear examiner (FFE) was defined as "an individual who conducts forensic comparisons of questioned footwear impressions and known footwear items and communicates their findings in written reports and during testimony in courts of law."

A total of 84 FFEs were considered participants in this study (Table S1). This does not include two FFEs who are omitted from analyses because they submitted fewer than 20 QKsets (a requirement communicated to participants in the FAQs; see *Appendix C5.5*). Fifty-five of the participants completed all 100 assigned comparisons, 16 completed at least 40 comparisons, and 13 completed at least 20 comparisons. For the subset of analyses in which we compute and compare individual rates for each participant, we limit analyses to the 71 participants who completed at least 40% of the assigned comparisons (2 test packets).

# QKsets	*# Participants*	*Use*
1-2	2	Omit from study
20-34	13	Use in analyses; do not use in measuring examiner-specific rates or comparing examiners
40-80	16	Use in all analyses
Complete (100)	55	

Table S1. Participants by number of comparisons completed.[*]

Participants were solicited at relevant conferences and via professional organization announcements. No individual was directly solicited and no qualified participants (based on aforementioned requirements) were barred from participation. Participants were required to complete an IRB approved consent form and background questionnaire prior to study access. The questionnaire responses were used to assess performance relative to examiner variables such as training, experience and certification. The responses were also used to inform an understanding of the participating examiners' operational procedures. A short post-test survey was conducted after the study was concluded to collect participants' overall assessments of the study; 67 of the 84 participants completed the post-test survey.

See *Appendix P* for the responses from the background questionnaire and short post-test survey.

Appendix C4 Footwear Items and Impressions

This appendix details the data collection process developed and used in this study. Data collection for this study was conducted in 2018, and involved a detailed process designed to ensure image quality, ground truth source attribution, and samples collected to be broadly representative of attributes found in casework. This process was time consuming. On average (after the procedure was clearly defined), data collection for each shoe took about 90 minutes: 20 minutes to image an outsole; 35 minutes to prepare, collect, and image a Q; 15 minutes to prepare and image a hand-rolled test impression; and 20 minutes to prepare and image a walking test impression; these do not include quality assurance or image preparation (see *Appendix C4.4*). Fig S1 and Fig S2 show an example of source images for one shoe, and the resulting QKset.

All images were captured either photographically or digitally scanned. The photographic images were captured following the SWGTREAD recommended best practices [31].

Images were provided to participants as both digital images, and as printed photographs and transparencies. The digital images were calibrated to 600 pixels per inch (ppi) and provided (via the project online interface for download) to participants in both JPEG and TIFF formats. The images that were printed as photographs and transparencies were resampled to 300 ppi (for printing) and provided as physical packets via mail to the participants.

[*] *Note that after omitting one problematic QKset, the dataset used for analyses included 41 participants with 100 trials each and 14 participants with 99 trials each: see Appendix D4 for details.*

Fig S1. Source images for footwear item #1364, which was released as QK134 (shown in Fig S2). Note the use of preprinted labels in images. The image of the upper includes the specific substrate assigned to this shoe (H13S refers to this specific hardwood board). Only one of the five outsole images is shown here.

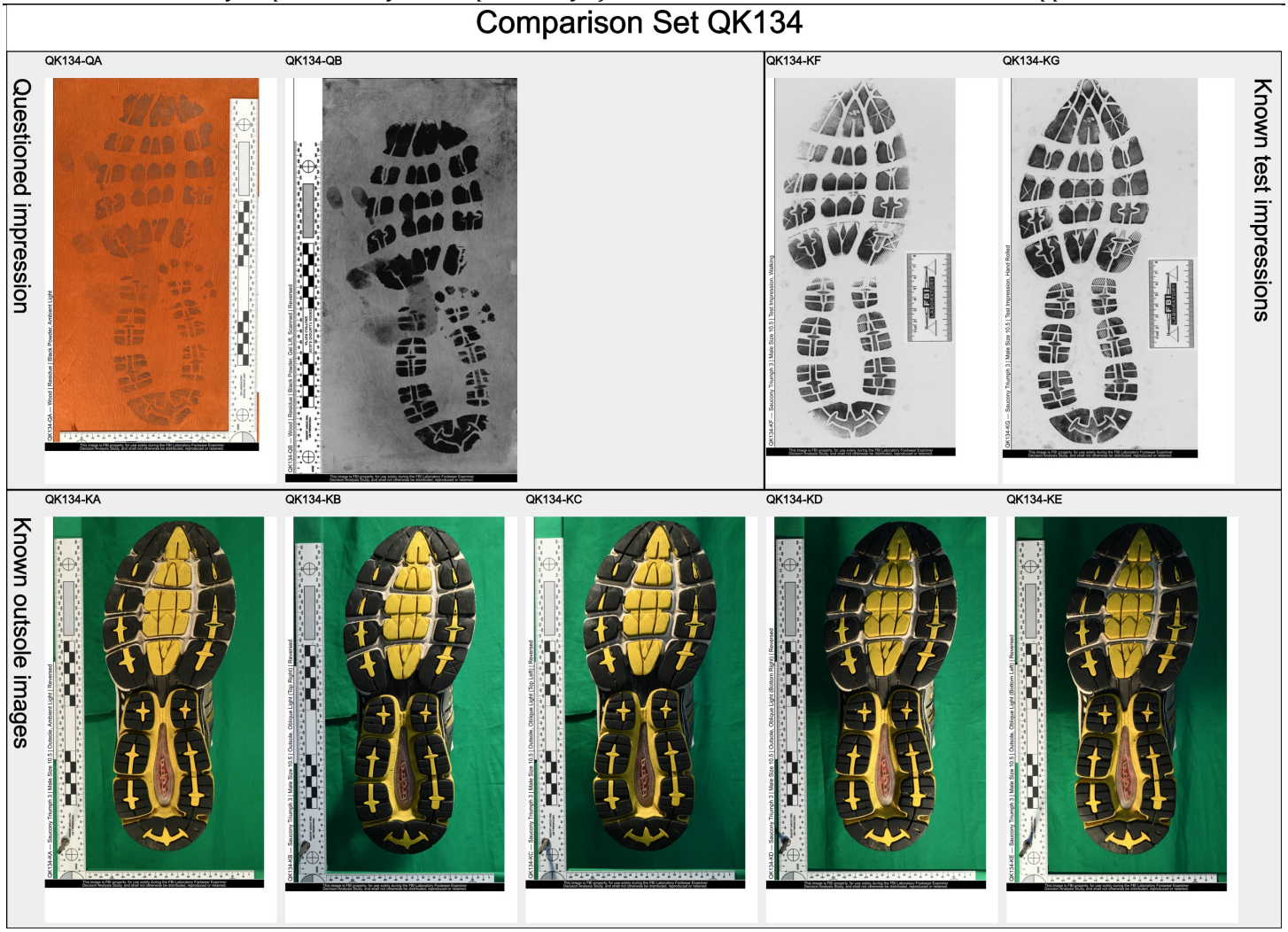

Fig S2. QK134, a mated QKset created from the source images shown in Fig S1.

Appendix C4.1 Footwear

The footwear used in the study (Table S2) came from four sources:

- Matt Marvin of Ron Smith & Associates, Inc. provided 64 New Balance 878 shoes (all male size 11), of which 54 were used in this study. Outsole images and walking test impressions for these shoes had already been collected prior to this study [32]. Our study team collected K hand-rolled test impressions and Q impressions.*
- The FBI provided 1171 Eastern Mountain Sports Day Hiker boots (variety of men's and women's sizes), of which 148 were used in this study. To ensure a statistically adequate sample, a random selection was separately acquired from each size within the men's and women's sub-samples. Hand-rolled test impressions for these boots had already been collected. Our study team collected K outsole images, walking test impressions and Q impressions.
- Seven pairs of new shoes were purchased, of which six individual shoes (three Converse Chuck Taylor All Star and three Vans Classic) were used in this study. Volunteers were recruited to wear the newly purchased footwear for up to two weeks to accumulate wear and randomly acquired characteristics (RACs) (see "Use of New Footwear" under *Appendix C4.3*).
- Volunteers and secondhand shops were used to collect the remainder of the footwear used in this study, which included 85 shoes or boots of 25 makes and models.

* *Any footwear examiners who had access to the New Balance 878 shoes prior to this study were not assigned any of those shoes in this study.*

Make and Model	*N*	*Use*			*DiffMakeModel group*	*Source*
		Mated	*Nonmated Q*	*Nonmated K*		
Adidas Cloudfoam Advantage Clean	4	1	2	2		New purchases, volunteers, secondhand shops
Adidas NEO Cloudfoam	2	1	1	1		
Asics Gel Nimbus	6	4	6	2	Brooks	
Brooks Adrenaline GTS 7	2	0	2	2		
Brooks Adrenaline GTS 12	2	0	0	2		
Brooks Adrenaline GTS X	2	0	0	2		
Converse Chuck Taylor All Star	12	4	8	8		
Dr. Martins 1914	4	2	2	2		
Eastern Mountain Sports Day Hiker	148	58	76	76		Federal Bureau of Investigation
ECCO Saunter GTX	4	2	2	2		New purchases, volunteers, secondhand shops
GORE-TEX Matterhorn	2	0	2	2	Timberland	
Keen Timmons Low Lace	4	2	2	2		
Koflach Mountaineering Boot (Vibram Sole)	2	0	1	2	Timberland	
Mizuno Wave Inspire 13	4	0	4	4		
New Balance 878	54	21	28	28		Ron Smith and Associates
Nike Air Force 1	2	0	2	2	Nike Air Force 1	New purchases, volunteers, secondhand shops
Nike Air Force 1 UltrafForce Mid	2	0	2	2		
Nike Air Max 95	2	0	2	2	Nike Air Max	
Nike Air Max Coliseum Racer	2	0	2	2		
Nike Metcon 2	4	2	2	2		
Saucony Triumph 3	4	2	2	2		
Saucony Triumph 11	4	2	2	2		
Sperry Top-Sider	4	2	2	2		
State Street Waterproof	2	0	2	2	Timberland	
Timberland Premium Waterproof Boots	4	1	4	3		
Vans Classic	4	2	2	2	Vans	
Vans Van Doren	4	2	2	2		
Total	290	108	162	162		

Table S2. Footwear by make and model. Counts are of distinct shoes (not pairs).

Outsoles

For the 54 New Balance 878 shoes, pre-existing outsole images were available. The outsole images of these shoes were collected prior to the initiation of this study by Matt Marvin of Ron Smith & Associates, Inc., using a Nikon D800 with a Micro Nikkor 60mm f2.8 lens. For each of these shoes, one composite HDR (high dynamic range) outsole image was created by merging nine images with varied directions of oblique light (using Adobe Photoshop's *Merge to HRD Pro* function).

For the remainder of the footwear (239 footwear items, of which 220 were used in this study), outsole images were captured using a Nikon D850 configured with a Micro Nikkor 60mm f2.8 lens, using autofocus, remote shutter release, and aperture priority exposure (using a range of f-stops, generally f16) with dual sided 7-image bracketing (0.3 exposure increments) while illuminating the outsole from five directions, as detailed in Table S3 and shown in Fig S2; all these images contained a visible metric scale. Footwear was mounted on a receiving stand adjusted for each differently sized item of footwear (Evident Forensic Supply #9859). Camera and lighting components were statically affixed to a motorized REPRO Industria Fototecnica Firenze copy stand outfitted with adjustable light fixture arms. The conical reflective light fixtures were outfitted with flood bulbs. For downstream image calibration, a NIST certified metric scale was placed in a standardized location in all outsole images. A spirit level was employed to ensure that the camera, outsole and scale were properly aligned prior to image capture. Multiple alignment checks were systematically implemented as follows: (i) the camera was leveled horizontally and vertically, (ii) the outsole was leveled medially-laterally and heel to toe; as much as possible given the curvature of some outsoles, (iii) the NIST certified metric scale was aligned at, and parallel to, the highest outsole point (typically at or near the footwear arch) and (iv) a final parallel alignment check was performed across the camera, outsole and scale. Randomly assigned footwear identifiers were embedded within all images via the inclusion of preprinted 2" x 3.5" white, semi-rigid cards placed at outsole level (identical to those used for test impressions). All identifying information was cropped from the images prior to comparison set compilation and dissemination. In total, 35 outsole images were captured for each of the 239 footwear items, for a total of 8,365 original images.

To ensure a consistent, stable and vibration-free imaging environment, the following quality control measures were implemented to prevent the need to physically manipulate the imaging environment or footwear once in place and properly prepared. A dedicated workstation running a camera settings software and remote shutter control was employed during outsole capture. This quality control measure further enhanced standardization across all photos for a single footwear by precluding the need to physically handle the camera once setup was finalized. A power strip outfitted with individual on/off switches was used to power on and off the four light fixtures attached to the copy stand. As a control check to ensure proper lighting, a nail was placed head down on the scale prior to imaging; the shadow cast by the nail effectively provided a sundial-like lighting control check for subsequent use during quality control and comparison set compilation. For added consistency and potential

reflectance mitigation, a solid colored cloth was placed across unnecessary background components (e.g., copy stand surface). To mitigate potential loss of data, images were simultaneously recorded to the camera memory card and an external hard drive.

Number of Light Sources and Position	*Bracketed Shots*
1. Four light sources (i.e., all 4 lights on). One each positioned as below.	7 *each* Dual sided 0.3 exposure increments
2. A single light source placed at the top left of the toe (~10:00 o'clock). All other lights off.	
3. A single light source placed at the top right of the toe (~2:00 o'clock). All other lights off.	
4. A single light source placed at the bottom left of the heel (~8:00 o'clock). All other lights off.	
5. A single light source placed at the bottom right of the heel (~4:00 o'clock). All other lights off.	

Table S3. Lighting configurations and capture sequence.

Test Impressions

Two types of known test impressions were used in this study: *walking test impressions* created in a walking-stepping fashion while wearing the shoe, and *hand-rolled test impressions* created manually. Volunteers with foot sizes within ± ½ the shoe size were recruited for the walking test impressions.

Walking Test Impressions

For the 54 New Balance 878 shoes, walking test impressions had already been collected prior to this study. For the remainder of the footwear in the study, walking test impressions were produced using the following procedure:

1. As needed, excess outsole detritus was removed using a soft bristled brush.
2. A 9.5" x 13.5" clear adhesive lift sheet (CSI Forensic Supply #2-6003) was placed tacky side up on a polished 12" x 24" marble tile.
3. Black fingerprint powder (Sirchie Silk Hi-Fi Volcano #BPP0964) was liberally applied to the outsole surface using a fiberglass fingerprint brush trimmed to approximately 1.5 inches (Zephyr #1-0200). Powder was applied using a continuous twisting-sweeping motion in multiple directions to ensure adequate coverage.
4. Post powder application, the side of the shoe was tapped several times to remove excess powder.
5. Without disturbing the powder, and while seated, a size appropriate volunteer secured the shoe on his/her foot. The volunteer then rose and, as naturally as possible, walked over the adhesive sheet. Care was taken to completely plant the entire outsole surface in a heel to toe fashion while walking across the adhesive sheet (i.e., not a flat step). The volunteer was instructed to lift and hold the shoe and affixed adhesive sheet up from the working surface after stepping through the heel to toe motion.
6. A team member then removed the adhesive lift sheet in one steady motion from heel to toe.
7. A 2" x 3.5" white, semi-rigid card (Avery 5371) containing the test impression identifier was immediately applied directly to the tacky surface of the adhesive lift sheet. A second 2" x 3.5" white card containing an eight cm scale was also place directly on the adhesive lift sheet (thereby permanently embedding both the test impression identifier and scale for use in downstream calibration). A marked ruler was utilized to ensure consistent card placement across test impressions.
8. A 10" x 14" polyester clear overlay sheet (CSI Forensic Supply #2-8003) was slowly applied to the adhesive side of the lift sheet in a heel-toe direction while applying pressure in a medial-lateral sweeping motion (to minimize air bubbles).
9. The walking test impression was subsequently digitized in grayscale at 600 ppi using an Epson Expression 11000XL Graphic Arts Scanner.

The 54 pre-existing walking test impressions from the New Balance 878 shoes were produced using a procedure similar to that detailed above with two primary exceptions: (i) a depletion series of test impressions was created and (ii) a white clean room tacky mat was employed as the receiving substrate. No powder reapplication occurred during the depletion series. Subject matter experts employed by the curator of the footwear collection selected the impression assessed as containing the best outsole representation. The impression was subsequently photographed using a Nikon D800 with a Micro Nikkor 60mm f2.8 lens.

Hand-rolled Test Impressions

For the 148 Eastern Mountain Sports Day Hiker boots, hand-rolled test impressions had already been collected prior to this study. The procedure developed for those hand-rolled test impressions was adapted for the remaining collection in this study as follows:

1. As needed, excess outsole detritus was removed using a soft bristled brush.
2. A 9.5" x 13.5" clear adhesive lift sheet (CSI Forensic Supply #2-6003) was placed tacky side up on a polished glass-topped dry erase worktable, with approximately one-half inch extending beyond the work surface edge.
3. Black fingerprint powder (Sirchie Silk Hi-Fi Volcano #BPP0964) was liberally applied to the outsole surface using a fiberglass fingerprint brush trimmed to approximately 1.5 inches (Zephyr #1-0200). Powder was applied using a continuous twisting-sweeping motion in multiple directions to ensure adequate coverage.

4. Post powder application, the side of the shoe was tapped several times to remove excess powder.
5. The shoe was held over the adhesive lift sheet in an outsole-work surface parallel orientation.
6. Maintaining a parallel orientation, the shoe was then firmly pressed against the adhesive lift sheet.
7. While applying *continual* pressure against the work surface and using the excess one-half inch portion of the adhesive lift sheet extending beyond the work surface, the shoe and adhesive lift sheet were slowly pulled off the edge of the workbench toward the operator.
8. As the outsole began to clear the work surface, firm medial-lateral pressure was applied manually to the outsole through the non-tacky surface of the clear adhesive lift sheet.
9. While continuing to advance the outsole off the work surface edge, pressure was applied to the remaining outsole in this fashion until the full outsole was in contact with the adhesive lift sheet.
10. The shoe was then turned over and additional pressure was manually applied where needed to ensure full contact over the entire outsole surface.
11. When complete, the adhesive lift sheet was removed in one steady motion from heel to toe.
12. A 2" x 3.5" white, semi-rigid card (Avery 5371) containing the test impression identifier was immediately applied directly to the tacky surface of the adhesive lift sheet. A second 2" x 3.5" white card containing an eight cm scale was also place directly on the adhesive lift sheet (thereby permanently imbedding both the test impression identifier and scale and for downstream calibration). A marked ruler was utilized to ensure consistent card placement across test impressions.
13. A 10" x 14" polyester clear overlay sheet (CSI Forensic Supply #2-8003) was slowly applied to the adhesive side of the lift sheet in a heel-toe direction while applying hand pressure in a medial-lateral sweeping motion (to minimize air bubbles).
14. The hand-rolled test impression was subsequently digitized in grayscale at 600 ppi using an Epson Expression 11000XL Graphic Arts Scanner.

Appendix C4.2 Questioned Impressions

All questioned impressions were produced under controlled laboratory conditions by either subject matter experts and or team members trained by subject matter experts. No impression evidence from operational casework was used. Questioned impressions were produced using varying combinations of substrates, matrices, processing techniques, and lift methods. Both new and used substrates were included. Full and partial impressions with varying degrees of distortion, obfuscation, overlapping impressions and background (substrate) noise were generated. No 3D materials were presented to participants during this study. See Table S4 for details of the questioned impressions (Qs).

Participants were provided the type of limited information about the questioned impression that would be available in casework, including substrate, matrix, processing method, lift technique, and lighting.

Multiple data collection control measures were implemented to regulate the quality and characteristics of each questioned impression. A total of 253 Q impressions from 239 items of footwear were collected (generally with 2-3 distinct images of each impression, and 5-7 exposure-bracketed captures of each image). To evaluate and select among the Q impressions using a quantifiable and reproducible method, we developed the quality rubric summarized in *Appendix F1*, and reported in [17]. Out of that pool, 162 Q impressions were selected and used in the study. Each Q impression was included in one nonmated QKset, and about 2/3 of the Qs were also included in one mated QKset.

Q impressions were collected on 59 distinct substrates, in the following categories:

- Clothing/Material: t-shirts, jeans, terrycloth
- Countertop/Tabletop: granite, marble, formica, melamine, porcelain
- Door: painted, paneled
- Glass: clear, mirrored
- Metal: vehicle exterior, galvanized, stainless
- Paper/Cardboard: copier, newspaper, cardboard, paper bags
- Plastic: construction sheeting
- Rug/Carpet: low pile
- Tile: ceramic, vinyl, marble
- Wood: laminate flooring, finished and unfinished lumber, OSB

Q impressions were prepared using four matrices: blood, mud, residue, and soil/dust.

Based on the substrate-matrix combination, in some cases, additional steps were taken to process, lift, and collect additional reproductions of the Q impression.

Note that 3 mated and 3 nonmated QKsets used a walking test impression collected with black powder as a Q: these were collected from unused footwear items in order to assess how FFEs performed on impressions that contained no RACs; test impressions were used in order to provide near-ideal Q impressions. See the next section for discussion.

Substrate	*Matrix*	*QKsets*		*Nonmate class category*					*Deposition*				*Extent*		*Superimposition*	
		M	*NM*	*Same make /model/size*	*Size Diff 0.5*	*Size Diff 1*	*Diff make /model*	*Diff Foot*	*Jump*	*Kick*	*Run*	*Walk*	*Full*	*Partial*	*No*	*Yes*
Test impression	Black powder	3	3	3								6	6		6	
Cloth	Blood	2	3	2		1					1	4	2	3	5	
Countertop	Blood	1	3	2	1				3			1	1	3	4	
	Mud	2	2	2					2			2	4		4	
	Residue	3	5	3			2		5			3	2	6	4	4
	Soil/Dust	5	5	4	1				6			4	2	8	10	
Door	Blood		1		1					1				1	1	
	Mud		3				3			3			3		3	
	Soil/Dust	1	13	5		2	6			14			3	11	14	
Glass	Blood	1	2			1		1				3	2	1	3	
	Residue	9	10	9	1							19	7	12	14	5
	Soil/Dust	2	3		2		1					5	2	3	5	
Metal	Blood	2	2	1				1				4		4	2	2
	Mud	2	2	2								4	4		2	2
	Residue	3	5	2	1		2		2			6	2	6	6	2
	Soil/Dust	2	4	2	1		1		3	2		1	3	3	3	3
Paper	Blood	10	13	6	1	4	2		4			19	16	7	21	2
	Soil/Dust	3	5	5								8	3	5	5	3
Plastic	Residue	2	2	2					2			2		4	4	
Rug	Soil/Dust	1	1		1							2	2		2	
Tile	Blood	2	2	2							2	2		4	4	
	Mud	1	2			2						3	2	1	3	
	Residue	16	28	14	6	4	4		5		2	37	17	27	26	18
	Soil/Dust	12	15	10	3	1	1				4	23	12	15	17	10
Wood	Blood	2	3	3								5	5		3	2
	Mud	1	2	2								3	3		3	
	Residue	12	14	13			1				2	24	16	10	17	9
	Soil/Dust	7	9	6	2		1		2			14	9	7	9	7
Total		107	162	100	21	15	24	2	34	20	11	204	128	141	200	69
% of QKsets (n=269)		40%	60%	37%	8%	6%	9%	1%	13%	7%	4%	76%	48%	52%	74%	26%

Table S4. Questioned impression distribution by substrate and matrix, with respect to class characteristic categories, manner of deposition, extent, and superimposition. (Does not include the mated QKset omitted from analyses (which was glass, blood, walking, partial contact, no superimposition).)

All other questioned impressions were processed, collected, photographically imaged or digitized following SWGTREAD best practices [31,33,34]. With the exeption of the gel lifters, the questioned impressions were captured using a Nikon D850 configured with a Micro Nikkor 60mm f2.8 lens, using autofocus, remote shutter release, and aperture priority exposure (using a range of f-stops, generally f16) with dual sided 7-image bracketing (0.3 exposure increments); all these images contained a visible metric scale. Specific imaging techniques were employed in some cases, which are detailed below.

Questioned Impression Preparation

When preparing the Qs, efforts were made to vary the quantities of substrate and outsole surface area reproduced to create impressions of varying quality; a depletion series of several steps, in some cases, was used to acheive the desired variation. In all cases, the item of footwear was worn by a team member with foot sizes within ± ½ the shoe size who prepared the questioned impression using one of the four deposition methods (jumping, kicking, running, or walking).

- Blood impressions were made using synthetic blood. Blood was manually applied to the outsole or the shoe was worn by a team member who stepped in a pool of blood. After preparing blood impressions, they were left to dry for up to 24 hours. If patent, they were photographed in situ under ambient light; if latent, they were processed using a method described below.
- Mud impressions were made by a team member who stepped in a tray of mud. After preparing mud impressions, they were photographed in situ under ambient light.
- Residue impressions were made using various oily matrices (e.g., Pam cooking spray and shoe shine). The matrix was manually applied either to the outsole or the substrate. After preparing residue impressions, they were photographed in situ under ambient light (if patent); if latent, they were processed using a method described below.
- Soil/dust impressions were made using various dry dusty matrices (e.g., available dust accumulated by walking inside an office building and dust collected from vacuum sweepers). The matrix was applied either to the outsole (while a team member wore the shoe and stepped into the matrix) or the substrate. When applied to the substrate, the shoe was worn by

a team member who stepped into the matrix. After preparing soil/dust impressions, they were photographed in situ using both ambient and oblique light.

Questioned Impression Processing

Some Qs were processed physically using fingerprint powder or chemically using Leucrocrystal Violet (LCV). In addition to these processing methods, some were lifted. These additional steps generated additional Q reproductions. Up to three reproductions of each Q impression were possible (in situ, after physical or chemical processing, and lifted) and provided in QKsets. The number of reproductions included in the QKsets varied from one to three.

Physical and Chemical Methods

- Fingerprint powder was applied to all residue impressions with a fiberglass fingerprint brush (trimmed to approximately 1.5 inches, Zephyr #1-0200) using a continuous twisting-sweeping motion in multiple directions to develop the impression detail. For darker substrates, gray fingerprint powder (Arrowhead Forensics #A-2332W) was applied; for lighter substrates, black fingerprint powder (Sirchie Silk Hi-Fi Volcano #BPP0964) was applied. After powdering the impressions, they were photographed.
- LCV working solution (Arrowhead Forensics #A-PF024) was applied to some of the blood impressions using spray bottles (including a Prevail disposal sprayer kit (Forenic Source # 1004109)). After processing the impressions, they were photographed.

Lifting Methods

- Gel lifters, in both black (BVDA #B-12500) and white (BVDA #B-15500) were used to lift powdered impressions. Black were used to lift gray-powdered impressions and white for black-powdered impressions. Prior to lifting, the clear cover sheets were removed and the lifters were allowed to relax for a minimum of five minutes. Then, the lifters were applied manually along the length of the impression using a sweeping motion from one end to the other. A team mate ensured adhesion by applying pressure across the lift manually and or using a fingerprint roller. Lifts were left on the impression for a minimum of 10 minutes and then removed in same manner they were laid down. The lifted impressions were placed adhesive side up (without cover sheets) in high quality photo paper boxes for transport from the location where they were prepared to the FBI Laboratory for imaging. They were imaged using the BVDA GLScan within 24 hours of lifting.
- Stat-Lifts (Arrowhead Forensics #A-5031) were used to lift some of the soil/dust impressions. The backing sheets were removed from the lifters. While applying pressure to one end of the lifter (to prevent movement), the team mate manually applied pressure along the length of the impression using a sweeping motion from one end to the other. A team mate ensured adhesion by manually applying pressure across the lift. Lifts were removed in same manner they were laid down. The lifted impressions were photographed using oblique light.

Appendix C4.3 Use of New Footwear Items

Six QKsets were created using unused footwear items:

- Three nonmated QKsets paired a Q from an unworn footwear item with a K from another unworn footwear item;
- Three mated QKsets paired a Q from an unworn footwear item with a K from the same footwear item after it was worn for up to two weeks.

Each participant was assigned one of each of these. The purpose of this was to assess how results were affected when given footwear items that had no RACs. In order for participants to see the presence or absence of RACs, the Qs were collected to be as close to ideal as possible: walking test impressions were collected using black powder.

Appendix C4.4 Image Preparation

The preparation of images for use in the study involved the following steps:

Quality Assurance

- Verify all footwear characteristics are as specified, checking footwear identifiers shown in each image against filenames. The filename for each image file included key information for use in quality assurance. For example, the top center image in Fig S1 was named "FBB_1364_STRM105L_KB04.TIF", which includes the sensitive item number (1364), the model (STR= Saucony Triumph 3), size and foot (Male 10.5 Left), and image type (KB=outsole image with directional light from top left). When the image was assigned to a QKset, the public name (QK134-KB) and sensitive name (1364-M) for the QKset were added as prefixes, resulting in "QK134-KB__1364-M--FBB_1364_STRM105L_KB04.TIF", which could be manually reviewed or automatically parsed.
- Verify focus.
- Verify directionality of light sources for outsole images.

Calibration and Resampling

- The resolution of photographic images was calibrated in Adobe Photoshop based on pixel counts between defined points on the rulers in the images, and checked by a second team member.
- The resolution of scanned images is fixed: gel lifts were scanned at 1044ppi, and test impressions were scanned at 600ppi.
- Images were resampled to 600ppi in ImageMagick using the Lanczos algorithm.

Manual Processing

- Images were manually cropped in Adobe Photoshop to remove all identifying handwritten text or printed labels (generally to 7x14", 178x356mm)
- Images with significant color cast were corrected in Adobe Photoshop by setting the white point using the white area of the ruler.
- Dark images were adjusted by histogram correction (setting the brightest point in the histogram to white)

Automated Processing

- All of the automated image processing was performed programmatically using ImageMagick.
- If rulers were cropped out due to identifying labels, rulers were programmatically added. Rulers were programmatically added to all New Balance 878 outsole images, and gel lifts.
- Outsole images and gel lifts were flipped (so that they orient with the impression on the ground).
- Labels were programmatically added; all labels shown in the images were extracted from information in the filenames.
- Images were converted to TIFF (losslessly compressed using LZW) and JPG (maximum quality setting).
- Proof sheets were created programmatically. (Each QKset included a proof sheet, which combined all Q and K images for each QKset into a single image.)

Printing

- The images to be printed were resampled to 300ppi in ImageMagick using the Lanczos algorithm.
- For test impression images, levels were corrected in Adobe Photoshop (Black=40, Gamma=.75, White=210) to optimize for printing on 7mil resin-based inkjet transparency film (Inkpress #ITF17100) using Epson Surecolor P7000 24" printers.
- For all other images, levels were corrected (Black=0, Gamma=1.25, White=233) and they were sharpened in Adobe Photoshop (smart sharpen filter, amount: 90%, radius: 1 pixel) to optimize for printing on 100% alpha-cellulose glossy photographic paper (Noritsu paper #H07320400) using Fujifilm Frontier DL600 printers,.

Appendix C4.5 Selection and Assignment of QKsets

The overall test size is a tradeoff between four factors: the number of participants, the number of QKsets per participant, the total number of QKsets in the study, and the number of participants per QKset:

- In study design, a notable limitation is that the number of participants is not known well in advance, and the number of participants who will complete the study cannot be known until the end of the data collection period. We estimated that there would be at least 60 participants, with a likely value of about 80.
- The number of QKsets per participant is limited by the demands on participants. We asked for 100 comparisons per participant, which required a significant amount of time from the volunteers. In the frequently asked questions we made the following estimate: "We expect the amount of time to vary significantly from examiner to examiner, as well as by the difficulty of the comparison. We assume in general 15-30 minutes per comparison, so the total time we expect would be 25-50 hours over 12 months, or about ½ hour to 1 hour per week over the course of the year."
- A large total number of QKsets in the study is desirable to be as broadly representative as possible, and limit the effects of individual QKsets. Based upon experience gained in earlier forensic examiner studies, it is clear that sample-specific effects are a driver of accuracy and reproducibility rates: a small total number of QKsets risks individual QKsets having a disproportunate effect on results, and also limits the breadth of variation to be assessed.
- A large number of participants per QKset is desirable to assess reproducibility and measure participant-specific effects. If a study's only goal would be to compare participants, each QKset would be assigned to all participants — which limits the total number of QKsets in the study. In prior forensic examiner studies, we have observed that 20-30 participants per set provides a good target to measure reproducibility among participants.

In order to balance these tradeoffs, we determined that 100 comparisons per participant would be as much as we could reasonably request from volunteers, and estimated that there would be at least 60 participants, with a likely value of about 80. Based on this, we assigned each QKset to $^1/_3$ of the participants, which required a total of 300 QKsets (270 distinct QKsets, 30 repeated). Based on an estimate of 60-80 participants, we estimated that there would be about 20-26 participants per QKset. The actual counts were 84 participants; including repeats, responses totaled 16-54 trials per QKset (mean 24.6, median 23). Table S5 and Table S6 show the counts and proportions used in study design and data selection.

In the latent print black box study [20], assignments of image pairs (analogous to QKsets in this study) were randomized among participants. (To be precise, the assignments were based on a "basic incomplete block diagram" algorithm, but use of this method was hampered by not knowing the number of participants in advance, and not knowing the number of participants who would drop out of the study.) A limitation of that study was that we did not control the number of participants per image pair (which ranged from 2 to 37, median 23), the proportion of mated vs nonmated image pairs per participant (which ranged from 49% to 74% of assignments), or the difficulty or other attributes of the image pairs assigned to each participant (which meant that some participants may have had a much easier or more difficult test than others, and would have seen different proportions of types of image pairs). For these reasons, on subsequent studies [22,25,35,36], we have balanced assignments so that each sample is assigned to approximately the same number of participants, and each participant receives comparable proportions of sample-specific attributes, which in this study include mating, quality, and the method used to select nonmates.

In the latent print black box, nonmates were selected based on the FBI's national automated fingerprint identification system, which both provides a means to select nonmates that cannot be trivially excluded, and provides a means of selection that is directly applicable to operational casework. No such system exists for footwear. Selection of nonmates was therefore based on class similarity between the Q and K (as shown in Table S6): most nonmates were identical in make, model, and size; some nonmates were identical in make and model but differed by up to one size; some nonmates were of similar make and model.

Variable	*Design Value*	*Explanation*
Total QKsets/examiner	100	40 mated, 60 nonmated (if all assigned QKsets are completed)
Repeated QKsets/examiner	10	4 mated, 6 nonmated
Unique #QKsets/examiner	90	36 mated, 54 nonmated
Total QKsets	300	Including 30 repeated QKsets
Distinct QKsets	270	
%examiners/QKset	33%	
Distinct Qs	162	Same as QKsets-NM (the QKsets-M use the same Qs)
Distinct Ks	270	Same as QKsets
Distinct shoes needed	270-324	270 if each shoe used as a Q in a NM QKset is also used as a K in a (different) NM QKset 324 (1.2x distinct QKsets) if all Qs and Ks used in NM QKsets are distinct

Table S5. Fixed counts used in test sizing. Note these are the planned values used in study design: See *Appendix D1* for actual test yield.

Variable	*Description*	*Distinct QKsets*	
		%	*#*
QKsets-M	Mated QKsets	40%	108
QKsets-NM	Nonmated QKsets	60%	162
NM-sameClass	Nonmates with same make/model/size	37%	100
NM-diffSize	Nonmates with same make/model, different size	14%	36
NM-similar	Nonmates with similar make/model or different foot	10%	26

Table S6. Proportions used in test sizing and data selection.

The repeated assignments of QKsets were included in different packets (with different QKset numbers) to lessen the chance of participants recognizing the QKset contents: 88% of repeats had at least one intervening packet; 38% had at least two intervening packets. The repeated assignments were limited to the Eastern Mountain Sports Day Hiker boots and New Balance 878 shoes, based on the assumption that the less-frequently-used types of footwear might have been more readily recognizable.

Appendix C5 Participant Instructions

The following was summarized from the participant instructions. Detailed participant instructions were forwarded in hard copy directly to participants as part of each test packet. Complete instructions were also available via the study software online interface.

Appendix C5.1 Summary

As a participant in the Footwear Examiner Decision Analysis Study (aka Footwear Black Box Study), you will be asked to perform a total of 100 footwear evidence comparisons over a period of approximately one year. For each comparison set, you will be asked to compare one questioned impression with one known item of footwear, using printed photographs and or digital images. The printed photographs will be sent to you by FedEx. The Footwear Black Box Study (FBBS) website (https://footwear.idealinnovations.com) will provide you access to the digital images (available for download in both TIFF and JPEG formats) as well as a user interface for reporting your conclusions.

Appendix C5.2 General Guidance

When you first log into the FBBS site, you will be presented a terms and conditions screen. To participate you must agree to the following:

- to conduct the comparisons in this study with the same regard and diligence used when conducting footwear evidence comparisons in operational casework,
- not to conduct the comparisons in this study collaboratively,
- not to share or distribute the test materials associated with this study to anyone (including coworkers and colleagues),
- to discard all handwritten notes and/or printed materials prepared by you during this study at the completion of each packet,
- to delete all digital images downloaded by you at the completion of each packet,
- not to mark on any of the printed materials included in a comparison set, and
- to promptly return all printed materials in a packet as soon as you have submitted your conclusions for all of the comparison sets in that packet.

Appendix C5.3 Packets and Comparison Sets

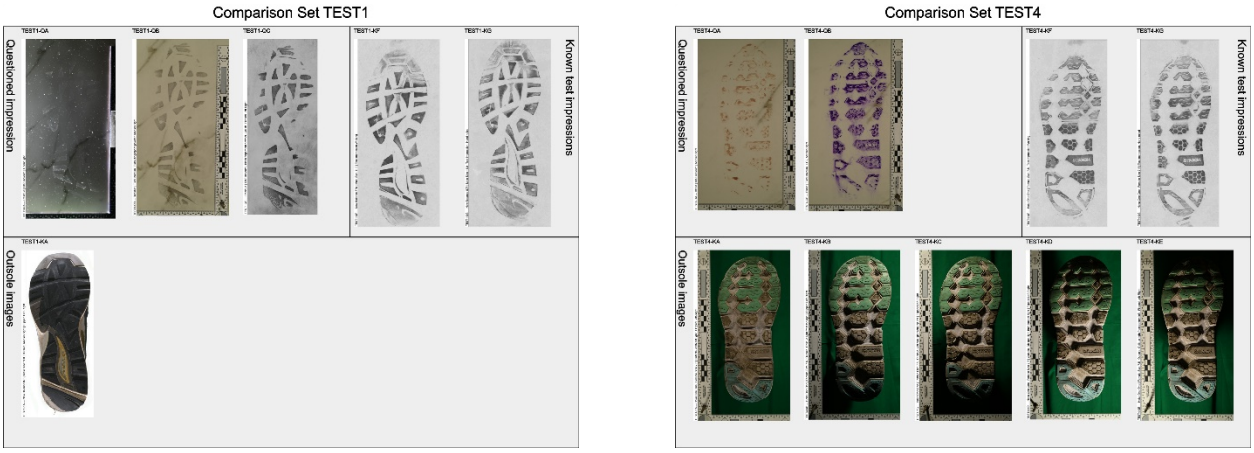

Fig. S1. Examples of proof sheets for two comparison sets.

You will complete 100 comparisons through the distribution of five ***packets***, each of which contains 20 *comparison sets*. You will receive each packet in a FedEx box, which will contain 20 comparison sets (each packaged in a separate envelope). Save the FedEx box since it will be used to return each packet using a prepaid shipping label. You will receive only one packet at a time.

Fig. S1 shows examples of two comparison sets. Each comparison set contains:

- A proof sheet showing exactly the images contained in the comparison set
- Up to three images of a single questioned impression (Q)
- Test impressions and outsole images of a single known item of footwear (K)
  - Each comparison set contains two test impressions, one walking and one hand rolled.
  - Each comparison set contains one or five outsole images:
    - *Most comparison sets contain five outsole images, one with ambient light, and four with oblique light from four different directions (top left, top right, bottom right, bottom left).*
    - *Some comparison sets contain one outsole image, which is a high-dynamic range (HDR) image produced by combining nine images with light from different directions.*

The two test impressions are printed on transparent media; all other images are printed on glossy photo paper.

All images depicting outsoles and lifts are reversed (i.e., the images were flipped horizontally) so that they orient with the impression on the ground. Such images are labelled "Reversed". For example, note in Fig. S1 "TEST1" that the features in all of the images are oriented in the same direction, which corresponds to a left shoe. Note that all Qs in a given comparison set depict the same impression. For example, in Fig. S1 "TEST4", the single Q is a bloody impression that was reproduced in two images (unenhanced and after processing with LCV).

Appendix C5.4 Comparison Determinations

You may conduct your comparisons using the printed images provided and/or by downloading the corresponding digital images from the FBBS website. You have the option of downloading the digital images in either TIFF or JPEG formats. Both the TIFF and JPEG images are high quality and contain the same number of pixels, but the JPEG images are less than half the size of the TIFF images.

Please do not mark on the printed materials (including the photographs) or envelopes so that they can be reused. If you accidentally mark on or damage any of the contents of a comparison set, please include a note with the specifics of the damaged item(s) to alert the study administrator so that the study administrator can replace the damaged item(s) prior to disseminating

the comparison set to another participant. Each comparison set has a *QKset number* (QK001 through QK999) that you will use to associate the printed and digital materials. The QKset number is on every image, proof sheet, and envelope.

You must report your decisions for each comparison set in the FBBS website. Your decisions will be recorded by responding to the questions and statements numbered 1 through 14. Note that your answers affect which questions you are shown. For example, if you select "1B. No" in question 1 (indicating that the questioned impression is not suitable for a meaningful comparison with the known item of footwear), questions/statements 2-7 will be hidden since they are not applicable. If you leave the FBBS website, your answers will be saved. Until you submit your responses, they are changeable so you may return to a comparison set and continue working. Once you submit your responses for a comparison set, they are final and submitted to the study administrator. You may not access submitted comparison sets again. In order to limit possible misunderstandings, please complete at least five comparisons before submitting any of them. The software will enforce this.

If any questions are unclear, please email the study administrator at footwear@idealinnovations.com for clarification before submitting.

#1 — Suitability	***Always shown***
1. Is the questioned impression suitable for a meaningful comparison with the known item of footwear?	Select the option that best characterizes your assessment of the suitability of the questioned impression for comparison with the known item of footwear in this set. An impression that lacks sufficient detail to enable a meaningful comparison with the known item of footwear is commonly referred to as "no value" or "unsuitable".
1A. Yes (...)	If you discerned impression detail (e.g., geometric shapes) in the questioned impression and determined that the detail was sufficient in size and quality to enable you to conduct a meaningful comparison with the known item of footwear provided in this set, select "Yes".
1B. No	If you were unable to discern any footwear impression detail or if you discerned impression detail in the questioned impression but you determined that the detail was insufficient in size and quality to enable you to conduct a meaningful comparison with the known item of footwear provided in this set, select "No".
#2 — Conclusion	***Only shown if (1) Suitability = (1A) Yes***
2. Select the most appropriate conclusion for this comparison set.	Select one of the following conclusions which best characterizes your determinations based on the comparison of the questioned impression and the known item of footwear in this set. Note that this conclusion scale is a modification of the SWGTREAD conclusion scale, which does not provide the opportunity for examiners to provide a completely neutral opinion: therefore, the "inconclusive" conclusion (2E) was added by the research team to accommodate that situation.
2A. Identification	This is the highest degree of association expressed by a footwear examiner. The questioned impression and the known item of footwear share agreement of class and randomly acquired characteristics of sufficient quality and quantity. In the opinion of the examiner, the particular known item of footwear is the source of (and made) the questioned impression. Another item of footwear being the source of the questioned impression is considered a practical impossibility.
2B. High degree of association	The questioned impression and the known item of footwear must correspond in the class characteristics of design, physical size, and general wear. For this degree of association, there must also exist: corresponding wear that, by virtue of its specific location, degree and orientation make it unusual, and/or one or more corresponding randomly acquired characteristics. In the opinion of the examiner, the characteristics observed exhibit a strong association between the questioned impression and known item of footwear; however, the quality and/or quantity were insufficient for an identification. Other footwear with the same characteristics observed in the questioned impression are included in the population of possible sources only if they display the same wear and/or randomly acquired characteristics observed in the questioned impression.
2C. Association of class characteristics	The class characteristics of both design and physical size must correspond between the questioned impression and the known item of footwear. Correspondence of general wear may also be present. In the opinion of the examiner, the known item of footwear is a possible source of the questioned impression and therefore could have produced the impression. Other footwear with the same class characteristics observed in the impression are included in the population of possible sources.
2D. Limited association of class characteristics	Some similar class characteristics were present; however, there were significant limiting factors in the questioned impression that did not permit a stronger association between the questioned impression and the known item of footwear. These factors may include but were not limited to: insufficient detail, lack of scale, improper position of scale, improper photographic techniques, distortion or significant lengths of time between the date of the occurrence and when the item of footwear was recovered that could account for a different degree of general wear. No confirmable differences were observed that could exclude the item of footwear. In the opinion of the examiner, factors (such as those listed above) have limited the conclusion to a general association of some class characteristics. Other footwear with the same class characteristics observed in the impression are included in the population of possible sources.
2E. Inconclusive	Similarities and/or differences may have been observed between the questioned impression and the known item of footwear, but significant limitations in the evidence prevented any specific association or non-association. In the opinion of the examiner, it could not be determined whether or not the known item of footwear is the source of the questioned impression.
2F. Indications of non-association	The questioned impression exhibits dissimilarities when compared to the known item of footwear; however, the details or features were not sufficiently clear to permit an exclusion. In the opinion of the examiner, dissimilarities between the questioned impression and the known item of footwear indicate non-association; however, the details or features were not sufficient to permit an exclusion.
2G. Exclusion	This is the highest degree of non-association expressed in footwear impression examinations. Sufficient differences were noted in the comparison of class and/or randomly acquired characteristics between the questioned impression and the known item of footwear. In the opinion of the examiner, the particular known item of footwear was not the source of, and did not make, the questioned impression.
#3 — Outsole Design	***Only shown if (1) Suitability = (1A) Yes***
3. Do the questioned impression and the known item of footwear correspond in outsole design?	Select one of the following options which best characterizes your determination regarding the correspondence of the gross outsole design* between the questioned impression and the known item of footwear in this comparison set.
3A. Yes, and the questioned impression was made by an item of footwear from the same foot.	If you determined that the features in the questioned impression correspond to the gross outsole design features on and orient with the known item of footwear, select 3A. This determination indicates that the questioned impression was made by an item of footwear that is of the same brand and model and from the same foot as the known item of footwear.
3B. Yes, but the questioned impression was made by an item of footwear from the opposite foot (i.e. right vs left).	If you determined that the features in the questioned impression correspond to the gross outsole design features on the known item of footwear, but it orients with the opposite foot (e.g., the questioned impression was made a left shoe but the known item of footwear is a right shoe), select 3B. This determination indicates that the questioned impression was made by an item of footwear that is of the same brand and model but from the opposite foot as the known item of footwear.
3C. No	If you determined that the features in the questioned impression do not correspond to (and are different than) the outsole design features on the known item of footwear, select 3C.
3D. Unsure	If you are unsure whether or not the features in the questioned impression correspond to the outsole design features on the known item of footwear, select 3D.

* *outsole design: the manufactured pattern or arrangement of design elements on a footwear outsole*

#4 — Mold Variations	***Only shown if (1) Suitability = (1A) Yes, and (3) Outsole Design = (3A) Yes (same outsole design, same foot)***
4. Did you observe any mold variations which indicate that the source of the questioned impression was made using a different mold than that used to produce the known item of footwear?	Select one of the following options which best characterizes your determination regarding the correspondence of specific outsole design features (i.e., mold variations[*]) between the questioned impression and the known item of footwear in this comparison set.
4A. Yes	If you selected 3A and you can distinguish the questioned impression from the known item of footwear based on mold variations, select 4A.
4B. No	If you selected 3A but you cannot distinguish the questioned impression from the known item of footwear based on mold variations, select 4B.
4C. Unsure	If you selected 3A but you are unsure if you can distinguish the questioned impression from the known item of footwear based on mold variations, select 4C.
#5 — Physical Size	***Only shown if (1) Suitability = (1A) Yes, and (3) Outsole Design = (3A) Yes (same outsole design, same foot)***
5. Do the questioned impression and the known item of footwear correspond in physical size?	Select one of the following options which best characterizes your determination regarding the correspondence of the physical size[†] between the questioned impression and the known item of footwear in this comparison set.
5A. Yes	If you determined that the features in the questioned impression correspond to the physical size features on the known item of footwear, select 5A.
5B. No	If you determined that the features in the questioned impression do not correspond to (and are different than) the physical size features on the known item of footwear, select 5B.
5C. Unsure	If you are unsure whether or not the features in the questioned impression correspond to the physical size features on the known item of footwear, select 5C.
#6 — Wear	***Only shown if (1) Suitability = (1A) Yes, and (3) Outsole Design = (3A) Yes (same outsole design, same foot)***
6. Do the questioned impression and the known item of footwear correspond in degree and position/location of wear?	Select one of the following options which best characterizes your determination regarding the correspondence of the degree[‡] and position/location[§] of wear [**] between the questioned impression and the known item of footwear in this comparison set.
6A. Yes	If you determined that the features in the questioned impression correspond to the degree and position/location of wear on the known item of footwear, select 6A.
6B. No	If you determined that the features in the questioned impression do not correspond to (and are different than) the degree and position/location of wear on the known item of footwear, select 6B.
6C. Unsure	If you are unsure whether or not the features in the questioned impression correspond to the degree and position/location of wear on the known item of footwear, select 6C.
#7 — Randomly Acquired Characteristics (RACs)	***Only shown if (1) Suitability = (1A) Yes, and (3) Outsole Design = (3A) Yes (same outsole design, same foot)***
7. If you observed any randomly acquired characteristics (RACs) that CORRESPOND between the known item of footwear and the questioned impression, click here to mark them.	If you observed any features in the questioned impression that CORRESPOND to RACs[††] on the known item of footwear in this comparison set, select this option and you will be provided the opportunity to mark them in a separate window. Only mark RACs if they are present in BOTH the questioned impression AND images of the known. Place a green circle on the outsole image using a left mouse click, making an effort to center the circle on the midpoint of the RAC. There is no relationship between the size of the green circle and the size of the RAC; it is simply a method for pinpointing the position of the RAC on the outsole. Repeat this process for all corresponding RACs. Use the "undo" and "clear" buttons to remove any unwanted circles. "Undo" removes one circle at a time in the order they were added. "Clear" removes all of the circles.
#8 — Difficulty	***Only shown if (1) Suitability = (1A) Yes***
8. Rate the difficulty associated with this comparison set.	Select one of the following options which best characterizes your perceived level of difficulty associated with the comparison of the questioned impression to the known item of footwear in this set.
8A. Very Easy / Obvious	
8B. Easy	
8C. Moderate	
8D. Difficult	
8E. Very Difficult	
#9 — Use of Software	***Always shown***
9. Did you use additional software (such as Adobe Photoshop) to view or process/enhance any of the high-resolution images in this comparison set?	Please indicate if and how you used software (other than the FBBS website) at all when conducting your comparison in this set.
9A. Yes, software was used to process/enhance one or more image(s)	
9B. Yes, software was used, but only to view images	
9C. No	
#10 — Use of Printed Photos	***Always shown***
10. Did you use the printed photographs/transparencies in making this comparison?	Please indicate if you used the printed photographs/transparencies provided in the envelope at all when conducting your comparison in this set.
10A. Yes	
10B. No	

[*] mold variation: a specific variation in the outsole design components between two molds of the same brand and model shoe, which can be evidenced by fewer outsole design elements, more outsole design elements, differences in the intersection of design elements with the edge of the outsole, and texture differences

[†] physical size: the dimensions, shapes, spacing and relative positions of the footwear outsole design components; this is not the same as the manufacturer's footwear size

[‡]degree of wear: the extent to which a footwear outsole has been eroded

[§]position/location of wear: a defined area of erosion on a footwear outsole

[**] wear: erosion of the surfaces of a footwear outsole during use

[††] randomly acquired characteristic: a feature on a footwear outsole resulting from random events (e.g., cut, scratch, tear, hole, stone hold, abrasion and the acquisition of debris)

#11 — Limitations	***Always shown***
11. Select any limitations associated with this comparison set.	Please indicate any limitations that kept you from making a more definitive conclusion OR that were a notable source of difficulty in making the comparison. Check all that apply. Leave blank if not applicable.
11A. Quality/clarity of the questioned impression	
11B. Insufficient quantity/area of outsole reproduced in the questioned impression	
11C. Distortion/movement in the questioned impression	
11D. Background/substrate interference in the questioned impression	
11E. Images/photographs of the questioned impression	
11F. Images/photographs of the outsole of the known item of footwear	
11G. Images/transparencies of the test impressions from the known item of footwear	
11H. Insufficient number of corresponding RACs	
11I. Lack of clarity of RACs in the questioned impression	
#12 — Typical of Casework	***Always shown***
12. Was the questioned impression in this comparison set typical of impressions encountered by you in operational casework?	Efforts were made by the research team to create questioned impressions which mimic those encountered in operational casework to include: varying the degree of quality, varying the quantity of information, introducing distortion/movement and background/substrate interference. We would like to know if these efforts were successful.
12A. Yes	
12B. Yes, but it is considered unusual and encountered infrequently	
12C. No	
#13 — Orientation	***Always shown***
13. Rotate the questioned impression until it is oriented with the toe pointing up.	Rotate the questioned impression until it is oriented with the toe pointing up or select one of the options below.
#14 — Comment	***Always shown***
14. Comment (optional)	Provide a comment only if there is an issue or a limitation in this comparison set that you could not adequately address using any of your responses above.

Appendix C5.5 Details Regarding the Samples Used in this Study

Questioned Impressions:

- All questioned impressions used in this study were collected under controlled conditions.
- There may be up to two weeks of wear between the dates the questioned impressions and the knowns (test impressions and outsoles) were collected.
- With the exception of the gel lifts, the questioned impressions were photographed using a Nikon D850 (outfitted with a Micro Nikkor 60mm f2.8 lens) mounted on either a tripod or a copy stand. Efforts were made to align the plane of the camera sensor with the plane of the impression. Either ambient light (available room light), flood lights, or oblique light (using a Foster+Freeman Crime-Lite 82L forensic light source) was utilized to illuminate the questioned impressions. A remote shutter release was used.
- All gel lifts were imaged in grayscale at 1044 pixels per inch (ppi) using the BVDA GLScan.
- There were no intentional efforts by the research team to create distortion by improperly photographing the questioned impressions.

Test impressions from known items of footwear:

- Each comparison set contains two test impressions, one walking (i.e., the powdered item of footwear was worn by an individual and he/she walked across an adhesive sheet) and one hand rolled (i.e., an adhesive sheet was applied manually to the powdered outsole while not being worn).
- All test impressions were imaged in grayscale at 600 ppi using an Epson Expression 10000XL Graphic Arts flatbed scanner.

Outsole images from known items of footwear:

- Most comparison sets contain five outsole images: one illuminated with ambient light, and four illuminated with oblique light from four different positions (top left, top right, bottom right, bottom left), lit by photoflood lights, captured using a Nikon D850 (Micro Nikkor 60mm f2.8 lens). For these outsole images, only the ambient light image was calibrated to 600ppi; the obliquely-lit outsoles should be identical but were not separately verified. Aperture priority (generally f16) and autofocus settings were employed. A remote shutter release was used.
- For the comparison sets that contain one outsole image of the known item of footwear, that image is a high-dynamic range (HDR) image produced by combining nine separate images captured using a Nikon D800 while illuminating the outsole obliquely from various angles.

Image Processing:

- Images were captured using a DSLR camera (either a Nikon D850 or D800), a BVDA GLScan or an Epson Expression 10000XL Graphic Arts flatbed scanner.

- The resolution of all digital images (TIFF and JPEG) is 600 ppi. Images captured with a DSLR camera were calibrated using the rulers in the images, then resampled to 600 ppi. The images generated using the BVDA GLScan were downsampled from 1044 ppi to 600 ppi.
- In images where the original rulers used for calibration could not be included in the distributed image (e.g. too far from the impression to be included after cropping), a ruler was digitally added (and labelled as such).
- All images were downsampled to 300 ppi prior to printing (both the photographs and the transparencies).

Overlapping Footwear Impressions:

- In some of the comparison sets, more than one footwear impression is depicted in the questioned impression image(s). In these cases, determine if there is a footwear impression which corresponds in outsole design to the known item of footwear provided in the set. If so, compare the footwear impression which bears the corresponding outsole design; ignore the remaining footwear impression(s). Your responses should be focused only the corresponding outsole design impression.

Appendix C6 Frequently Asked Questions (FAQ)

To guarantee that participants all received identical instructions during the study, no changes to instructions were made after the start of the test period. Any questions related to the study were answered in a Frequently Asked Questions (FAQ) document that was shared with all participants. The final version of the FAQ is included below (last updated 19 December 2018).

1. I would really like to do this but am a bit concerned about the amount of time required for 100 examinations. Any idea of the likely time involved?

We expect the amount of time to vary significantly from examiner to examiner, as well as by the difficulty of the comparison. We assume in general 15-30 minutes per comparison, so the total time we expect would be 25-50 hours over 12 months, or about ½ hour to 1 hour per week over the course of the year.

2. What if we start the study but are unable to complete it?

If you start and are unable to complete all 100 comparisons, we will include results for any participants who complete at least 20 comparisons. However, keep in mind that by not completing all 100 comparisons you will not be eligible for the monetary rewards.

3. Is the study open to non-English speakers?

Because the questions and instructions are all in English, participants must be reasonably fluent in English to participate in order to minimize the potential for misunderstandings.

4. How will the results be reported?

Results will be published in a peer-reviewed journal, and presented at appropriate forensic conferences.

5. Can we find out our own results?

No: we are required to keep the results anonymous. We have processes in place so that results are anonymous even within the analysis team.

6. Many of us are in government service, which either means we would not be able to accept money awards, or any money received would have to go the agency.

We understand. When we randomly select the award winners, they can tell us then if they would like the award personally, give it to their agency, or are unable to accept the award (in which case we will repeat until we select someone who accepts the award).

7. Will there be time restrictions on the completion of each of the 5 packets (i.e. must submit within a month of receipt etc). Likewise, if examiners are unable to complete one packet on time/schedule, are they able to complete the rest of the trial? *(added 4 Oct 2018)*

We will not have time restrictions per packet: when we receive one packet back from you, we will send out the next packet. At the conclusion of the testing period (intended to be 12 months) we will stop accepting responses and test materials must be returned, whether or not 100 comparisons have been completed.

8. Is there a preference for examiners that use a digital comparison method to be tested? We are currently transitioning from a manual methods (i.e. physical photographs/impressions with acetate overlays) to digital methods, but many of our examiners do not have significant experience with digital tools for markup/comparison. *(added 4 Oct 2018)*

No. There is no preference for how the participant performs the comparisons — manual or digital. The test materials are designed to accommodate examiners who perform only manual methods, only digital methods, or some combination thereof. The participant will not be required to markup the comparison set as traditionally done in operational casework (i.e., prepare case/bench notes and laboratory report). The participant will be required to answer predefined questions regarding each comparison set and mark any corresponding randomly acquired characteristics (RACs) via the online form accessed from the Footwear Black Box website.

Each comparison set will include natural size (I.e., life size, 1:1) photographs of the images within each set, excluding the known footwear test impressions. The known footwear test impressions will be provided natural size and on transparent material. The participant will have the opportunity to download digital images of all of the images associated with each comparison set. Within the online form, the participant will be asked if they used any additional software (e.g., Adobe Photoshop) to view or enhance/process any of the images in the comparison set.

9. Is the material all accessed and examined via computer and if so, will I need any particular software for this? Does this study require access to a sharefile site? Our current IT rules do not allow us to access sharefiles without high-level/special permissions, so I am interested in knowing that aspect ahead of time. *(added 4 Oct 2018)*

Physical photographs will be mailed to each participant. However, for those individuals who desire to have access to the source images, they are available for download via the Footwear Black Box website, which is accessible from most modern Internet browsers (Chrome, Firefox, or Internet Explorer). The study responses are collected via the Footwear Black box website; no additional software must be installed. As long as your agency's/employer's Internet firewalls do not prevent you from accessing the Footwear Black Box website, no additional permissions should be necessary.

10. May we keep the printed and digital images and use them for internal training? *(added 4 Oct 2018)*

No. Printed images must be returned for reuse, using the prepaid shipping labels provided. When each packet is returned, participants must attest that they have deleted all computer files and paper copies.

11. May we mark on the printed images? *(added 4 Oct 2018)*

No. We plan to reuse the printed images in packets sent to other participants, so any marks would interfere with that person's test. If you inadvertently mark or damage an image, please put a note in the envelope for that comparison test pointing it out to the study administrator.

12. What resolution are the images? *(added 4 Oct 2018)*

All of the images in each comparison set were calibrated to natural size (i.e., life size, 1:1) at 600 pixels per inch (ppi). These 600 ppi images are available for download via the Footwear Black Box website. The images were printed at 300 ppi.

13. Can trainee examiners participate? We have a number that have completed all the required footwear examination training and are performing casework, but their opinions are checked by an expert examiner prior to reporting. *(added 4 Oct 2018)*

Participation eligibility is defined in this way:

A forensic footwear examiner (FFE), for the purposes of this study, is defined as an individual who conducts forensic comparisons of questioned footwear impressions and known items of footwear and communicates their findings in written reports and during testimony in courts of law.

So, if you are authorized to issue laboratory reports in your name and to testify as an expert witness to your findings in courts of law, you're eligible.

14. How does this study differ from the white-box black-box study that was conducted by WVU last year? *(added 4 Oct 2018)*

This study is expected to be larger and more comprehensive than the West Virginia University (WVU) study. The details of the WVU study have not yet been publicly released.

15. What types of questioned impressions will be included? *(added 4 Oct 2018)*

This study will only include two-dimensional footwear impressions. Efforts have been made to produce impressions in a variety of substrates and matrices that mimic those encountered in operational casework. Efforts have been made to produce footwear impressions with varying levels of quality and varying amounts of the outsole reproduced. Also, many different types and models of footwear will be included in this test.

16. What kind of comparisons are expected (result in a written report, scale of conclusion, binary result)? *(added 4 Oct 2018)*

Responses will be entered via an online form at this study's website. For each comparison set, the participant will be required to complete a series of questions, selecting from pre-defined answers. If the participant determines that the questioned impression is suitable for a meaningful comparison with the known item of footwear, he/she will be requested to select the most appropriate conclusion from a predefined conclusion scale. (Refer to Question 17 for details of the conclusion scale.) Additionally, the participant will be asked other questions regarding their assessment of the comparison set, such as limitations, difficulty, and the extent of correspondence. No written report (like that produced during operation casework) will be generated during this study.

17. What conclusion scale does the test use? *(added 4 Oct 2018)*

For each comparison set, the participant will be asked to first determine if the questioned impression is suitable for a meaningful comparison with the known item of footwear. If the participant responds "yes", then he/she will be required to select the most appropriate conclusion from one of the following conclusions.

1. Identification
2. High degree of association (probably made)
3. Association of class characteristics (could have made)
4. Limited association of class characteristics
5. Inconclusive
6. Indications of non-association (probably did not make)
7. Exclusion (elimination)

The above conclusion scale is a modification of the SWGTREAD 2013 scale with the addition of "inconclusive". Inconclusive, for the purposes of this study, is defined as follows: "It could not be determined whether or not the known item of footwear is the source of the impression. No specific association or non-association was possible due to limitations in the evidence."

18. Are you interested in participation from international jurisdictions? You said "Non-U.S. examiners may participate if they use the SWGTREAD conclusion scale (either 2006 or 2013 scale)" — what if we use something similar or a variation? *(added 4 Oct 2018)*

Yes, we encourage non-U.S. examiners to participate. You are eligible to participate if you utilize a multi-point scale that was inspired by either of the SWGTREAD scales. (Refer to Question 17 for details of the conclusion scale.) This eligibility requirement is in place to ensure that participants are comfortable reporting their conclusions within the reporting constraints provided in this study.

19. If there are multiple examiners from a lab, would each be receiving separate comparison packets, or would it be ok to share the packets as long as each examiner works on the comparisons individually? *(added 19 Dec 2018)*

The packets are assigned to individual examiners, and may not be shared (or viewed) by others: it is important that the results are for individual examiners, and therefore there may be no input (even casual review) from others.

Appendix D Test Yield and Conclusion Rates

Appendix D1 QKset counts

A total of 270 distinct QKsets were created for this study: 108 mated and 162 nonmated (one of the mated QKsets was omitted from analyses; see *Appendix D4*). To assess repeatability (intra-examiner variation), 30 of these QKsets were assigned twice to the same participants (identical images, but different QKset numbers)— these reassignments are termed "2nd assignments" throughout and are generally omitted from analyses, unless otherwise specified. Note that one mated QKset was ultimately omitted because the questioned impression (on a glass substrate) was inadvertently flipped left-right when photographing, resulting in 269 distinct QKsets used in analyses.

Overall, each participant who completed the study was assigned 100 total QKsets for comparison, which was comprised of 90 distinct QKsets (36 mated and 54 nonmated) and 10 repeated QKsets (4 mated and 6 nonmated).

Appendix D2 Images and QKsets

To create the QKsets used in this study, 162 distinct questioned impressions were compiled. Of these, 108 were included as Qs in both a mated and a nonmated QKset in an effort to standardize the image quality and the attributes between mated and nonmated sets—note that the same participant never received both QKsets (mated and nonmated containing the same Q) as part of their assigned comparisons. The remaining 54 questioned impressions were used only in nonmated QKsets.

Appendix D3 Nonmate selection

In an effort to assess performance for nonmates with varying degrees of correspondence, we selected and assigned nonmates based on the following categories. Of the 162 distinct nonmated QKsets:

- 100 QKsets: the Q and K had the same make, model, and size
- 36 QKsets: the Q and K had the same make and model, but differed in up to one manufacturer size
  - 19 QKsets: difference of ½ manufacturer's size
  - 15 QKsets: difference of one manufacturer's size
  - 2 were male vs female footwear, with an effective difference of ½ size or less
- 24 QKsets: the Q and K differed in make or model (but were similar in design)
- 2 QKsets: the Q and K were from opposite feet (left vs right)*
  - 1 different make and model (but were similar in design)
  - 1 same make, model, and size†

Appendix D4 Response Counts

Table S7 details the total number of QKsets utilized in this study and the number of responses that were collected from participants; Table S8 delineates the number of QKsets and responses as a function of nonmate category. Each QKset was assigned to one third of the participants. The *Baseline Dataset* includes responses from 16-30 participants per QKset (mean 22.4, median 23) — overall (including repeats), responses were received from 16-54 participants per QKset (mean 24.6, median 23).

		All	*Mated*	*Nonmated*
QKsets	Distinct QKsets	269	107	162
	Repeated QKsets	30	12	18
	Total QKsets	299	119	180
Responses	Not repeated	5,454	2,189	3,265
	1st assignment	578	228	350
	2nd assignment	578	228	350
	Total responses	6,610	2,645	3,965
	Baseline — total responses, omitting 2nd assignment	6,032	2,417	3,615

Table S7. Counts of QKsets and responses both overall and by mating (mated/nonmated). The repeated QKsets were assigned to the same examiners twice (with identical images but different QK numbers).

		Same made/ model/size	*Same make/model, Different size*	*Different make/model*	*Different foot*
QKsets	Distinct QKsets	100	36	24	2
	Repeated QKsets	18	0	0	0
	Total QKsets	118	36	24	2
Responses	Not repeated	1,894	810	518	43
	1st assignment	350	0	0	0
	2nd assignment	350	0	0	0
	Total responses	2,594	810	518	43
	Baseline responses (omitting 2nd assignment)	2,244	810	518	43

Table S8. Counts of nonmated QKsets and responses by nonmate category. The repeated QKsets were assigned to the same examiners twice (with identical images but different QK numbers).

In total, we collected responses from 84 participants on 6,610 trials. This does not include three trials from two participants who each submitted only one or two QKsets; we also omitted 16 trials from one mated QKset for which the Q (on a glass substrate) was inadvertently reversed left-right when photographed.‡ Table S9 details the datasets used for analyses—unless otherwise specified, analyses are generally conducted on the *Baseline Dataset*.

---

* *During design and assignments, QKsets from opposite feet were considered part of different make/model, but were separated out during analysis because of the striking difference in results.*

† *In this nonmated QKset, the Q and K were both collected from a left shoe, but the Q was in advertently captured through glass and therefore was flipped left/right and appeared to be from a right foot.*

‡ *Note that although 55 participants completed all 100 assigned QKsets, after the omitted trials 14 of those 55 participants only had 99 trials used in analyses.*

Dataset	*Trials*	*Distinct QKsets*	*Repeated QKsets*	*Participants*	*Description*
Baseline Dataset	6,032	269	0	84	Default dataset for analyses—omits 2nd assignments
All Data	6,610	269	30	84	Includes 2nd assignments
Repeat Dataset	1,156	30	30	64	578 pairs of 1st and 2nd assignments
Examiner Comparison Dataset	5,749	269	0	71	Subset of *Baseline Dataset* for measuring individual rates and comparing examiners. Omits 13 participants who did not complete at least 40 QKsets each.
Reproducibility Dataset	6,032	269	0	84	132,074 inter-examiner decision pairs derived from a self join of the 6,032 responses in the *Baseline Dataset (*each individual response is paired with every other response on the same QKsets*)*

Table S9. Datasets used in analyses.

Table S10 shows the total response counts by conclusion category.

	Baseline data (omitting 2nd assignments)						*All data (including 2nd assignments)*					
	Responses		*Mated*		*Nonmated*		*Responses*		*Mated*		*Nonmated*	
ID	734	12.2%	725	30.0%	9	0.2%	842	12.7%	831	31.4%	11	0.3%
HighAssn	460	7.6%	410	17.0%	50	1.4%	504	7.6%	451	17.1%	53	1.3%
Assn	1534	25.4%	682	28.2%	852	23.6%	1721	26.0%	736	27.8%	985	24.8%
LimitedAssn	887	14.7%	312	12.9%	575	15.9%	957	14.5%	328	12.4%	629	15.9%
Inc	197	3.3%	60	2.5%	137	3.8%	206	3.1%	62	2.3%	144	3.6%
NotSuitable	170	2.8%	39	1.6%	131	3.6%	171	2.6%	39	1.5%	132	3.3%
NonAssn	389	6.4%	43	1.8%	346	9.6%	440	6.7%	47	1.8%	393	9.9%
Excl	1661	27.5%	146	6.0%	1515	41.9%	1769	26.8%	151	5.7%	1618	40.8%
Definitive (ID & Excl)	2395	39.7%	871	36.0%	1524	42.2%	2611	39.5%	982	37.1%	1629	41.1%
Probable (HighAssn & NonAssn)	849	14.1%	453	18.7%	396	11.0%	944	14.3%	498	18.8%	446	11.2%
Class Assoc (Assn & LimitedAssn)	2421	40.1%	994	41.1%	1427	39.5%	2678	40.5%	1064	40.2%	1614	40.7%
Neutral (Inc & NotSuitable)	367	6.1%	99	4.1%	268	7.4%	377	5.7%	101	3.8%	276	7.0%
Total	6032		2417		3615		6610		2645		3965	

Table S10. Responses by conclusion category.

Appendix D5 Accuracy and Error Rates

This appendix provides support for Section 4.1, Conclusion Rates, Accuracy, and Errors.

Although measures of accuracy, error rates, and predictive values are longstanding and used universally in a variety of fields, they were originally formulated and are generally applied to binary decision tasks with explicit "positive" and "negative" outcomes. However, implementation becomes more ambiguous for decision tasks involving more than two levels. In the forensic literature, there has been disagreement about how to handle "inconclusive" responses for three-level conclusion scales and whether they should be included in the denominator for computing accuracy and error rates (see the following sources for some examples/discussion: [2,20,21,37]). However, in 2018 the OSAC Human Factors Committee proffered the following guidance for reporting such results for performance tests of forensic examiners using three-level conclusion scales [37]:

> *Importantly, false positives and false negatives are reported three ways: (1) as a percentage of all presentations (% PRES); (2) as a percentage of all comparisons, i.e., excluding those comparisons where the impressions were deemed to be of no value (% COMP); and (3) as a percentage of all conclusive calls, i.e., excluding both no value comparisons and inconclusive (% CALLS). PCAST advocates reporting error rate data as a percentage of conclusive calls (ignoring no value and inconclusive comparison), on grounds that cases where examiners reached a conclusion are those likely to be used in a criminal proceeding, and hence the rates of error for those conclusions are most relevant. Our view is that forensic scientists should be prepared to present error rate data for their methods in a variety of ways.*

More recently, the OSAC Human Factors Committee expanded their guidance to include larger categorical scales (e.g., five or seven level scales) and recommend an analogous approach for computing accuracy, wherein %CALLS includes the additional non-inconclusive categories [21]. For this study, we have opted to implement and expand upon the recommendations outlined by the OSAC Human Factors Committee, and we present accuracy and error rates for our eight-level modified SWGTREAD conclusion scale as follows:

- %PRES: includes all presentations in the denominator (does not omit any trials)
- %COMP: includes all comparisons in the denominator (omits any trials resulting in a determination of *NotSuitable*)
- %CALLS: includes all non-neutral calls in the denominator (omits any trials resulting in a determination of *NotSuitable* or a decision of *Inc*)
- %DEF: includes only definitive conclusions in the denominator (omits any trials resulting in a response of *NotSuitable, HighAssn*, *Assn*, *LimitedAssn*, *Inc*, or *NonAssn*)

Note that the %DEF metric was added for the purposes of this study.

Table S11 details the distribution of conclusions for mated and nonmated trials in the *Baseline Dataset*; Table S12 delineates conclusions by nonmate category.

	Mated					*Nonmated*				
	#	*%PRES*	*%COMP*	*%CALLS*	*%DEF*	*#*	*%PRES*	*%COMP*	*%CALLS*	*%DEF*
Excl	146	6.0%	6.1%	6.3%	16.8%	1,515	41.9%	43.5%	45.3%	99.4%
NonAssn	43	1.8%	1.8%	1.9%	---	346	9.6%	9.9%	10.3%	---
Not Suitable	39	1.6%	---	---	---	131	3.6%	---	---	---
Inc	60	2.5%	2.5%	---	---	137	3.8%	3.9%	---	---
LimitedAssn	312	12.9%	13.1%	13.5%	---	575	15.9%	16.5%	17.2%	---
Assn	682	28.2%	28.7%	29.4%	---	852	23.6%	24.5%	25.5%	---
HighAssn	410	17.0%	17.2%	17.7%	---	50	1.4%	1.4%	1.5%	---
ID	725	30.0%	30.5%	31.3%	83.2%	9	0.2%	0.3%	0.3%	0.6%
Total Presentations (PRES)	2,417					3,615				
Total Comparisons (COMP)	2,378					3,484				
Total Calls (CALLS)	2,318					3,347				
Total Definitives (DEF)	871					1,524				

Table S11. Response counts by conclusion. Errors and incorrect conclusions are highlighted; neutral and debatable conclusions are shown in gray. (*Baseline Dataset*)

	Nonmated																			
	Same make/model/size					*Same make/model, ± ½-1 size*					*Different make/model*					*Different foot*				
	#	*%PRES*	*%COMP*	*%CALLS*	*%DEF*	*#*	*%PRES*	*%COMP*	*%CALLS*	*%DEF*	*#*	*%PRES*	*%COMP*	*%CALLS*	*%DEF*	*#*	*%PRES*	*%COMP*	*%CALLS*	*%DEF*
Excl	724	32.3%	33.0%	34.4%	98.9%	367	45.3%	46.9%	48.6%	99.7%	382	73.7%	82.2%	86.0%	100.0%	42	97.7%	97.7%	97.7%	100.0%
NonAssn	246	11.0%	11.2%	11.7%	---	78	9.6%	10.0%	10.3%	---	21	4.1%	4.5%	4.7%	---	1	2.3%	2.3%	2.3%	---
Not Suitable	51	2.3%	---	---	---	27	3.3%	---	---	---	53	10.2%	---	---	---	0	0.0%	---	---	---
Inc	88	3.9%	4.0%	---	---	28	3.5%	3.6%	---	---	21	4.1%	4.5%	---	---	0	0.0%	0.0%	---	---
LimitedAssn	405	18.0%	18.5%	19.2%	---	135	16.7%	17.2%	17.9%	---	35	6.8%	7.5%	7.9%	---	0	0.0%	0.0%	0.0%	---
Assn	676	30.1%	30.8%	32.1%	---	170	21.0%	21.7%	22.5%	---	6	1.2%	1.3%	1.4%	---	0	0.0%	0.0%	0.0%	---
HighAssn	46	2.0%	2.1%	2.2%	---	4	0.5%	0.5%	0.5%	---	0	0.0%	0.0%	0.0%	---	0	0.0%	0.0%	0.0%	---
ID	8	0.4%	0.4%	0.4%	1.1%	1	0.1%	0.1%	0.1%	0.3%	0	0.0%	0.0%	0.0%	0.0%	0	0.0%	0.0%	0.0%	0.0%
Total PRES	2,244					810					518					43				
Total COMP	2,193					783					465					43				
Total CALLS	2,105					755					444					43				
Total DEF	732					368					382					42				

Table S12. Response counts for nonmated trials by category. Errors and incorrect conclusions are highlighted; neutral and debatable conclusions are shown in gray. (*Baseline Dataset*. Responses for 162 distinct nonmated QKsets: 100 identical, 36 different size, 24 different make or model, 2 different foot)

Table S13 and Table S14 report the accuracy rates and error rates of conclusions reached by participating FFEs in this study. In addition to point estimates for each rate, we also provide confidence intervals for each metric. Confidence intervals (CIs) are reported using Clopper-Pearson, a commonly utilized binomial CI approach that produces conservative estimates of the interval [38]. Note that since rates are not evenly distributed by QKset or by participant (heteroscedastic), any approach for measuring confidence intervals is necessarily imperfect. Furthermore, the Clopper-Pearson estimate, like most other CI methods, assumes independence among decisions; because our data includes commonalities of examiners and image pairs, we expect the confidence intervals presented here may be narrower than appropriate for the data.

Metric	*Abbreviation*	*Definition*	*Rate*	*C.I.*	*Counts*
True positive rate		Proportion of mated QKset trials resulting in *IDs*			
	TPR_{PRES}	(including all mated QKset presentations in the denominator)	30.0%	[28.2%-31.9%]	(725/2417)
	TPR_{COMP}	(omitting mated QKset presentations that resulted in *NotSuitable*)	30.5%	[28.6%-32.4%]	(725/2378)
	TPR_{CALLS}	(omitting mated QKset presentations that resulted in *NotSuitable* or *Inc*)	31.3%	[29.4%-33.2%]	(725/2318)
	TPR_{DEF}	(including only mated QKset presentations that resulted in *ID* or *Excl*)	83.2%	[80.6%-85.7%]	(725/871)
Correct association rate*		Proportion of mated QKset trials resulting in *HighAssns*			
	CAR_{PRES}	(including all mated QKset presentations in the denominator)	17.0%	[15.5%-18.5%]	(410/2417)
	CAR_{COMP}	(omitting mated QKset presentations that resulted in *NotSuitable*)	17.2%	[15.7%-18.8%]	(410/2378)
	CAR_{CALLS}	(omitting mated QKset presentations that resulted in *NotSuitable* or *Inc*)	17.7%	[16.2%-19.3%]	(410/2318)
True positive + correct association rate		Proportion of mated QKset trials resulting in *IDs* or *HighAssns*			
	$TPR+CAR_{PRES}$	(including all mated QKset presentations in the denominator)	47.0%	[45.0%-49.0%]	(1135/2417)
	$TPR+CAR_{COMP}$	(omitting mated QKset presentations that resulted in *NotSuitable*)	47.7%	[45.7%-49.8%]	(1135/2378)
	$TPR+CAR_{CALLS}$	(omitting mated QKset presentations that resulted in *NotSuitable* or *Inc*)	49.0%	[46.9%-51.0%]	(1135/2318)
True negative rate		Proportion of nonmated QKset trials resulting in *Excls*			
	TNR_{PRES}	(including all nonmated QKset presentations in the denominator)	41.9%	[40.3%-43.5%]	(1515/3615)
	TNR_{COMP}	(omitting nonmated QKset presentations that resulted in *NotSuitable*)	43.5%	[41.8%-45.1%]	(1515/3484)
	TNR_{CALLS}	(omitting nonmated QKset presentations that resulted in *NotSuitable* or *Inc*)	45.3%	[43.6%-47.0%]	(1515/3347)
	TNR_{DEF}	(including only nonmated Qkset presentations that resulted in *ID* or *Excl*)	99.4%	[98.9%-99.7%]	(1515/1524)
Correct non-association rate*		Proportion of nonmated QKset trials resulting in *NonAssns*			
	CNR_{PRES}	(including all nonmated QKset presentations in the denominator)	9.6%	[8.6%-10.6%]	(346/3615)
	CNR_{COMP}	(omitting nonmated QKset presentations that resulted in *NotSuitable*)	9.9%	[9.0%-11.0%]	(346/3484)
	CNR_{CALLS}	(omitting nonmated QKset presentations that resulted in *NotSuitable* or *Inc*)	10.3%	[9.3%-11.4%]	(346/3347)
True negative + correct non-association rate		Proportion of nonmated QKset trials resulting in *IDs* or *NonAssns*			
	$TNR+CNR_{PRES}$	(including all nonmated QKset presentations in the denominator)	51.5%	[49.8%-53.1%]	(1861/3615)
	$TNR+CNR_{COMP}$	(omitting nonmated QKset presentations that resulted in *NotSuitable*)	53.4%	[51.7%-55.1%]	(1861/3484)
	$TNR+CNR_{CALLS}$	(omitting nonmated QKset presentations that resulted in *NotSuitable* or *Inc*)	55.6%	[53.9%-57.3%]	(1861/3347)

Table S13. Summary accuracy rates and Clopper-Pearson 95% confidence intervals. Numerators and denominators for each calculation are show to avoid ambiguities. Starred metrics were developed for this study to accommodate the use of a multilevel conclusion scale. (*Baseline Dataset*)

Metric	*Abbreviation*	*Definition*	*Rate*	*C.I.*	*Counts*
False positive rate		Proportion of nonmated QKset trials resulting in *IDs*			
	FPR_{PRES}	(including all nonmated QKset presentations in the denominator)	0.2%	[0.1%-0.5%]	(9/3615)
	FPR_{COMP}	(omitting nonmated QKset presentations that resulted in *NotSuitable*)	0.3%	[0.1%-0.5%]	(9/3484)
	FPR_{CALLS}	(omitting nonmated QKset presentations that resulted in *NotSuitable* or *Inc*)	0.3%	[0.1%-0.5%]	(9/3347)
	FPR_{DEF}	(including only nonmated Qkset presentations that resulted in *ID* or *Excl*)	0.6%	[0.3%-1.1%]	(9/1524)
Incorrect association rate*		Proportion of nonmated QKset trials resulting in HIghAssn*s*			
	IAR_{PRES}	(including all nonmated QKset presentations in the denominator)	1.4%	[1.0%-1.8%]	(50/3615)
	IAR_{COMP}	(omitting nonmated QKset presentations that resulted in *NotSuitable*)	1.4%	[1.1%-1.9%]	(50/3484)
	IAR_{CALLS}	(omitting nonmated QKset presentations that resulted in *NotSuitable* or *Inc* from the denominator)	1.5%	[1.1%-2.0%]	(50/3347)
False positive + incorrect association rate		Proportion of nonmated QKset trials resulting in *IDs* or HighAssn*s*			
	$FPR+IAR_{PRES}$	(including all nonmated QKset presentations in the denominator)	1.6%	[1.2%-2.1%]	(59/3615)
	$FPR+IAR_{COMP}$	(omitting nonmated QKset presentations that resulted in *NotSuitable*)	1.7%	[1.3%-2.2%]	(59/3484)
	$FPR+IAR_{CALLS}$	(omitting nonmated QKset presentations that resulted in *NotSuitable* or *Inc*)	1.8%	[1.3%-2.3%]	(59/3347)
False negative rate		Proportion of mated QKset trials resulting in *Excls*			
	FNR_{PRES}	(including all mated QKset presentations in the denominator)	6.0%	[5.1%-7.1%]	(146/2417)
	FNR_{COMP}	(omitting mated QKset presentations that resulted in *NotSuitable*)	6.1%	[5.2%-7.2%]	(146/2378)
	FNR_{CALLS}	(omitting mated QKset presentations that resulted in *NotSuitable* or *Inc*)	6.3%	[5.3%-7.4%]	(146/2318)
	FNR_{DEF}	(including only mated QKset presentations that resulted in *ID* or *Excl*)	16.8%	[14.3%-19.4%]	(146/871)
Incorrect non-association rate*		Proportion of mated QKset trials resulting in *NonAssns*			
	INR_{PRES}	(including all mated QKset presentations in the denominator)	1.8%	[1.3%-2.4%]	(43/2417)
	INR_{COMP}	(omitting mated QKset presentations that resulted in *NotSuitable*)	1.8%	[1.3%-2.4%]	(43/2378)
	INR_{CALLS}	(omitting mated QKset presentations that resulted in *NotSuitable* or *Inc*)	1.9%	[1.3%-2.5%]	(43/2318)
False negative + incorrect non-association rate		Proportion of mated QKset trials resulting in *IDs* or *NonAssns*			
	$FNR+INR_{PRES}$	(including all mated QKset presentations in the denominator)	7.8%	[6.8%-9.0%]	(189/2417)
	$FNR+INR_{COMP}$	(omitting mated QKset presentations that resulted in *NotSuitable*)	7.9%	[6.9%-9.1%]	(189/2378)
	$FNR+INR_{CALLS}$	(omitting mated QKset presentations that resulted in *NotSuitable* or *Inc*)	8.2%	[7.1%-9.3%]	(189/2318)

Table S14. Summary error rates and Clopper-Pearson 95% confidence intervals. Numerators and denominators for each calculation are show to avoid ambiguities. Starred metrics were developed for this study to accommodate the use of a multilevel conclusion scale. (*Baseline Dataset*)

Although accuracy and error rates are important measures for characterizing the performance of FFEs, they require *a priori* knowledge of ground truth regarding the source of the questioned impression. In casework, this information is unknown. Instead, the quantity of interest becomes the likelihood that a reported conclusion is in fact correct, which can be computed using posterior probabilities— as provided in Table S15 and Table S16. These posterior probabilities modify the error rates by the mate prevalence or the proportion of trials that were mated (see [20] for a more detailed discussion and necessary equations) in order to determine the chance that a conclusion is correct (e.g., the chance that a decision of *ID* was reported on a mated trial). Fig S3 plots PPV and NPV across a range of mate prevalences (from 0% to 100%) since this proportion can vary between laboratories and depending upon the case factors.

	Metric	*Abbr*	*Definition*	*Rate*	*Counts*	*Rescaled (50:50 MP)*
Baseline Dataset	Positive predictive value	PPV	Proportion of *IDs* that were correct (i.e. on mated QKsets)	98.8%	(725/734)	99.2%
	False discovery rate	FDR	Proportion of *IDs* that were incorrect (i.e. on nonmated QKsets); (1-PPV)	1.2%	(9/734)	0.8%
	Positive predictive association	PPA	Proportion of *IDs* and *HighAssns* that were correct (i.e. on mated Qksets)	95.1%	(1135/1194)	96.6%
	Negative predictive value	NPV	Proportion of *Excls* that were correct (i.e. on nonmated QKsets)	91.2%	(1515/1661)	87.4%
	False omission rate	FOR	Proportion of *Excl* that were incorrect (i.e. on mated QKsets); (1-NPV)	8.8%	(146/1661)	12.6%
	Negative predictive association	NPA	Proportion of *Excls* and *NonAssns* that were correct (i.e., on nonmated QKsets)	90.8%	(1861/2050)	86.8%

Table S15. Posterior probabilities of accuracy and error. Rates are based upon 40.1% mate prevalence for responses in the *Baseline Dataset*, and are also shown rescaled to 50:50 mate prevalence (MP).

	Metric	*Abbr*	*Definition*	*Rate*	*Counts*	*Rescaled (50:50 MP)*
Same make, model, and size	Positive predictive value	PPV	Proportion of *IDs* that were correct (i.e. on mated QKsets)	98.9%	(725/733)	98.8%
	False discovery rate	FDR	Proportion of *IDs* that were incorrect (i.e. on nonmated QKsets); (1-PPV)	1.1%	(8/733)	1.2%
	Positive predictive association	PPA	Proportion of *IDs* and *HighAssns* that were correct (i.e. on mated Qksets)	95.5%	(1135/1189)	95.1%
	Negative predictive value	NPV	Proportion of *Excls* that were correct (i.e. on nonmated QKsets)	83.2%	(724/870)	84.2%
	False omission rate	FOR	Proportion of *Excl* that were incorrect (i.e. on mated QKsets); (1-NPV)	16.8%	(146/870)	15.8%
	Negative predictive association	NPA	Proportion of *Excls* and *NonAssns* that were correct (i.e., on nonmated QKsets)	83.7%	(970/1159)	84.7%

Table S16. Posterior probabilities of accuracy and error, limited to QKsets in which the Q and K were of the same make, model, and size (i.e. all mates, but a subset of nonmates from *Baseline Dataset*; 51.9% mate prevalence). Since accuracy was notably higher for nonmated QKsets that differed in make/model, NPV and NPA are notably lower as compared to Table S16 (n=2,244 nonmated trials; 2,417 mated trials).

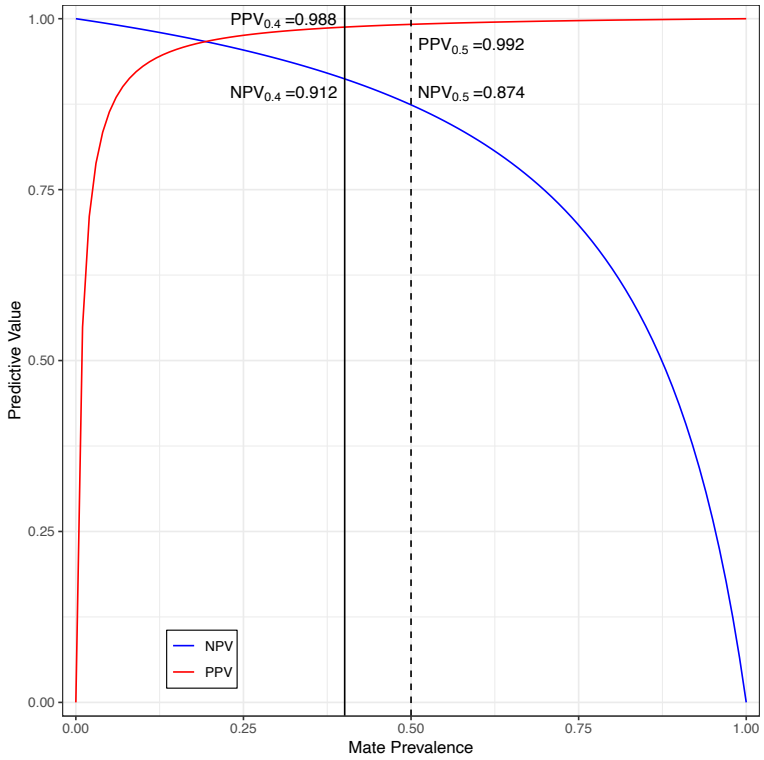

Fig S3. Positive (red) and negative (blue) predictive values as a function of mate prevalence. The solid line indicates the mate prevalence for the *Baseline Dataset* (40.1% mated trials, 59.9% nonmated trials); the dashed line indicates a mate prevalence of 50.0%, which assumes uninformative priors for the posterior probability estimate, as is often the case for operational casework.

Appendix E Errors and Incorrect Conclusions

Appendix E1 Erroneous IDs — False Positives (FPs)

Eleven erroneous IDs occurred in this study: the *Baseline Dataset* includes nine erroneous IDs, and an additional two FPs occurred in the repeatability data (responses to 2nd assignments not included in the *Baseline Dataset*). For the *Baseline Dataset*, this yields a false positive rate (FPR_{PRES}) of 0.2% (9 out of 3615 nonmated trials). We do not include second responses in calculating rates because if we did some QKsets would be counted twice, resulting in a biased rate.

The FPs were made by just four participants, but two participants made multiple such errors:

- One participant made six FPs: five in the *Baseline dataset*, and an additional one in the repeatability data. This participant also had the highest rate of incorrect *HighAssn* responses.
- One participant made three FPs: two in the *Baseline dataset*, and an additional one in the repeatability data.
- Two participants each made one FP in the *Baseline dataset*.

Both of the participants who made more than one FP have at least five years of experience, conduct footwear examinations less than weekly, completed a formal training program lasting 6-12 months, are not IAI certified, work for a non-US government agency, have not completed a proficiency test in the past year, and infrequently or never report IDs in casework; a total of six participants meet these criteria, but none of the other four reported any erroneous IDs.

One of the participants who made one FP also had the highest rate of erroneous *Excls* (14 FNs; FNR=39%). The other participants who made FPs made no FNs. See Figure 5 for the associations between error rates for these participants.

With respect to QKsets, the false positives in the *Baseline Dataset* were reported on eight distinct QKsets, with one QKset yielding two erroneous *ID*s (QK203, Figure 1). This QKset presented a full soil/dust questioned impression on tile/linoleum (quality grade A) and contained superimposed impressions. Of the 19 participants assigned this QKset, 14 indicated that it was typical of casework; regarding difficulty of the comparison, 5 indicated it was Easy, 12 Moderate, 1 Difficult, and 1 Very difficult.

Appendix E2 Incorrect HighAssns — Incorrect Associations (IAs)

Overall, participants in this study incorrectly reported *HighAssn* 50 times across all 3,615 nonmated trials in the *Baseline Dataset*, yielding an incorrect association rate (IAR_{PRES}) of 1.4%. Regarding nonmate categories:

- 46 incorrect associations were reported on nonmated trials presenting known footwear with the same make/model/size
- 4 incorrect associations were reported on nonmated trials presenting known footwear with the same make/model but different size
- 0 incorrect associations were reported on nonmated trials presenting known footwear with different make/model or foot

The incorrect associations were not limited to just a few participants. Rather, 26 participants made at least one incorrect association– 32% (16/50) of these were reported by just 3 participants (3.6% of all participants). One participant reported 10 incorrect associations in the *Baseline Dataset*; this was the same participant who rendered 6 total FPs.

With respect to QKsets, 32 out of 162 nonmated QKsets resulted in at least one incorrect association. One QKset accounted for 8 *HighAssn* conclusions (highest in this study) in the *Baseline Dataset*, but no false positives. This QKset (see Fig S4) was a close nonmate presenting known footwear with the same make/model/size; the questioned impression was a new shoe. This QKset presented a full impression (quality grade B) and had an average difficulty rating of moderate; 71% of participants assigned this QKset indicated that it was typical of casework

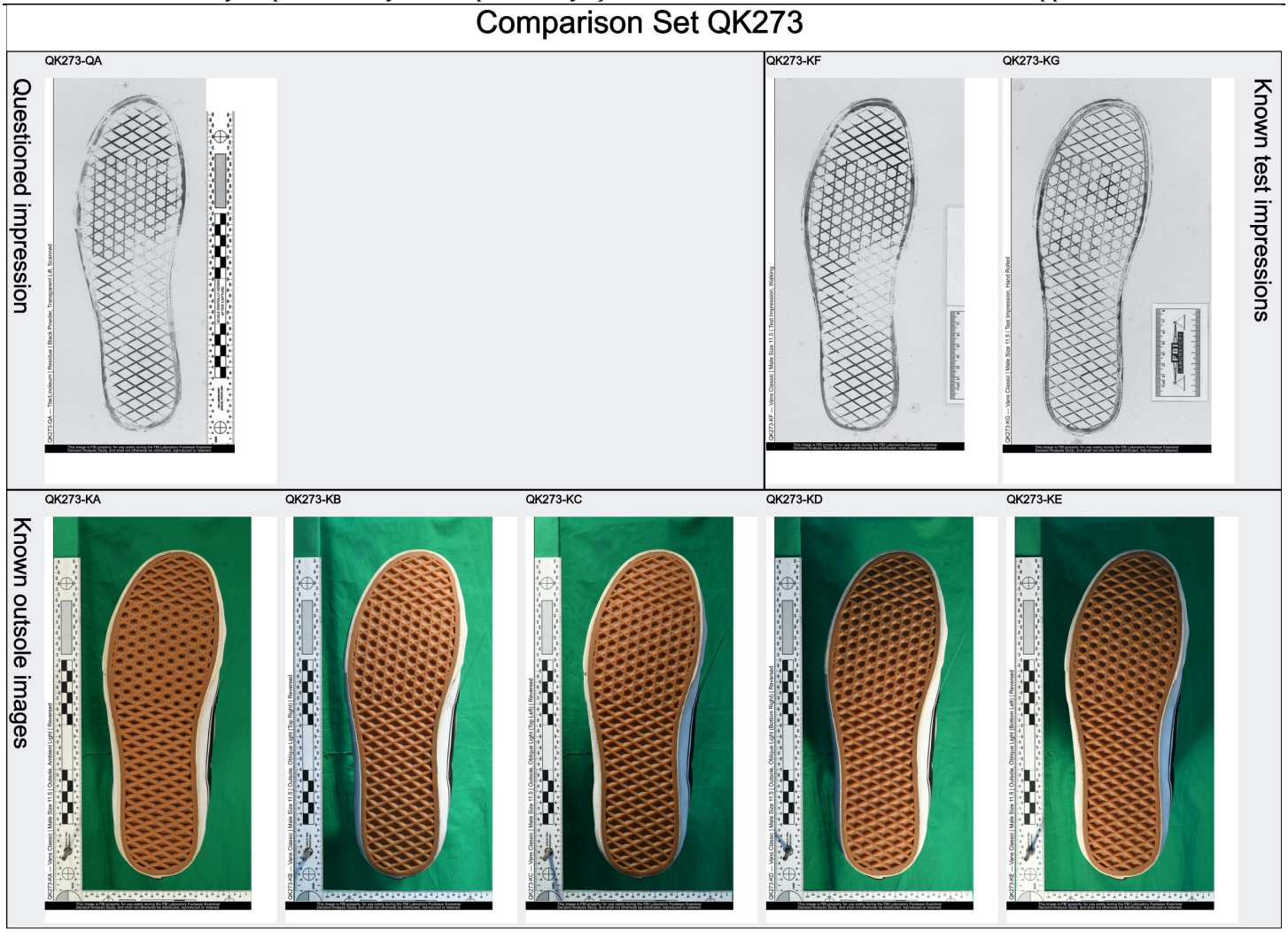

Fig S4. QK273: nonmated QKset (close nonmated; same make/model/size) that resulted in 8 IAs in the *Baseline Dataset* (and was not included in the *Repeat Dataset*). Note questioned impression was created using a new shoe. Conclusion rates for this QKset: 0 *ID*, 8 *HighAssn*, 16 *Assn*, 0 *LimitedAssn*, 0 *Inc*, 0 *NotSuitable*, 0 *NonAssn*, 0 *Excl*.

Appendix E3 Erroneous Excls — False Negatives (FNs)

Overall, participants in this study erroneously reported *Excl* 146 times across all 2,417 mated trials in the *Baseline Dataset*, yielding a false negative rate (FNR_{PRES}) of 6.0%. The repeatability data included an additional five erroneous *Excls.*

The false negatives were not limited to just a few participants; two-thirds of all participants (56/84) reported at least one erroneous *Excl* conclusion in the *Baseline Dataset*. Although these errors were generally widespread, one participant made 14 false negative errors. This participant has more than 10 years of experience, conducts a few footwear examinations yearly, has testified as a footwear expert, did not complete a formal training program, is not IAI certified, and infrequently or never reports *Excl*s in casework; four other participants meet these criteria and did not exhibit high individual false negative rates (one reported just a single erroneous *Excl*, and the remaining three did not report any).

With respect to erroneous exclusions, 60% of mated QKsets yielded at least one false negative in the *Baseline Dataset*, but some had disproportionately higher frequencies. In particular, a single mated QKset accounted for 20 FNs (QK213; Figure 8 in the main paper, detail in Fig S5). QK213 had a FNR of 77% and presented a full blood questioned impression on cloth (quality grade F) and had an average difficulty rating of easy; 62% of participants assigned this QKset indicated that it was typical of casework. Note that the only other mated QKset on cloth resulted in a single erroneous *Excl* (5% FNR). The second highest FNR (44%, QK083) was on a QKset on plastic sheeting. Future studies may consider evaluating the effects of such malleable substrates in examinations.

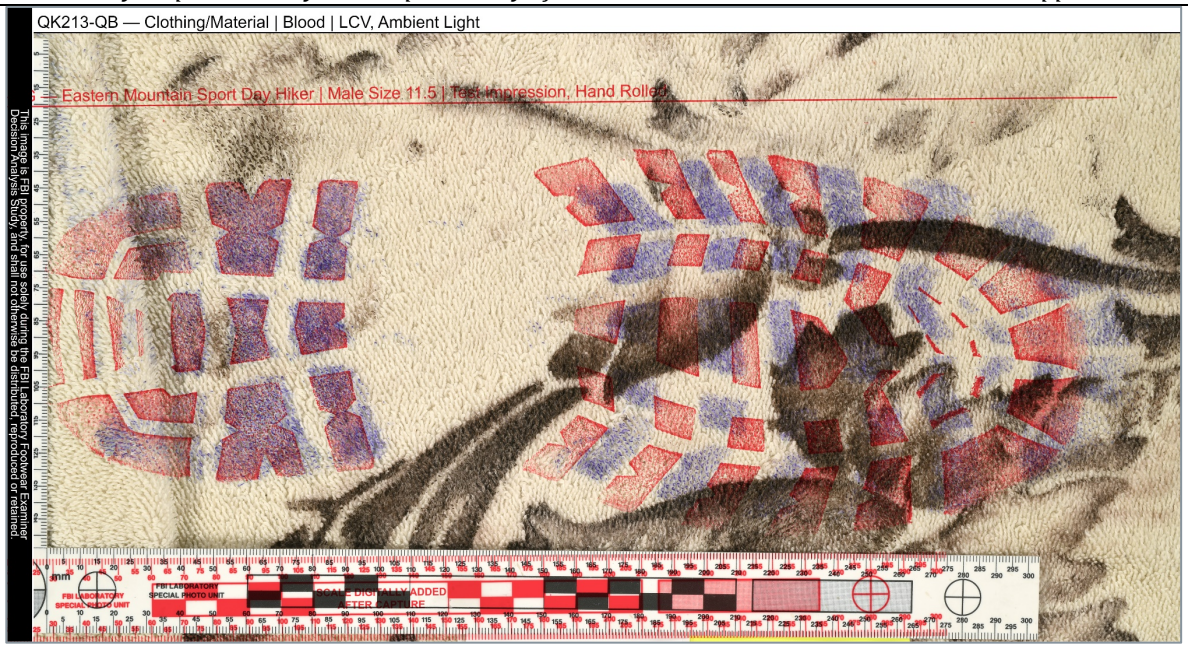

Fig S5. Detail of QK213, which resulted in 20 FNs. See Figure 8 (main paper) for additional images for this QKset. KG (colored red) superimposed on QB (blood impression on terry cloth towel, processed with LCV), showing that the physical dimensions of the Q impression are notably different from those of the test impression.

Appendix E4 Incorrect Non-associations

Overall, participants in this study incorrectly reported *NonAssn* 43 times in the *Baseline Dataset*, yielding an incorrect non-association rate (INR_{PRES}) of 1.8%.

The incorrect non-associations were not limited to just a few participants: 29 of the 84 participants made at least one incorrect non-association; one participant reported four incorrect *NonAssn*s, and 11 participants each reported 2 incorrect *NonAssn*s .

With respect to QKsets, 30 out of 107 mated QKsets resulted in at least one incorrect *NonAssn*: three QKsets each yielded three incorrect *NonAssn*s, and seven QKsets each yielded incorrect *NonAssn*s.

Appendix E5 Repeatability of Errors and Incorrect Conclusions

Table S17 shows the paired trials in the *Repeat Dataset* that included errors or incorrect conclusions. Out of 578 pairs of 1st and 2nd responses, one resulted in repeated errors (one instance in which erroneous *Excls* were repeated, labelled "FN-FN"), and one resulted in an incorrect response that was changed to an error (one instance in which an incorrect *HighAssn* was changed to an erroneous *ID*, labelled "IA-FP"). In all the other instances of incorrect or erroneous conclusions in the *Repeat Dataset*, the incorrect/erroneous conclusions were not repeated. Correct conclusions were changed to incorrect at about the same rate in which incorrect conclusions were corrected.

Mating	*Error type*	*1st Response*	*2nd Response*	*Paired Trials*	*Paired Trials (by error type)*
Mated	FN-FN	Excl	Excl	1	1
	FN (not repeated)	ID	Excl	2	13
		HighAssn	Excl	1	
		Assn	Excl	1	
		Excl	ID	2	
		Excl	HighAssn	1	
		Excl	Assn	4	
		Excl	LimitedAssn	2	
	IN (not repeated)	ID	NonAssn	1	7
		Assn	NonAssn	1	
		LimitedAssn	NonAssn	2	
		NonAssn	ID	1	
		NonAssn	Assn	1	
		NonAssn	LimitedAssn	1	
Nonmated	IA-FP	HighAssn	ID	1	1
	FP (not repeated)	ID	Assn	1	2
		Assn	ID	1	
	IA (not repeated)	HighAssn	Assn	2	8
		HighAssn	LimitedAssn	1	
		HighAssn	NonAssn	1	
		HighAssn	Excl	1	
		Assn	HighAssn	1	
		Inc	HighAssn	1	
		NonAssn	HighAssn	1	

Table S17. Repeatability of errors and incorrect conclusions. (*Repeat Dataset*: 21 mated and 11 nonmated pairs of trials)

Appendix E6 Examiner Comments on Errors and Incorrect Conclusions

Participants were permitted to make comments as part of their comparison responses. A total of 1,274 comments were received (on the 6,610 total trials). Of the 262 trials that resulted in errors or incorrect conclusions, 53 had comments. On review of the these comments, 47 comments indicated a basis for the incorrect conclusions, summarized in Table S18.

	Mating	*Conclusion*	*Review Category*	*# of comments*
Error	Mated	Excl	Erroneous Excl based on mold	1
			Erroneous Excl based on RACs	5
			Erroneous Excl based on size	4
			Erroneous Excl based on wear	5
			Erroneous Excl based on wear/mold	2
			Erroneous Excl based on wear/mold/size	1
			Erroneous Excl based on wear/size	1
Incorrect	Mated	NonAssn	Incorrect NonAssn based on design	1
			Incorrect NonAssn based on minor size differences, wear, RACs	1
			Incorrect NonAssn based on mold	1
			Incorrect NonAssn based on RACs	6
			Incorrect NonAssn based on size	5
			Incorrect NonAssn based on wear	1
	Nonmated	HighAssn	Incorrect HighAssn based on mold	1
			Incorrect HighAssn based on RACs	4
			Incorrect HighAssn based on RACs and wear	2
			Incorrect HighAssn based on Schallamach	1
			Incorrect HighAssn based on texture of outsole	1
			Incorrect HighAssn based on wear	4

Table S18. Summary of comments made by participants on comparisons that resulted in errors or incorrect conclusions. (47 trials)

Appendix F Quality and Difficulty

This appendix provides support for Section 4.4, Effects of Questioned Impression Quality.

Appendix F1 Quality Metric Definition

To aid in data selection for this study, we developed a novel framework to assess and rate the quality of questioned impressions using a footwear-specific rubric. For a detailed description of this process, please refer to [17]; for convenience, the quality rubric and scoring process are summarized here.

Table S19 is the footwear questioned impression quality rubric that was developed to aid in data selection for this study; the rubric includes 10 questioned impression attributes, which were each ranked using a three-level ordinal scale ranging from 0 (poor) to 2 (good). Each questioned impression was generally assessed by two forensic footwear examiners on the study team who assigned a score to each attribute in the rubric. The team members were asked to assess the questioned impression (Q) based on their observations on all available reproductions of each impression, without reference to the known footwear. The raw scores for each attribute were averaged and then summed to obtain a single composite quality score describing the quality and quantity of information available in the questioned impression; quality scores ranged from a minimum of 0 (worst, score of 0 across all attributes) to a maximum of 20 (best, score of 2 across all attributes).

ISO	*Attribute*	*Consideration*	*Assessment Score*		
			0	1	2
Character	Quantity	Estimate the relative amount of the outsole that is reproduced in the impression.	Much less than half	About half	Most or all (heel to toe)
Fidelity	Pattern	Can you discern the geometric shapes that form the pattern in the impression?	No	Somewhat	Yes
	Contrast	Rate the contrast between the impression and the background.	Poor	Moderate	Good
	Distortion	How much distortion is present in the impression?	Significant amount	Some	None
	Substrate	Do features of the substrate (e.g., texture, voids, and background pattern) interfere with visualizing the impression detail?	Yes	Somewhat	No
	Matrix	Does the amount or type of matrix (e.g., too much or too little blood) prevent visualizing the impression detail?	Yes	Somewhat	No
	Overlap	Can you distinguish the primary impression from the overlapping impression(s)?	Hard to distinguish	Easy to distinguish	No overlapping impressions
	Clarity	Is the clarity of the impression sufficient to visualize fine detail (e.g., outsole texturing and potential RACs)?	No	Only in some areas	Yes
Character/Fidelity	Left vs Right	Can you determine if the impression was made by a left or a right shoe?	No	Possibly, but uncertain	Yes
Utility	Suitability	Classify the impression according to expected suitability for comparison.	Unsuitable for comparison	Suitable for class inclusion or exclusion	Suitable for identification

Table S19. Footwear questioned impression quality rubric (reprinted with permission from [17]). "ISO" refers to [39], which defines sample quality in terms of character, fidelity, and utility.

Appendix F2 Quality Distribution

Fig S6 details the distribution of quality scores for each of the 269 distinct QKsets, illustrating the range of questioned impression qualities included this study. To facilitate analyses, these quality scores were then assigned to one of five quality grade quintiles, ranging from F (poorest quality) to A (best quality) (Table S20).

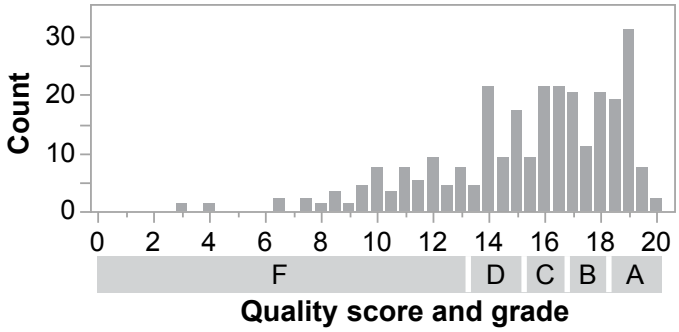

Fig S6. Distribution of quality scores for QKsets. Quality grades A-F are defined as quintiles of this distribution. N=269 QKsets; mean=15.3, median=16. (The equivalent distribution for the 162 distinct Qs is not notably different from the QKset distribution: mean=15.5, median=16).

Quality grade	*# QKsets*	*Min(Quality score)*	*Max(Quality score)*	*# Distinct Qs*
A	59	18.5	20	34
B	51	17	18	29
C	51	15.5	16.5	30
D	51	13.5	15	29
F	57	3	13	40

Table S20. Quality grade quintiles. Note that the A and F bins are slightly larger than the middle bins because the ordinal data could not be split cleanly into equal quintiles without separating quality scores.

Appendix F3 Quality and Neutral Responses

As illustrated in Fig S7, neutral responses (*NotSuitable* and *Inc*) were strongly associated with quality grade. Neutral responses were disproportionately rendered for the poorest quality impressions (quality grade F)— 69.4% (234/337) of all neutral responses were reported for QKsets that included a Q with quality grade F. By comparison, 4.5%-8.9% of neutral responses were reported for quality grades A-C and 19.9% were reported for quality grade D.

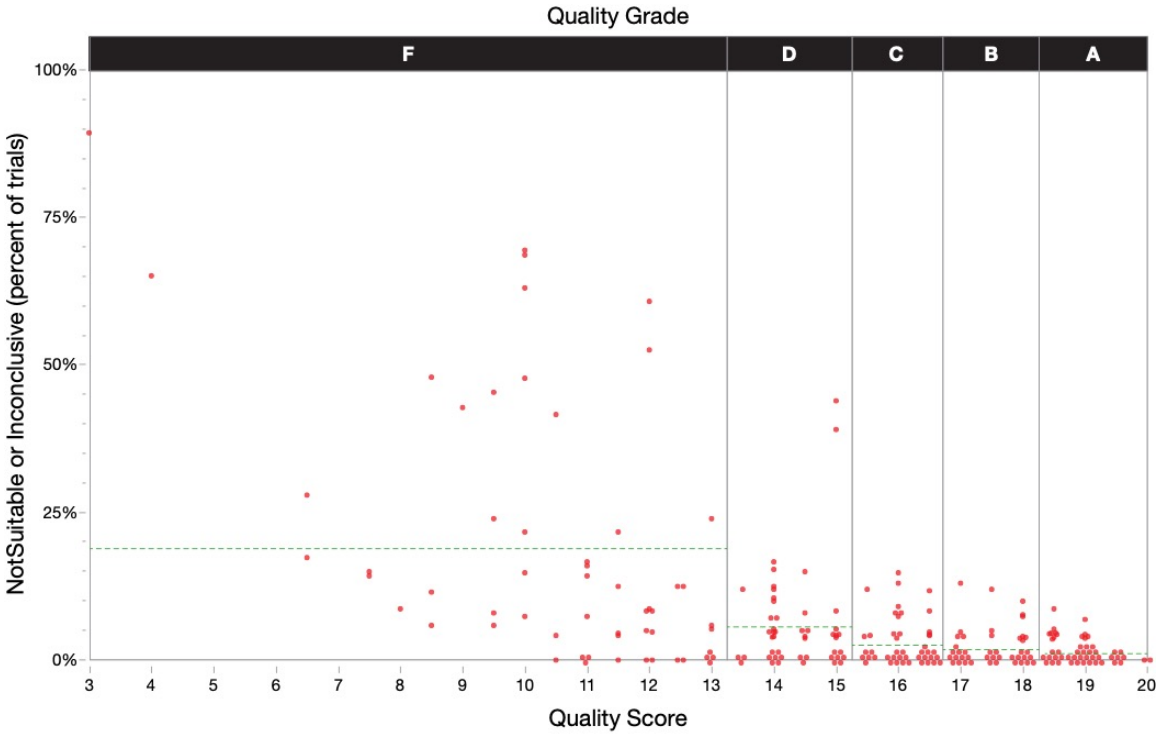

Fig S7. Association between quality and proportion of trials resulting in *NotSuitable* or *Inc* conclusions. Each point represents one QKset. Dashed lines indicate means by quality grade. Points are jittered to minimize superimpositions. (*Baseline Dataset*. n=269 distinct QKsets)

Appendix F4 Quality and Conclusions

Fig S8 shows the association between quality score and conclusion rates in the *Baseline Dataset*, for different categories of class similarity within nonmates. This provides a detailed view of the summary shown in Figure 9 (main paper).

For mated trials, the proportion of *ID*s (true positives) is strongly associated with the quality score: as quality increases, the proportion of true positives likewise increases, and the proportion of class associations decreases. Nonmated trials show a similar (but weaker) effect, in which higher quality is generally associated with higher rates of *Excls* (true negatives). However, for nonmated QKsets in which the Q and K are from different make, model, or foot, some QKsets are unanimously excluded even for quality scores as low as 13 (quality grade F).

As discussed in the previous section, neutral responses are disproportionately associated with poor quality for both mates and nonmates: here we see that even poor quality questioned impressions were often assessed as suitable for comparison.

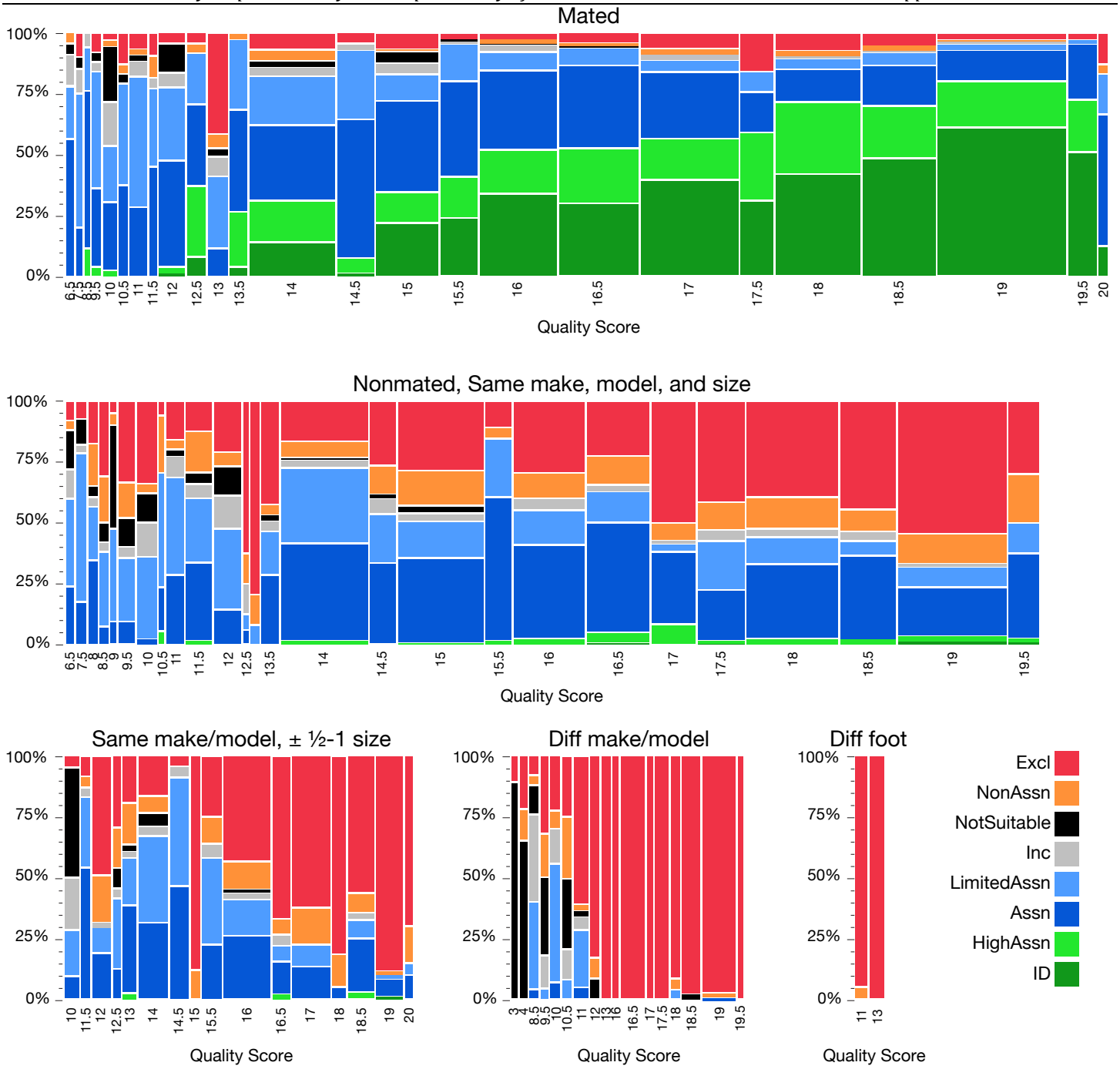

Fig S8. Conclusion rates by quality, for different categories of class similarity within nonmates. (*Baseline Dataset*)

Appendix F5 Difficulty and Conclusions

Fig S9 shows the association between difficulty ratings and conclusion rates in the *Baseline Dataset*. For each assigned QKset, participants were asked to rate their "perceived level of difficulty associated with the comparison of the questioned impression to the known item of footwear." Conclusions were strongly associated with difficulty: for both mates and nonmates, greater difficulty was associated with a smaller proportion of correct definitive conclusions, and more probable, association, or neutral responses. Note that erroneous *Excls* and *IDs* do not show obvious trends with respect to difficulty: the proportions of erroneous *Excls* are similar for all levels of difficulty, and erroneous *IDs* were assessed as Easy, Moderate, and Difficult (not Very Difficult).

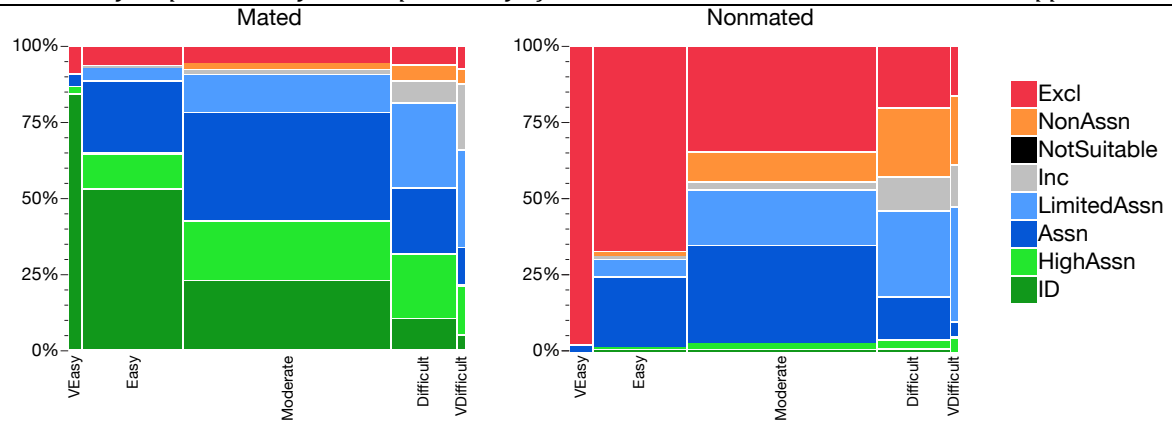

Fig S9. Conclusions by difficulty. (*Baseline Dataset*)

Fig S10 shows the association between difficulty ratings (individually and averaged by QKset) and quality grade: easier comparisons are associated with higher quality impressions. Note that quality and difficulty are not necessarily expected to track perfectly. For example, a poor quality questioned impression could result in a very easy class exclusion if compared to a known with obvious design differences; alternatively, a high quality questioned impression may result in a difficult comparison if compared to a known requiring detailed assessments of RACs. Furthermore, the footwear community does not have a standardized definition of difficulty (nor did this study). As such, the difficulty determination was necessarily subjective and may be impacted by a variety of factors including time spent on the comparison, difficulty in selecting a conclusion category, or effort required to observe and assess discriminating features.

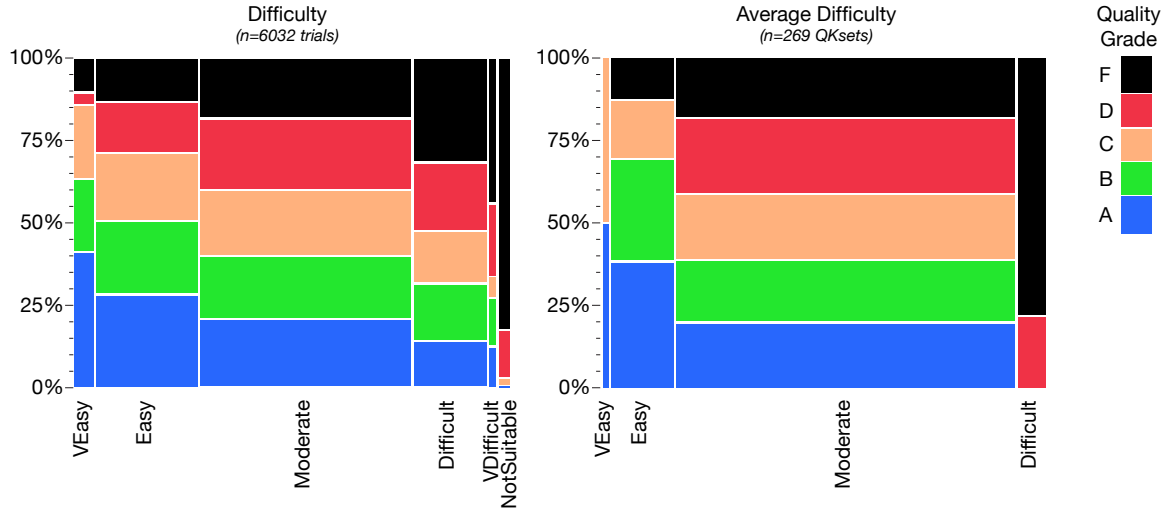

Fig S10. Association between difficulty and quality grades for individual examiner assessments (left) and average QKset difficulty (right). Average QKset difficulty was determined by assigning a numerical value to each difficulty ranking, computing a weighted difficulty sum for each QKset, and dividing by the total number of trials completed for that QKset.

Appendix G Consensus and "Appropriate" Conclusions

This appendix provides supporting material for Section 4.3, QKset-Specific Effects and Consensus.

Table S21 summarizes the counts of QKsets by different methods of assessing consensus. The only QKset with an erroneous majority (or supermajority) consensus was QK213 (77% FNR; Figure 8 (main paper) and *Appendix E3*, Fig S5). Three QKsets had a debatable majority consensus: *Assn* on nonmated QKsets where the Q and K were ½ size different; one of which had a 75% supermajority consensus.

				Nonmated					
		Total	*Mated*	*All*	*Same Make, Model, Size*	*SizeDiff ½*	*SizeDiff 1*	*Diff MakeModel*	*Diff Foot*
Majority	ID	29	29	-	-	-	-	-	-
	HighAssn	1	1	-	-	-	-	-	-
	Assn	44	22	22	19	3	-	-	-
	LimitedAssn	5	2	3	3	-	-	-	-
	NotSuitable	2	-	2	-	-	-	2	-
	Excl	58	1	57	24	9	5	17	2
	(No majority)	130	52	78	54	9	10	5	-
75% supermajority	ID	13	13	-	-	-	-	-	-
	Assn	9	5	4	3	1	-	-	-
	NotSuitable	1	-	1	-	-	-	1	-
	Excl	37	1	36	9	3	5	17	2
	(No supermajority)	209	88	121	88	17	10	6	-
Plurality	ID	38	38	-	-	-	-	-	-
	HighAssn	13	13	-	-	-	-	-	-
	Assn	86	38	48	38	7	3	-	-
	LimitedAssn	27	9	18	12	2	2	2	-
	NotSuitable	9	2	7	2	-	2	3	-
	NonAssn	1	-	1	1	-	-	-	-
	Excl	80	3	77	40	12	6	17	2
	(No single plurality category)	15	4	11	7	-	2	2	-
Median	ID	29	29	-	-	-	-	-	-
	HighAssn	18	18	-	-	-	-	-	-
	Assn	66	41	25	22	3	-	-	-
	LimitedAssn	52	11	41	28	6	5	2	-
	Inc	18	3	15	8	-	4	3	-
	NonAssn	18	-	18	14	3	1	-	-
	Excl	58	1	57	24	9	5	17	2
	(No single median category)	10	4	6	4	-	-	2	-
Average	ID	16	16	-	-	-	-	-	-
	HighAssn	27	27	-	-	-	-	-	-
	Assn	42	35	7	6	1	-	-	-
	LimitedAssn	58	25	33	26	3	4	-	-
	Inc	60	3	57	40	7	6	4	-
	NonAssn	33	1	32	22	7	-	3	-
	Excl	33	-	33	6	3	5	17	2
	Total QKsets	269	107	162	100	21	15	24	2

Table S21. Consensus conclusion counts, using various consensus methods. Erroneous consensus conclusions are shown in yellow; debatable consensus conclusions are shown in gray. (*Baseline dataset*)

Details regarding the 3 mated QKsets that had plurality conclusions of *Excl* (FNs):

- Full blood impression on clothing/material, poor quality (quality grade: F), and majority difficulty rating of easy
- Partial residue impression on plastic, relatively high quality (quality grade: B), and majority difficulty rating of moderate
- Full impression of new known footwear, relatively high quality (quality grade: B), and majority difficulty rating of moderate

Details regarding average conclusions in the *Baseline Dataset*:

- In contrast to majority and plurality which consider the density of responses per category, determining consensus via the average conclusion accounts for the entire distribution of responses. The average conclusion is determined by assigning a numerical score to each conclusion category (*Excl* = 0, *NonAssn* = 1, *NotSuitable* = 2, *Inc* = 2, *LimitedAssn* = 3, *Assn* = 4, *HighAssn* = 5, *ID* = 6); this score is used to compute the weighted sum of responses for a given QKset which is then divided by the total number of trials for that QKset in order to obtain an average conclusion category. (Note that the average category is rounded to the nearest conclusion)
- The consensus conclusion determined via average was never incorrect and and rarely erroneous with respect to ground truth: the same mated QKset (QK213, shown in Figure 8 (main paper) and *Appendix E3*, Fig S5) that produced a majority conclusion of *Excl*, likewise yielded an average conclusion of *Excl*.
- The average conclusion agreed exactly with majority conclusions for every QKset that exhibited a majority conclusion.
- The average and plurality conclusions were the same for 76% of the QKsets, differed by a single category for 20% of QKsets, and differed by more than one conclusion category for 4% of QKsets.

Details regarding interquartile range of conclusions in the *Baseline Dataset*:

- Richetelli et al. [11] computed IQR for each of their 12 comparisons and reported the proportion of responses that fell within the IQR as a measure of examiner agreement with this consensus range; for comparable results for the current study, see *Appendix M*.

Appendix H Reproducibility and Repeatability

This appendix provides support for Sections 4.5 (Reproducibility) and 4.6 (Repeatability).

Appendix H1 Reproducibility and Repeatability of suitability assessments

Prior to conducting comparisons between questioned impressions and known footwear in casework, FFEs evaluate the suitability of the questioned impression to determine whether there is sufficient detail to make a meaningful comparison. As such, participants in this study were asked to assess whether the Q in each assigned QKset was suitable for comparison; Table S22 reports the reproducibility of these assessments. Overall, 2.8% of trials in the *Baseline Dataset* yielded a response of *NotSuitable*. Participants typically disagreed with each other on determinations of *NotSuitable*: Just under one-third of *NotSuitable* responses were reproduced by a second examiner (30.9%). If we consider that participants may have conflated *NotSuitable* and *Inc* (since *Inc* is not in the SWGTREAD 2013 conclusion scale), participants typically disagreed with each other on the combined determination of *NotSuitable* and *Inc* (28.3% of responses were reproduced). These results indicate that suitability assessments are notably inconsistent and may benefit from standardization efforts to more clearly define what consititutes a suitable versus not suitable questioned impression.

Note that although most participants used *NotSuitable* and *Inc* responses, not all did: 33 participants never reported *NotSuitable*, 15 of whom never reported *NotSuitable* or *Inc*. Table S22 shows that limiting results to participants who used these conclusions increases reproducibility, but the reproducibility of *NotSuitable* responses is still less than 50% (42.4%). This behavior does not explain high rates of erroneous *IDs* or *Excls*: the participants with the highest error rates all used *NotSuitable* and/or *Inc*.

With respect to repeatability, there were only two determinations of *NotSuitable* in the *Repeat Dataset*; neither determination was repeated.

All participants		*Count*	*Examiner B* *NotSuitable*	*Suitable*
Examiner A	*NotSuitable*	170	30.9%	69.2%
	Suitable	5862	2.0%	98.1%

All participants		*Count*	*Examiner B* *NotSuitable or Inc*	*Other*
Examiner A	*NotSuitable or Inc*	367	28.3%	71.7%
	Other	5665	4.6%	95.4%

51 participants who ever report NotSuitable		*Count*	*Examiner B* *NotSuitable*	*Suitable*
Examiner A	*NotSuitable*	170	42.4%	57.6%
	Suitable	3926	2.6%	97.4%

69 participants who ever report NotSuitable or Inc		*Count*	*Examiner B* *NotSuitable or Inc*	*Other*
Examiner A	*NotSuitable or Inc*	367	31.5%	68.5%
	Other	4900	5.3%	94.8%

Table S22. Reproducibility of suitability assessments and inconclusive responses. Left tables differentiate *Suitable* vs *NotSuitable*; right tables treat *NotSuitable* and *Inc* as synonymous. Upper tables include all participants; lower left table omits participants who never responded *NotSuitable*; lower right table omits participants who never responded (*NotSuitable* or *Inc*). (n = 132,074 pairs of responses from different participants on the same QKsets). (*Baseline Dataset*)

Appendix H2 Reproducibility and Repeatability of conclusions

Table S23 and Table S24 provide the contingency tables of the reproducibility and repeatability results illustrated in Figure 10 and Figure 11 (main paper).

Reproducibility			Trials	Inter-examiner decision pairs	Examiner B (inter-examiner decision pairs)							
					ID	HighAssn	Assn	LimitedAssn	Inc	NotSuitable	NonAssn	Excl
Examiner A	Mated	ID	725	15,992	62.2%	18.8%	10.1%	3.4%	1.1%	0.5%	0.5%	3.2%
		HighAssn	410	9,145	32.9%	28.4%	22.9%	7.4%	1.7%	0.4%	1.4%	4.9%
		Assn	682	14,640	11.1%	14.3%	45.6%	17.4%	2.4%	1.1%	2.3%	5.8%
		LimitedAssn	312	6,909	8.0%	9.7%	36.9%	26.0%	5.0%	3.6%	2.8%	8.0%
		Inc	60	1,350	13.5%	11.6%	25.6%	25.3%	6.2%	7.7%	3.4%	6.7%
		NotSuitable	39	808	9.0%	4.2%	19.3%	31.2%	12.9%	18.3%	2.1%	3.0%
		NonAssn	43	963	8.3%	13.1%	35.2%	19.7%	4.8%	1.8%	3.3%	13.8%
		Excl	146	3,277	15.8%	13.7%	26.1%	16.9%	2.7%	0.7%	4.1%	19.9%
	Nonmated	ID	9	186	1.1%	0.5%	40.3%	11.3%	0.0%	0.0%	11.3%	35.5%
		HighAssn	50	1,088	0.1%	7.9%	46.0%	11.3%	2.1%	0.3%	9.1%	23.3%
		Assn	852	19,071	0.4%	2.6%	41.6%	19.1%	3.1%	1.1%	9.7%	22.5%
		LimitedAssn	575	12,904	0.2%	1.0%	28.2%	27.8%	6.2%	5.0%	9.6%	22.2%
		Inc	137	3,051	0.0%	0.8%	19.2%	26.2%	8.7%	9.7%	9.8%	25.6%
		NotSuitable	131	2,810	0.0%	0.1%	7.6%	22.7%	10.6%	34.4%	8.3%	16.2%
		NonAssn	346	7,531	0.3%	1.3%	24.6%	16.5%	4.0%	3.1%	12.6%	37.7%
		Excl	1,515	32,349	0.2%	0.8%	13.3%	8.9%	2.4%	1.4%	8.8%	64.3%

Table S23. Contingency table for reproducibility of examiner decisions, shown in Figure 10. (*Reproducibility Dataset*: 53,084 inter-examiner decision pairs derived from 2,417 decisions on 107 mated QKsets; 78,990 inter-examiner decision pairs derived from 3,615 decisions on 162 nonmated QKsets)

Repeatability			Trials	2nd response (% of trials)							
				ID	HighAssn	Assn	LimitedAssn	Inc	NotSuitable	NonAssn	Excl
1st response	Mated	ID	110	81%	10%	3%	3%	1%	0%	1%	2%
		HighAssn	41	17%	56%	22%	2%	0%	0%	0%	2%
		Assn	48	8%	6%	73%	6%	2%	0%	2%	2%
		LimitedAssn	13	15%	8%	15%	46%	0%	0%	15%	0%
		Inc	2	50%	50%	0%	0%	0%	0%	0%	0%
		NotSuitable	1	0%	100%	0%	0%	0%	0%	0%	0%
		NonAssn	3	33%	0%	33%	33%	0%	0%	0%	0%
		Excl	10	20%	10%	40%	20%	0%	0%	0%	10%
	Nonmated	ID	1	0%	0%	100%	0%	0%	0%	0%	0%
		HighAssn	6	17%	0%	33%	17%	0%	0%	17%	17%
		Assn	135	1%	1%	63%	15%	3%	1%	6%	11%
		LimitedAssn	47	0%	0%	49%	34%	0%	0%	13%	4%
		Inc	10	0%	10%	30%	30%	10%	0%	10%	10%
		NotSuitable	0								
		NonAssn	44	0%	2%	23%	18%	0%	0%	36%	20%
		Excl	107	0%	0%	8%	6%	2%	0%	14%	70%

Table S24. Contingency table for repeatability of examiner decisions, shown in Figure 11A&B. Percentages based on fewer than ten trials are grayed. (*Repeat Dataset*: 228 test-retest decision pairs on 12 mated QKsets; 350 test-retest decision pairs on 18 nonmated QKsets)

Fig S11 describes the reproducibility and repeatability of conclusions as a function of the absolute difference (Delta) in conclusion categories. To report this, we assigned each conclusion category a numerical score (Excl=1, NonAssn=2, NotSuitable=3, Inc=3, LimitedAssn=4, Assn=5, HighAssn=6, ID=7) and computed the absolute value of the difference between conclusions reported for each decision pairing (either between different participants or between repeated examinations by the same participant). Note this analysis does not distinguish between *NotSuitable* and *Inc*.

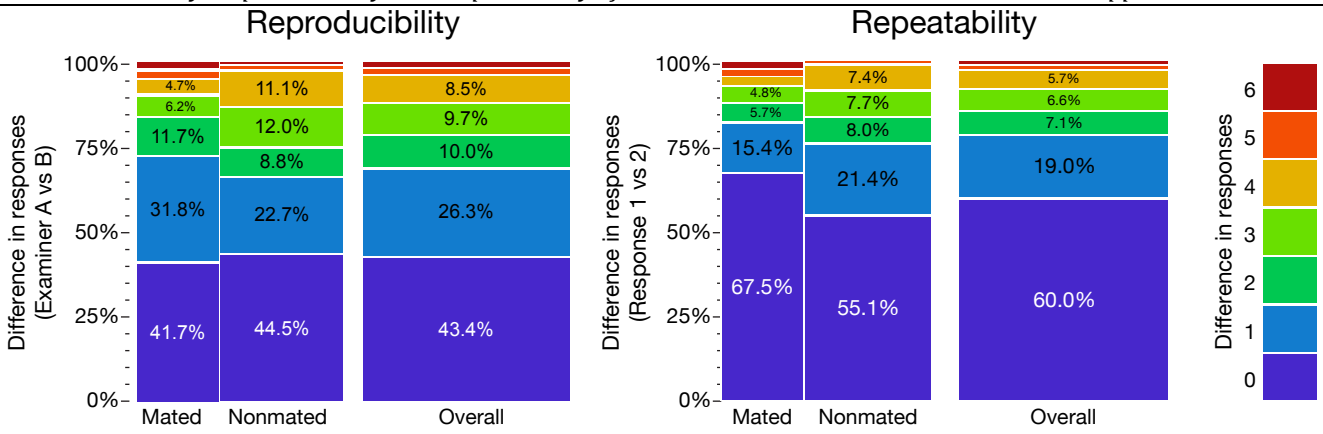

Fig S11. Reproducibility (left) and repeatability (right) of conclusions. Same data as Figure 10, summarized by a difference in conclusions: 0 indicates the same conclusion; 6 indicates diametrically opposed conclusions (*ID* vs *Excl*). (Reproducibility data: 53,084 inter-examiner decision pairs derived from 2,417 decisions on 107 mated QKsets; 78,990 inter-examiner decision pairs derived from 3,615 decisions on 162 nonmated QKsets. Repeatability data: 228 test-retest decision pairs on 12 mated QKsets; 350 test-retest decision pairs on 18 nonmated QKsets.)

Appendix H3 Reproducibility and Repeatability by Difficulty and Quality

The repeatability and reproducibility of conclusions tend to decrease as difficulty increases (Fig S12).

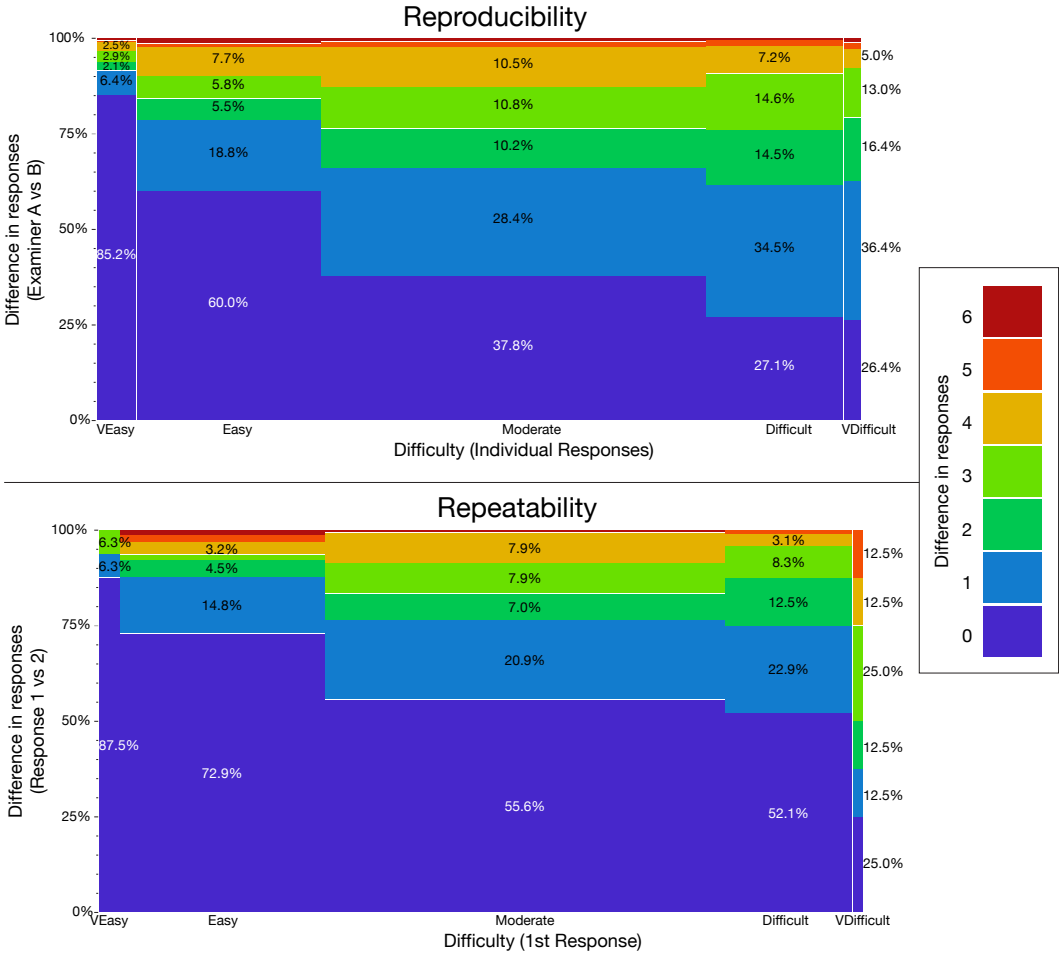

Fig S12. Reproducibility and repeatability: differences in conclusions by difficulty. (*Reproducibility Dataset* (top), and *Repeat Dataset* (bottom) — both omit initial responses of *NotSuitable,* which are not associated with difficulty.)

The reproducibility and repeatability of correct definitive conclusions (true positives and true negatives) are inversely associated with participants' assessments of comparison difficulty (Fig S13 and Fig S14). In general, as difficulty increases, the proportion of reproduced/repeated true positives and true negatives decreases and the proportion of class associations reported by a second examiner (or by the same examiner on a second comparison) increases. The exception is for repeatability of true negatives, which do not follow this trend; possible explanations for this may be the relatively small number of trials, and the fact that the nonmated QKsets included in the *Repeat Dataset* were the same make/model/size, which were associated with low true negative rates.

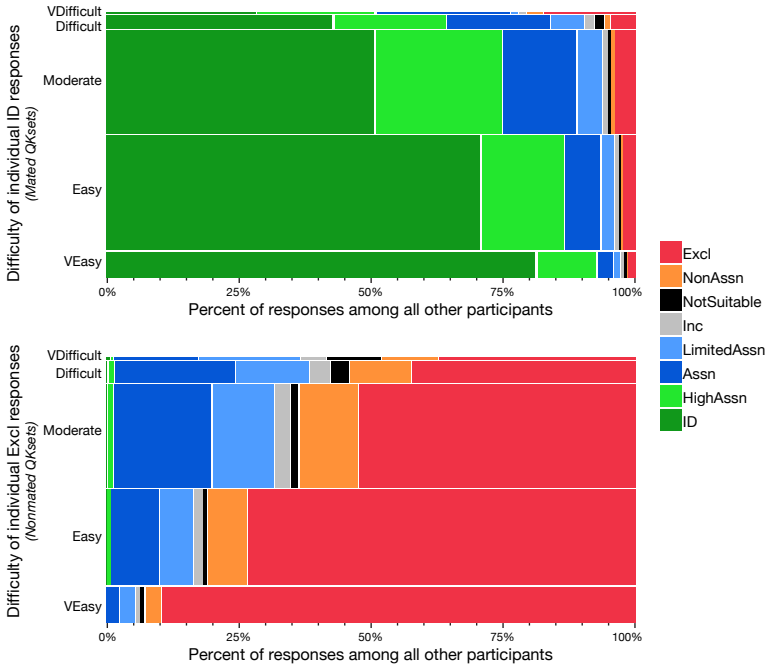

Fig S13. Reproducibility by difficulty, limited to (top) *IDs* on mated QKsets and (bottom) *Excls* on nonmated QKsets. (Top: 15,992 inter-examiner decision pairs derived from 725 *ID* responses on mated QKsets. Bottom: 32,349 inter-examiner decision pairs derived from 1515 *Excl* responses on nonmated QKsets.)

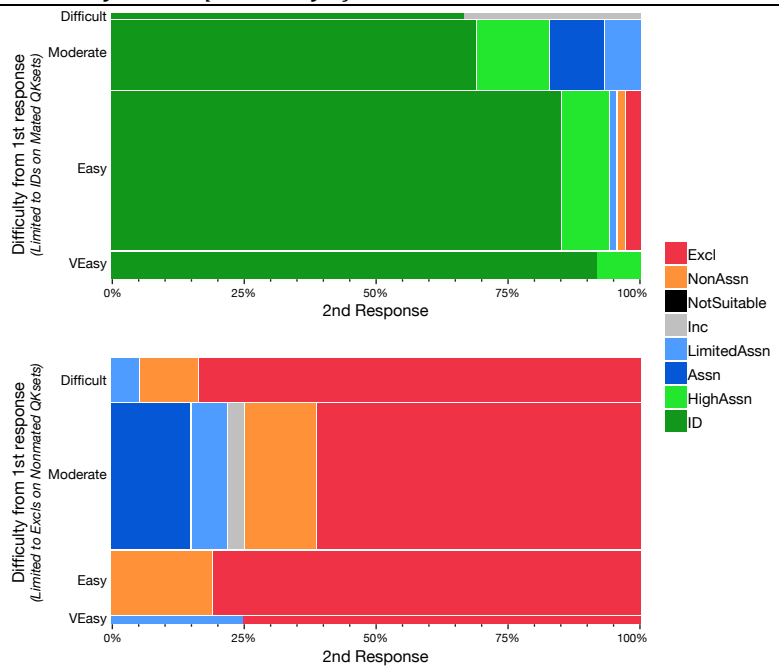

Fig S14. Repeatability by difficulty, limited to (top) *IDs* on mated QKsets and (bottom) *Excls* on nonmated QKsets. None were assessed as "Very Difficult." (Subset of *Reproducibility Dataset*. Top: 110 pairs of 1st and 2nd responses on mated QKsets in which the 1st response was *ID*. Bottom: 107 pairs of 1st and 2nd responses on nonmated QKsets in which the 1st response was *Excl*.)

Note that while repeatability and reproducibility show a clear association with participants' assessments of difficulty, the same is not true with the quality of the questioned impression. Fig S15 shows differences in conclusions are not notably associated with the quality of the questioned impressions, for both reproducibility and repeatability. Note that on nonmated trials repeatability increases for the poorest quality images (Fig S15, right): this is presumably due to high rates of class associations: for quality grade F 15 of the 18 repeat trials resulted in repeated class association decisions (*Assn* or *LimitedAssn*).

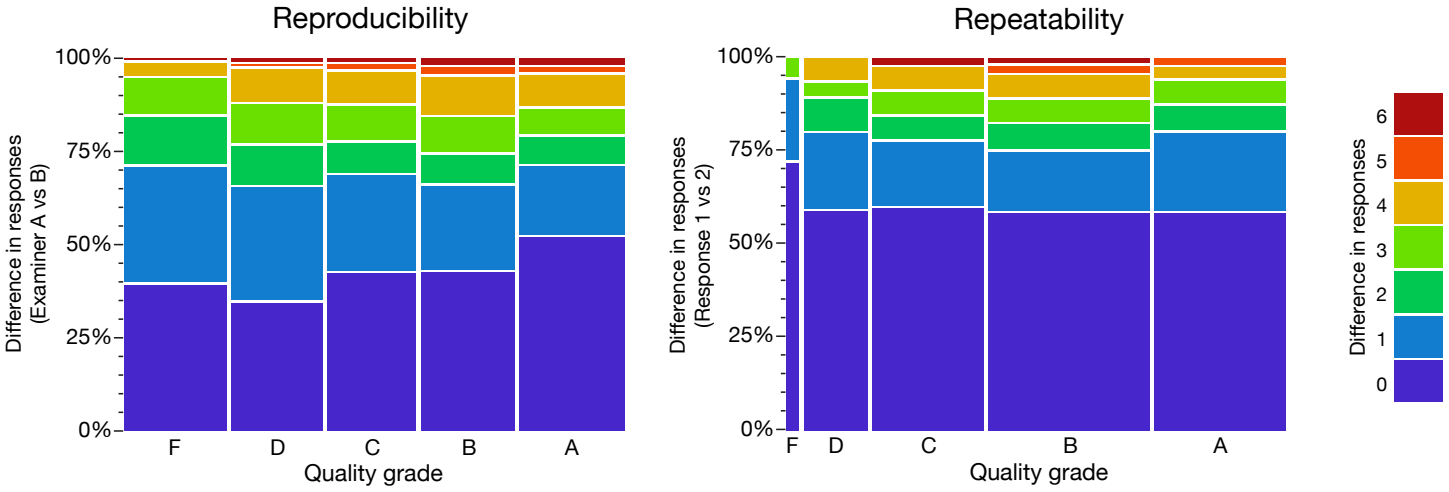

Fig S15. Reproducibility and repeatability: differences in conclusions by quality grade. (Left: *Reproducibility Dataset*; Right: *Repeat Dataset*.)

Appendix I Comparing Examiners

This appendix provides support for Section 4.1, Conclusion Rates, Accuracy, and Errors.

Note that when comparing examiner performance at an individual level, we limit analyses to the 71 examiners who completed at least 40 comparisons (Examiner Comparison Dataset).

Figure 5 and Figure 6 in the main paper compare participants graphically by depicting the rates of erroneous/inaccurate conclusions and true/correct conclusions. However, in order to facilitate an analytic comparision of individual performances, it is necessary to consider three inter-related decision factors: correctness, definitiveness, and the relative value/cost of making definitive (*ID* and *Excl*) versus probable (*HighAssn* and *NonAssn*) conclusions:

- Correctness: considers the participant's conclusions with respect to the ground truth regarding the source of the questioned impression for the given QKset (e.g., decisions of *HighAssn* or *ID* would be considered correct for mated QKsets).
- Definitiveness: considers the participant's reporting tendencies across the conclusion scale, not just on the extremes (e.g., a conservative participant may rarely report definitive conclusions and prefer to report probable conclusions, but this tendency should be accounted for in the assessment of performance).
- Relative value/cost: considers the added value of a correct definitive over probable conclusion or the added cost of an incorrect definitive over probable conclusion (e.g., a correct *ID* conclusion has potentially more value than a correct *HighAssn*, and an erroneous *ID* conclusion has potentially more severe consequences than an incorrect *HighAssn*).

For the purposes of these analyses, probable conclusions are weighted as half of definitive conclusions in both relative value and cost: there are no recommendations in the literature that inform these weights, so these values were selected as an approximation.

To account for these factors, we developed four weighted rates:

- Weighted TP-CA rate = $TPR_{PRES} + 0.5(CAR_{PRES})$
- Weighted TN-CN rate = $TNR_{PRES} + 0.5(CNR_{PRES})$
- Weighted FP-IA rate = $FPR_{PRES} + 0.5(IAR_{PRES})$
- Weighted FN-IN rate = $FNR_{PRES} + 0.5(INR_{PRES})$

Because these rates have notably different ranges, we converted each of these into ***ratios*** describing each participant's performance relative to the other participants in the study (displayed in Fig S16). Each ratio contrasts a participant's individual weighted reporting rate versus the average of that weighted reporting rate across all other participants: for example, a weighted TP-CA ratio of 2.0 means that participant's weighted TP-CA rate was twice the average. (We considered z-normalization of these values, but that was not well-suited to distributions with ceiling effects.)

Fig S16 uses these weighted ratios to depict the same information as Figure 5 and Figure 6 (in the main paper), but collapses the eight dimensions from those charts into four, providing a different method of comparing examiner performance. The symbols and colors correspond between Fig S16, Figure 5, and Figure 6.

As an example of these computations, consider the orange diamond located in the bottom right corner of the top two panels in Fig S16, which represents the participant that committed five false positive errors in the *Baseline Dataset*. This participant had a FPR of 9% and an IAR of 19%; therefore, this participant's weighted FN+IA ***rate*** is 18.5% (9%+0.5(19%)). The average weighted FN+IA rate across all other participants was 1.26%. Therefore, this participant's weighted FP-IA ***ratio*** was 14.7 (18.5%/1.26%), as shown along the y-axis in Fig S16 (top). This value indicates that this participant made nearly 15x the weighted number of *ID* and *HighAssn* conclusions that an average examiner made on assigned nonmated QKsets.

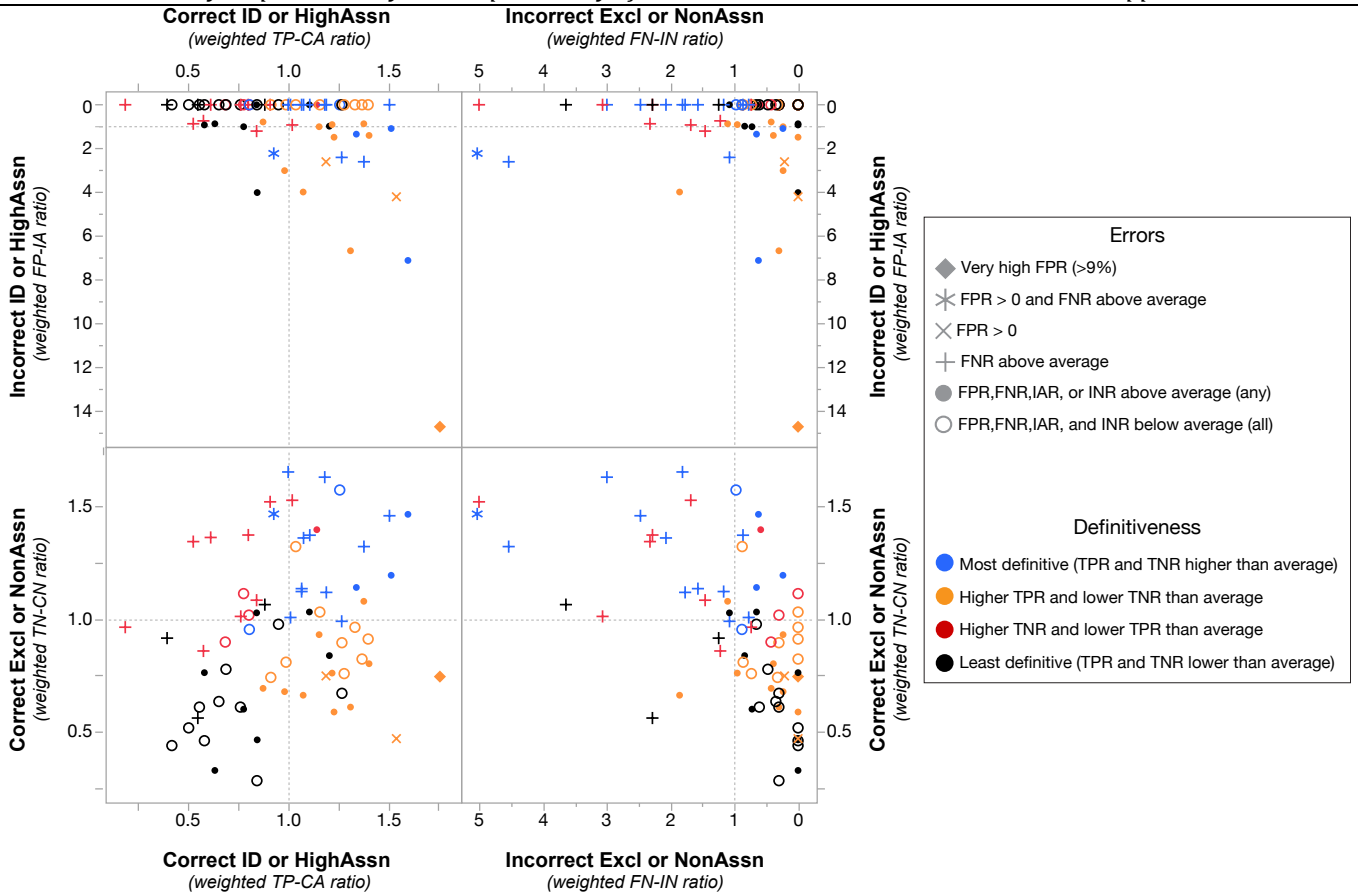

Fig S16. Comparison of participants by weighted performance ratios (n=71; *Examiner Comparison Dataset*).

An alternative method of comparing participants is shown in Fig S17. For each participant, if you compare their responses on all nonmated trials with their responses on all mated trials, this shows how often the participant gave a "higher" response (response closer to *ID* than *Excl*) to the mated trial. In this context, we used a scale with *ID*=2, *HighAssn*=1,*NonAssn*=-1, *Excl*=-2, and all other responses=0. Note that this does not differentiate between *Inc*, *NotSuitable*, *Assn*, and *LimitedAssn*, because those assess class rather than source.

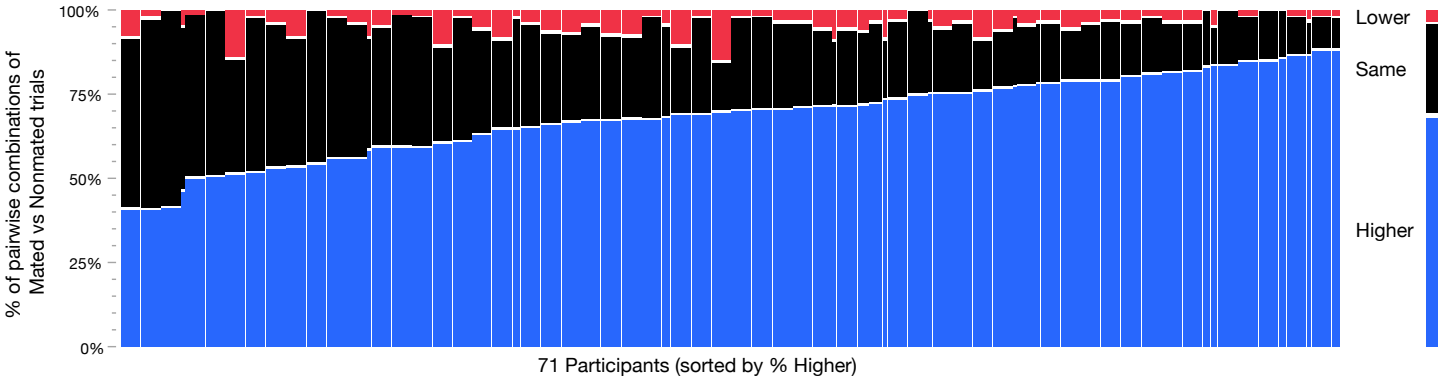

Fig S17. Comparing participants in terms of how often they gave a "higher" response (response closer to *ID* than *Excl*) to mated comparison sets than to nonmated comparison sets. (n=116,443 pairwise combinations of 3,449 nonmated vs 2,300 mated trials for the 71 participants in the *Examiner Comparison Dataset*).

Fig S18 compares participants using three distributions derived from Fig S17. These distributions differ only in how the "same" responses are handled: in Fig S18A the "same" responses are essentially treated as incorrect (mean 70, median 71); in Fig S18B the "same" responses are treated as correct (mean 96, median 97); in Fig S18C the "same" responses are omitted (mean 94, median 95). These provide three approaches for calculating the "empirical AUC" [40–43](Area under the receiver operating

characteristic curve), which is a standard approach for evaluating the effectiveness of classifiers, and provides an alternative method of assessing individual examiner effectiveness.

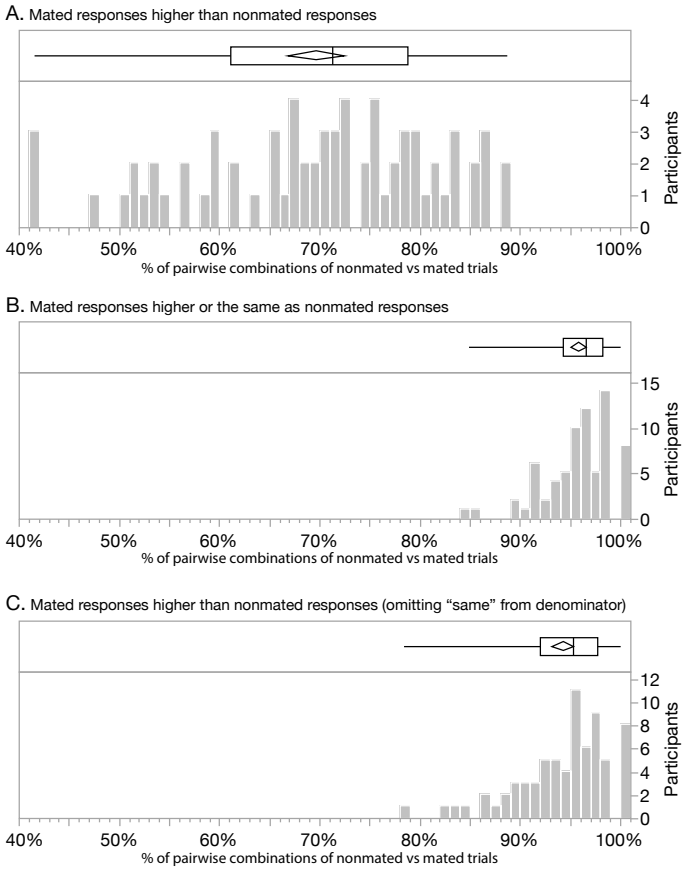

Fig S18. Distributions for the 71 participants in the *Examiner Comparison Dataset* derived from the data shown in Fig S17. Distribution of participants based on the proportion of pairwise combinations of nonmated vs mated trials in which (A) the mated responses are higher than the nonmated responses (height of blue columns in Fig S17); (B) the mated responses are higher or the same as the nonmated responses (combined height of blue and black columns in Fig S17); (C) the mated responses are higher than the nonmated responses, but omitting the "same" responses from the denominator (height of blue columns divided by the combined height of the blue and red columns in Fig S17).

Appendix I1 Definitiveness and Effectiveness

Fig S18 details participant reporting tendencies with respect to definitiveness. The left panel contrasts examiners' definitive conclusion rates versus their rates of reporting other non-neutral conclusions (*HighAssn*, *Assn*, *LimitedAssn*, and *NonAssn*). Participants who fall on the top-left to bottom-right bold diagonal never report neutral decisions (*NotSuitable* or *Inc*). The bottom-left to top-right dotted diagonal represents an even split between definitive conclusions and other conclusions: participants above this diagonal report definitive conclusions more often than other non-neutral conclusions (and vice versa for those who fall below this line). The relationship between definitiveness and accuracy becomes apparent when considering the open circles on this plot—these participants largely fall below the bottom-left to top-right dotted diagonal, indicating that they report fewer definitive conclusions than other conclusions, which manifests as fewer erroneous and incorrect conclusions than average.

The right panel contrasts examiners' definitive conclusion rates versus their probable conclusion rates. Participants who fall below the dashed diagonal report a majority of their conclusions as class associations and/or neutral responses (fewer than 50% of responses are definitive or probable). Not surprisingly, the majority of participants that fall below this diagonal also exhibit lower than average error rates (open circles). The bottom-left to top-right dotted diagonal represents an even split between definitive and probable conclusions. Notably, just three participants report more probable conclusions than definitive conclusions.

For example, in the left panel we see that the orange diamond (the participant that committed five false positive errors in the *Baseline Dataset*) was associated with almost no neutral responses (i.e., is close to the bold diagonal), and very close to a 50%-

50% split of definitive vs other non-neutral conclusions. In the right panel we see that the blue asterisk was associated with one of the highest rates of definitive conclusions and a very low rate of probable conclusions.

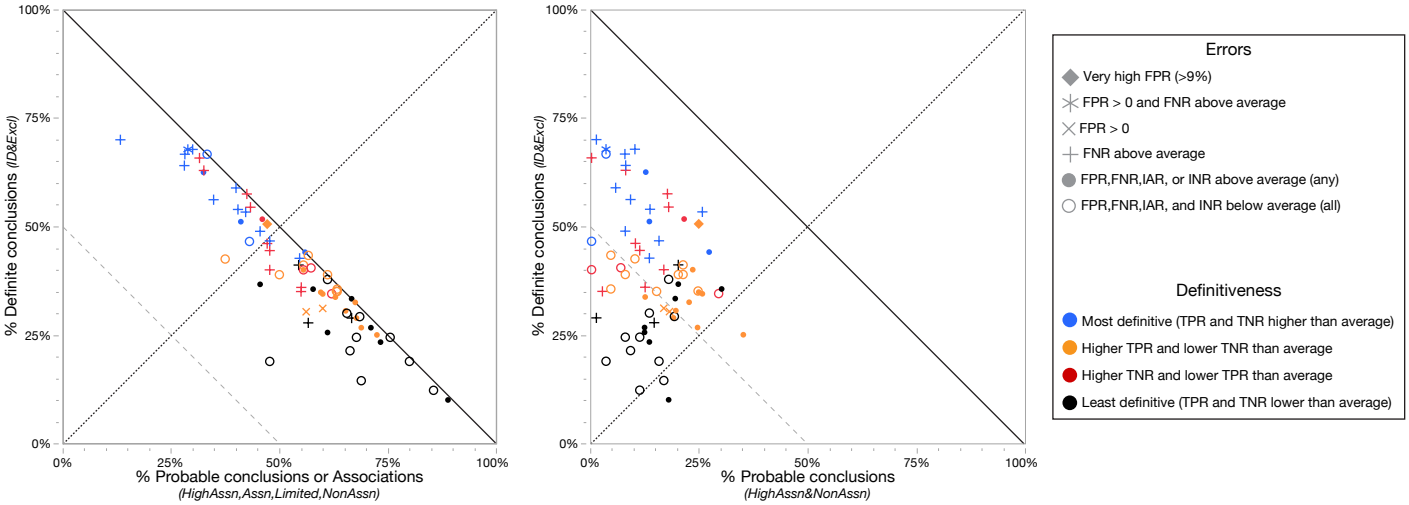

Fig S19. Definitiveness among examiners: definite conclusions vs. other conclusions (left) and definite conclusions vs. probable conclusions (right). The symbols and colors are the same as in Fig S16, Figure 5, and Figure 6. Left chart: bold diagonal indicates participants who made no responses of *NotSuitable* or *Inc*; distance from bold diagonal to (0,0) indicates proportion of responses that were *NotSuitable* or *Inc*; participants who made more definitive conclusions than probable conclusions and associations fall above/left of the dotted line. Right chart: distance from bold diagonal to (0,0) indicates proportion of responses that were (*Assn, LimitedAssn,NotSuitable,* or *Inc*); participants who made more (*Assn, LimitedAssn,NotSuitable,* or *Inc*) responses than definitive or probable conclusions fall below/left of the dashed line; participants who made more definitive conclusions than probable conclusions fall above/left of the dotted line. (*Examiner Comparison Dataset*)

Appendix I2 Examiner Effects vs. Sample Effects

This section is included here for completeness, to show the results of a type of analysis developed for [25]*; see that publication for more details regarding this approach. These analyses show visualizations of effects that are complementary to those shown in Section 4.2 of the main paper.*

The likelihood of a given conclusion can be modelled as a function of both examiner effects and sample effects: Fig S19 and Fig S20 show how individual examiners' conclusions are related to the collective assessments made by all examiners on the same image pairs. Each column depicts the comparisons made on one QKset, with the position on the x axis based on the conclusion rates for that QKset (across all examiners who were assigned that QKset). Each row depicts the comparisons made by an examiner who participated in the study, with the position on the y axis based on that examiner's conclusion rate for those comparisons (across all QKsets assigned to that examiner). For example, the column of green dots at x=100% in Fig S19 show the responses for a mated QKset for which every response was ID (or HighAssn); the row of dots at y=75% show the responses for one participant, for whom 75% of responses on mated QKsets were ID or HighAssn.

The diagonal lines represent probabilities of decisions as predicted by logistic regression based on the QKset (QK) and examiner ID rates, performed on a leave-one-out basis: the outcome for each trial was omitted when calculating the two rates for that trial. These results show wide variation in participants' conclusion rates (y axis), and suggest the presence of implicit individual decision thresholds: the x axis can be seen as a collective assessment among multiple examiners as to whether a given QKset includes sufficient information to make a given decision, and the diagonals show the extent to which how the individual participants decisions thresholds agree with the consensus.

Technical note: in [25]*, created to assess a latent print dataset, data was limited to non-unanimous conclusions — here, because the assignments were balanced by difficulty among examiners and because almost no QKsets were unanimous, all data is used, and the effect of limiting to non-unanimous conclusions would be minimal.*

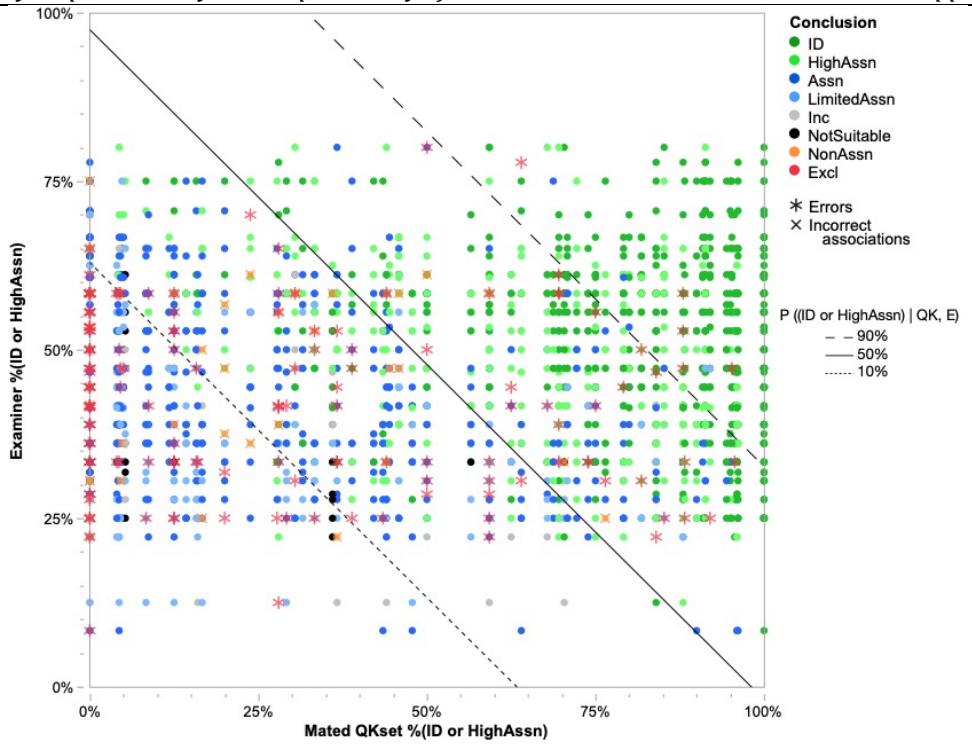

Fig S20. Sample effects vs examiner effects: 107 mated QKsets, plotted by (x axis) the percentage of examiners who made ID or HighAssn conclusions on that QKset, and (y axis) the percentage ID or HighAssn conclusions made by that examiner; the diagonal lines represent {90%, 50%, 10%} probabilities of ID or HighAssn decisions as predicted by logistic regression. (*Baseline Dataset*)

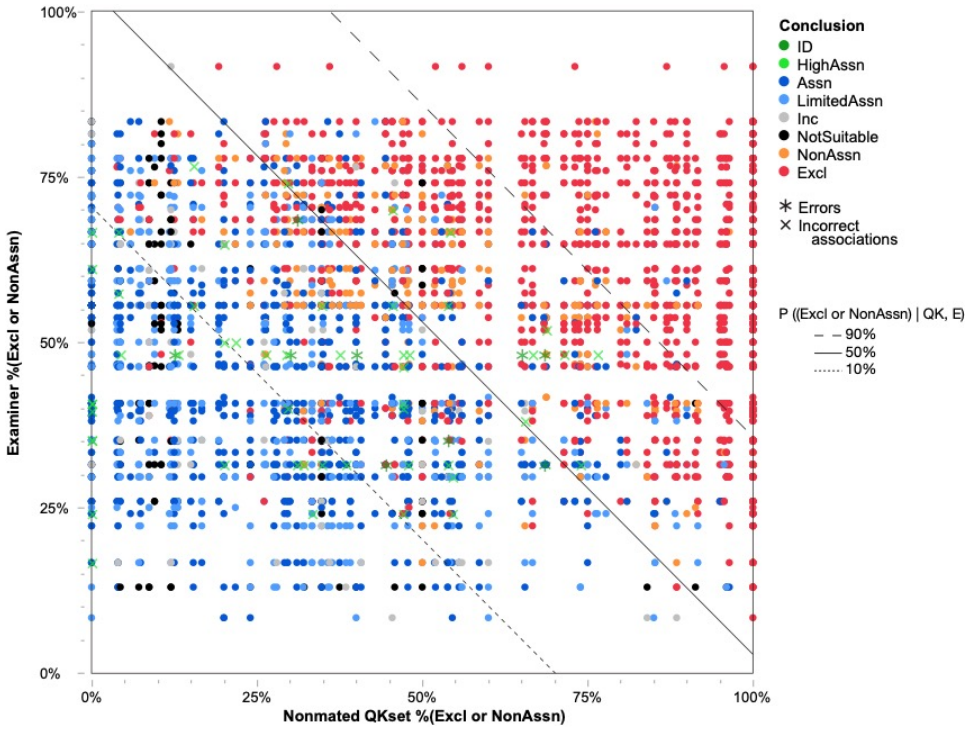

Fig S21. Sample effects vs examiner effects: 162 nonmated QKsets, plotted by (x axis) the percentage of examiners who made Excl or NonAssn conclusions on that QKset, and (y axis) the percentage Excl or NonAssn conclusions made by that examiner; the diagonal lines represent {90%, 50%, 10%} probabilities of Excl or NonAssn decisions as predicted by logistic regression. (*Baseline Dataset*)

Appendix J Associations Between Participant Attributes and Performance

This appendix reports associations between performance and attributes of the participants. *Appendix J1* reports associations between performance and the participants' responses on the background questionnaire. *Appendix J2* discusses the only other participant attribute that had a notable association with performance, participants' use of software during the study.

Appendix J1 Participant Background Associations with Performance

The performance of 71 FFEs, each of whom completed at least 40% of all assigned QKsets, was evaluated with respect to 16 background attributes of interest (Table S25), using variable importance analysis (VIA) and attribute-specific significance testing. VIA considers all variables simultaneously by leveraging both linear regression and random forest techniques, yielding importance scores; significance testing via the Kruskal-Wallis test was conducted for each attribute individually to assess for differences between groups, resulting in *p*-values and Benjamini-Hochberg (BH) *q*-statistics.

Our use of VIA importance scores, *p*-values, and BH *q*-statistics was developed for the Bloodstain Pattern Analysis Black Box study, and is formally detailed in *Appendix 2.8* of that publication [22], but summarized here for convenience:

- Our use of VIA couples linear regression with random forest analysis. Linear regression was used to associate subsets of background attributes with each performance measure, and random forest analysis was then used to determine the importance of each background attribute based on the goodness of fit measures from the linear regression models. Although linear regression is commonly used to associate various factors with an observation of interest, it is generally not a suitable technique when the ratio of observations to covariates is low, and/or when there is high correlation between predictor variables [44,45], both of which are the case for this study. Alternatively, random forest regression, a non-parametric technique, is robust even in the presence of small sample sizes and multicollinearity [45,46]. However, the primary issues for random forest, particularly as related to variable importance, are selecting an unbiased estimator of importance and highlighting important variables even in the presence of redundancy [45,46]. To overcome these limitations, linear regression was coupled with random forest analysis to conduct variable importance analysis.
- To evaluate whether performance varied between different groups of analysts for each background attribute, the Kruskal-Wallis (KW) test was used. This non-parametric alternative to the traditional t-test or ANOVA analysis does not require that responses be normally distributed [47], which was not necessarily expected for the performance measures utilized in this study. The Kruskal-Wallis test statistic is approximately chi-square distributed (k-1 degrees of freedom). Therefore, the test statistic is compared to the chi-square distribution to obtain a *p*-value. In addition, the Benjamini-Hochberg *q*-statistic (BH(q)) was computed given the large number of background attributes under consideration [48]. The BH(q) is essentially a *p*-value adjusted based upon the false discovery rate, rather than the family-wise false positive rate [48,49]. It is generally used to control for the detection of spurious effects with increased power when many individual tests are conducted, wherein a traditional Bonferroni-adjustment becomes overly stringent and conservative [49].

Effect thresholds were set for each of these significance measures (VIA importance scores, *p*-values, and *q*-statistics) as detailed in [22], summarized here for convenience:

- For variable importance analysis, the threshold for counting as an important variable is being classified as an extreme outlier with respect to percent increase in mean squared error. For the purposes of this study, an extreme outlier was defined as an attribute having an importance value greater than the third quartile (Q3, 75th percentile) by more than three times the interquartile range (Q3 + 3IQR).
- For the Kruskal-Wallis *p*-value, the association threshold is 0.05. Note that 5% of tests may meet this criterion by chance alone.
- To control for spurious effects, the BH(q) for the Kruskal-Wallis test was also considered when evaluating the degree of association between attributes and performance. Any background attribute whose KW *q*-statistic falls below a significance level cutoff of 0.10 meets the association threshold for BH(q). Theoretically, 10% of all detected significant results will truly be null at this significance level (e.g., if there are 5 significant attributes, just 0.5 of those may be expected to be false rejections of a true null).

Because we used different performance measures for this study, as compared to [22], we developed a new proposed association reporting hierarchy:

- If an attribute meets the criteria for all three association evaluations (variable importance analysis, Kruskal-Wallis *p*-value, and BH *q*-statistic), then the support for an association is considered ***notable*** and reported.
- If an attribute meets the criteria for two of the association evaluations (variable importance analysis, Kruskal-Wallis *p*-value, and/or BH *q*-statistic), then the support for an association is considered ***limited*** and reported with qualifications.
- Otherwise, there is insufficient support to indicate a meaningful association between the attribute and performance and no association is reported.

Attribute	Survey Q	Variable Type	Levels
Age	1	Ordinal	< 40, 40-49, 50+
Education	2	Ordinal	< Bachelors, Bachelors, Graduate
Experience	3	Ordinal	< 5 years, 5-10 years, 11-15 years, 16+ years
Examination frequency	4	Ordinal	Monthly, weekly, yearly
# times testified	5A	Ordinal	0, 1-9, 10-19, 20+
# other disciplines	6	Ordinal	< 2, 2, 3, 4, 5+
Training program	7	Binary	No formal, formal (6-12 mo), formal (1+ year)
Certification status	8	Binary	Have been certified, never certified
Last proficiency test	9	Ordinal	Never, more than 1 year ago, within 1 year
Casework impression types	13A/B	Categorical	2D, 3D, both (40/60, 60/40, or 50/50 split)
Type of examination	16	Categorical	Physical, digital, both
Employer	17	Categorical	US local, US State, Intl Gov, Other
Employer accreditation	17A	Categorical	Not accredited, accredited, unsure
# other FFEs	18	Ordinal	0, 1, 2, 3, 4
Conclusion scale	19	Categorical	SWGTREAD 2006, SWGTREAD 2009, other
Blind verification policy	22	Binary	No blind verification, blind verification

Table S25. Background attributes of interest with associated survey question numbers and variable information. Note that in some cases response categories appearing on the survey have been combined to ensure sufficient sample sizes for comparison (minimum of 5 for any category).

Table S26 details the results for variable importance analysis and significance testing of the 16 background attributes of interest, with respect to the four performance ratios described in *Appendix I*. The majority of background attributes — including level of education, experience, examination frequency, and certification status — did not exhibit support for an association with performance.

Just one background attribute, employer, exhibited strong support for an association with performance as a function of TN-CN ratio. Fig S21A displays the distribution of weighted TN-CN ratio as a function of employer. Based upon a Bonferroni-adjusted Dunn's post-hoc analysis, participants who were employed by US local agencies generally had a higher weighted ratio of correct *NonAssns* and *Excls* than those employed by international governments (p = 0.0019). This observation is likely driven by differences in reporting tendencies between the two groups, and a lack thereof between participants employed by US state or other agencies. Based upon the results of a Kruskal-Wallis analysis, participants employed by US local agencies were significantly more likely to report a definitive conclusion than those employed by an international government agency (p = 0.0012); furthermore, participants employed by US local agencies were significantly less likely to report a class association than their international government agency counterparts (p = 0.0427). Because participants from US local agencies were much more likely to report a definitive conclusion, this consequently increased their potential for a higher weighted TN-CN ratio, as opposed to participants from international governments who were more conservative and more often reported class associations (which do not contribute to the weighted TN-CN ratio).

In addition, there was limited support for an association between training program and FN-IN ratio, but this result should be interpreted with caution given that does not meet the criteria for the q-value significance test (as detailed in [22]) and we thus cannot preclude the possibility that this is a spurious effect. Fig S21B displays the distribution of the weighted FN-IN ratio as a function of training program; note the large differences in variance between Formal (>1 year) and the other groups—in this scenario the KW test is measuring differences in dominance rather than median. Based upon a Bonferroni-adjusted Dunn's post-hoc analysis, participants who completed 1+ year of formal training were much more likely to exhibit notably lower rates of erroneous and incorrect non-associations than those who completed 6-12 months of formal training (p = 0.0246). No difference was detected between participants who had no formal training and either of the groups who completed formal training.

The BH q-statistic is affected by the number of attributes considered: since testing large numbers of attributes raises the potential for some attributes to meet significance thresholds by chance, the BH q-statistic essentially raises significance thresholds based on the number of attributes. To evaluate the impact of the set of attributes used and whether associations were being diluted due to the number of attributes, we also conducted the background versus performance analysis on a subset of ten core background attributes — education, experience, examination frequency, number of times testified, training program, certification status, last proficiency test, employer, employer accreditation, and conclusion scale. The results on this subset were in complete alignment with those obtained on the full set of 16 attributes: there were no changes in significance for any of the three association evaluations (VIA, KW p-value, BH q-statistic), and accordingly no differences in the associations or strength of associations that we detected.

	TP-CA Ratio			TN-CN Ratio			FP-IA Ratio			FN-IN Ratio		
	VIA (4.84)	P (0.05)	Q (0.10)	VIA (3.60)	P (0.05)	Q (0.10)	VIA (4.81)	P (0.05)	Q (0.10)	VIA (5.14)	P (0.05)	Q (0.10)
Age	0.15	0.80	0.94	0.95	0.27	0.41	0.33	0.57	0.91	0.25	0.27	0.61
Education	1.49	0.40	0.91	0.56	0.35	0.45	0.18	0.48	0.91	0.27	0.43	0.71
Experience	0.24	0.37	0.91	0.61	0.83	0.83	0.19	1.00	1.00	2.17	0.65	0.80
Exam Frequency	0.63	0.29	0.91	0.50	0.39	0.45	0.27	0.49	0.91	0.35	0.44	0.71
Testified	0.15	0.73	0.94	0.80	0.44	0.47	**7.09**	0.89	0.95	0.28	0.11	0.35
Number Other Disc	0.87	0.18	0.91	0.93	0.11	0.37	2.17	0.45	0.91	0.38	0.77	0.88
Training Program	0.14	0.94	0.94	1.12	0.20	0.40	0.14	0.40	0.91	**13.60**	**0.03**	0.35
Certification Status	0.12	0.46	0.93	0.48	0.23	0.40	0.23	0.68	0.92	0.32	0.60	0.79
Last Proficiency Test	0.28	0.75	0.94	1.91	**0.01**	0.10	0.18	0.71	0.92	0.33	0.34	0.68
Casework Impression Types	1.31	0.71	0.94	2.34	0.10	0.37	2.27	0.75	0.92	1.47	0.09	0.35
Type of Examination	0.70	0.17	0.91	1.14	0.39	0.45	1.24	0.10	0.91	1.37	0.90	0.94
Employer	1.39	0.37	0.91	**33.53**	**0.00**	**0.07**	1.76	0.31	0.91	1.89	0.10	0.35
Employer Accreditation	1.86	0.90	0.94	0.73	0.06	0.30	0.57	0.57	0.91	1.47	0.94	0.94
Other FFEs	0.15	0.68	0.94	**6.94**	0.15	0.37	1.08	0.87	0.95	1.65	0.08	0.35
Conclusion Scale	2.33	0.94	0.94	1.16	0.28	0.41	0.87	0.56	0.91	0.56	0.57	0.79
Blind Verification Policy	0.28	0.26	0.91	0.56	0.16	0.37	1.21	0.50	0.91	0.24	0.24	0.61

Table S26. Attribute versus performance results for three association evaluations: variable importance analysis and significance testing (Kruskal-Wallis *p*-values and BH *q*-statistics). The association threshold for each measure is listed in parentheses. Cells highlighted yellow meet one of the three association criteria (insufficient to indicate a meaningful association). Cells highlighted blue meet two of the three association criteria (limited support for association). Cells highlighted green meet the all three association criteria (notable support for association).

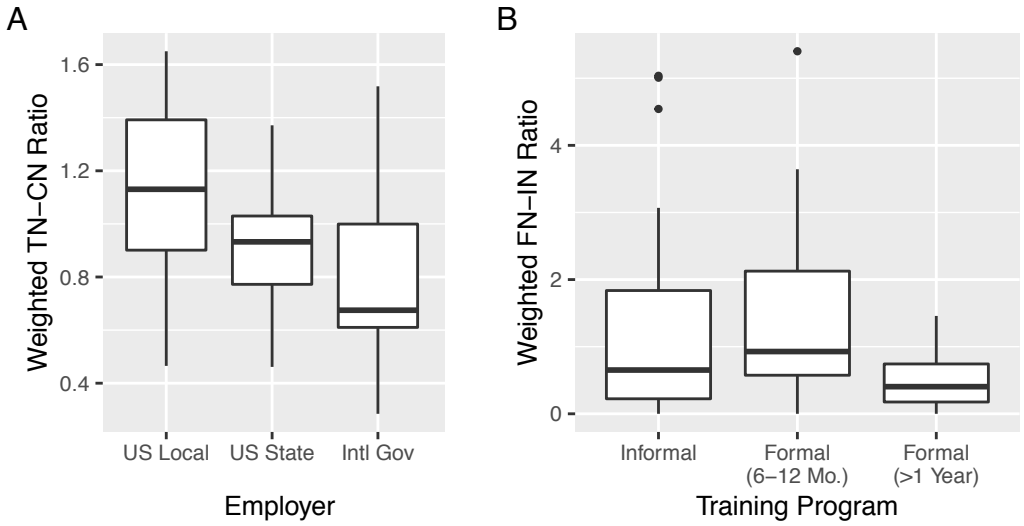

Fig S22. Distribution of weighted performance measures for background attributes exhibiting strong support (A) or limited support (B) for association with performance. (*Examiner Comparison Dataset*; Employer (plot A) omits 3 participants who do not fit in the 3 categories shown.)

Appendix J2 Participants' Use of Software

This section includes the only participant-specific attribute that had a notable association with performance. Note this is not included in Appendix J1 because those attributes are from the background survey, whereas this is a result of the study itself.

In comparison question #9, participants were asked "Did you use additional software (such as Adobe Photoshop) to view or process/enhance any of the high-resolution images in this comparison set?" Note that some of the study was conducted during COVID-19, which may have limited some particants' access to software.

Participants generally did not use software. Out of 6,032 trials in the *Baseline Dataset,*

- 788 responses (13%): "Yes, software was used to process/enhance one or more image(s);"
- 879 responses (15%): "Yes, software was used, but only to view images;"
- 4,365 responses (72%): "No."

Use of software had no notable association with accuracy of results for the specific trials it was used on.

Out of the 71 participants in the *Examiner Comparison Dataset*,

- 23 (32%) never used software;
- 28 (39%) did not use software in the majority of trials;
- 18 (25%) used software in the majority of trials;
- 2 (3%) used software in every trial.

There was a statistically significant association between the use of software and the rates of correct *IDs* and *HighAssns*. Fig S22displays the distribution of performance (according to the weighted TP-CA ratio) as a function of frequency of software usage in the comparisons in this study. Based upon a Bonferroni-adjusted post-hoc analysis, those who never used software generally had lower rates of correct associations than those who used software a majority of the time ($p = 0.0023$).

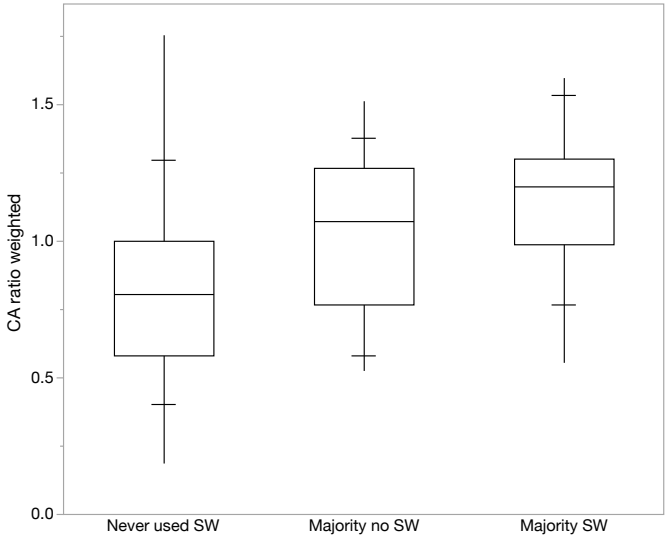

Fig S23. Distribution of performance measure (CA ratio) as a function of casework software usage. (*Examiner Comparison Dataset*)

Appendix K Effects of Unused Footwear Items

As discussed in *Appendix C4.3*, six QKsets were included to evaluate the effects of Qs from brand new footwear items. The study included three mated QKsets ("new-used") in which the Q was from an unused footwear item and the K was from the same footwear item after it was worn for up to two weeks, and three nonmated QKsets ("new-new") in which the Q was from an unused footwear item and the K was from a different unused footwear item (of the same make, model, size, and foot). Each participant who completed the study received one of each. These Qs were collected to be as close to ideal as possible, so the walking test impression protocol used in this study (see "Walking Test Impressions" under *Appendix C4.1*) was employed there. Fig S23 shows the resulting conclusions.

Two of these QKsets (QK035 and QK050) were created using Converse Chuck Taylor All Star (men's size 11) shoes. The other four QKsets (QK040, QK182, QK273, and QK292) were created using Vans Classic (men's size 11.5) shoes. These two specific footwear models both include manually applied foxing strips, toe guards, and heel labels; because they were applied by hand, the specific position of characteristics associated with these footwear components can vary and thus could provide a means of discriminating the footwear items of the same make, model, and size. The images depicting the questioned impressions, outsoles (including both the ones used to prepare the Qs and the unused Ks), and test impressions in these sets were all reviewed in detail by FFEs on the study team. In that review, the following apparent manufacturing artifacts were observed:

- Several manufacturing artifacts were observed on the Converse shoe which was used as both Q and K in QK035 (mated) and used as Q in QK050 (nonmated); some of these manufacturing artifacts also appear in the (nonmated) K for QK050.
- All the new Vans shoes share the same two manufacturing artifacts, linear features present in the medial and lateral sides of the ball region: the Qs and Ks in QK040 (mated), QK182 (mated), and QK273 (nonmated) all share these two manufacturing artifacts. For QK292 (nonmated), the Q and K shared manufacturing subclass characteristics in addition to the two present in all of the Vans, ***and*** differed in manufacturing subclass characteristics.

Even though these are based on a small number of QKsets, several of the results are notable:

- On the mated ("new-used") QKsets, 27% of responses were *IDs* (TPs), and 27% were *HighAssns* (CAs). Although these responses are consistent with ground truth, our review of the items suggests that apparent subclass manufacturing artifacts may have been mistakenly considered RACs, calling into question the basis for the *ID* (and arguably *HighAssn*) responses.

- On the mated ("new-used") QKsets, 17% of responses were *Excl* (FNs). QK040, one of the sets that contained a K that was worn for up to two weeks after Q collection, had a 32% FNR (3rd highest in the study) and an 8% INR. Review of the items in this QKset shed some light on the source of the FNs. The shoe acquired several discrete damage features (RACs) during the two-week period without evidence of overall wear; the participants may have improperly interpreted their findings as justification for *Excl*.
- On the nonmated ("new-new") QKsets, 14% of the responses were incorrect *HighAssns* (IAs), and there was one erroneous *ID* (FP). We observed manufacturing artifacts in both the (nonmated) Q and K, which may have been used as the basis for the incorrect and erroneous conclusions.
- On the nonmated ("new-new") QKsets, 12% of responses were *Excls* (TNs), and 8% were *NonAssn* (CNs), all of which can be explained by differences in manufacturing artifacts, including features associated with the the toe guards and heel labels, thus provide a means of discriminating between Q and K, explaining the *Excls* and *NonAssns*.

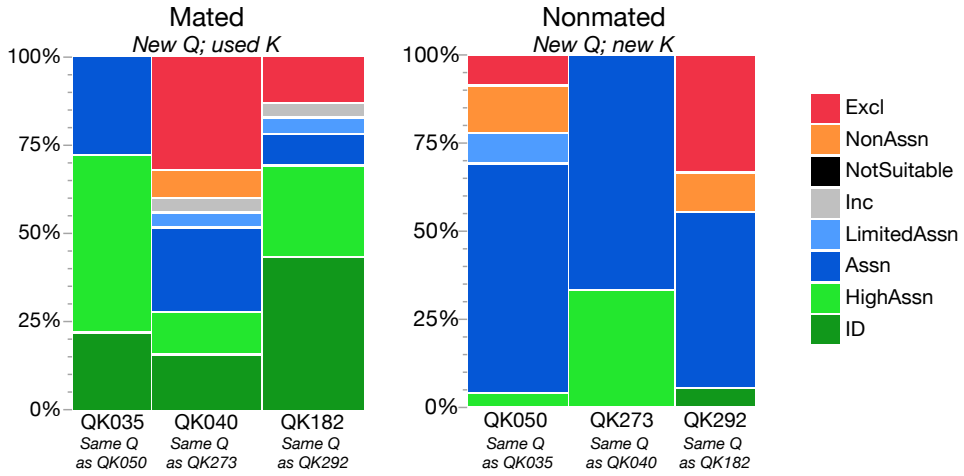

Fig S24. Conclusions for QKsets with unused Qs. (Subset of *Baseline Dataset*. Mated: 66 trials on 3 QKsets. Nonmated: 65 trials from 3 QKsets.)

Appendix L Effects of Wear between the Q and K

In the instructions, participants were informed "There may be up to two weeks of wear between the dates the questioned impressions and the knowns (test impressions and outsoles) were collected." The study included six mated QKsets in which the footwear items were worn between the Q and K (in addition to the three mated QKsets with unused Qs discussed in the previous section); 98 mated QKsets were not worn between the Q and K. Fig S24 shows the resulting distribution of conclusions. The summary chart (Fig S24 left) shows that the overall rates of *LimitedAssn* and *Assn* were higher (and rates of *HighAssn* and *ID* were lower) than for the other mated QKsets — as would be expected. Fig S24 (right) breaks down the results by QKset to illustrate that the distributions of conclusions varied markedly among the six QKsets, and only somewhat tracked with the quality of the Qs. Fig S24 (left) shows that the rate of erroneous *Excls* was almost identical between the two groups (but note that in the previous section Fig S23 shows one example of a QKset with a high FNR due to extensive wear).

In short, wear between the Q and K increased the relative proportion of class associations and decreased the proportion of definitive and probable conclusions.

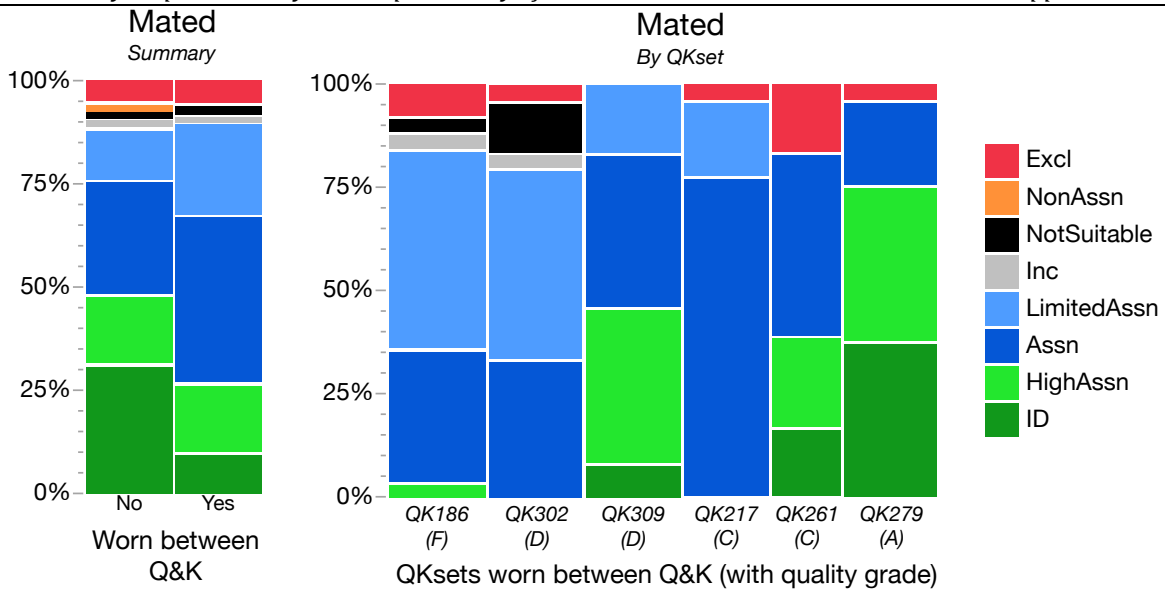

Fig S25. Conclusions for mated QKsets in which the footwear item was worn up to two weeks between the time the Q was collected and the time the K was was collected. (Subset of *Baseline Dataset*: 137 trials on 6 mated QKsets. Does not include the three mated QKsets ("new-used") discussed in the previous section in which the Q was unused and the K was from the same footwear item after it was worn for up to two weeks)

Appendix M Comparison of Results to WVU Study

With respect to studies evaluating the accuracy and reliability of forensic expert conclusions, PCAST recommends that "to ensure that conclusions are reproducible and robust, there should be multiple studies conducted by separate groups reaching similar conclusions" (Criterion 6 in Box 4: Key criteria for validation studies to establish foundational validity) [2]. As previously discussed, another footwear examination black box study was recently conducted and published by a research team from West Virginia University in response to the PCAST recommendations [10–12]. The following sections compare the results obtained in the current study with those obtained by the WVU study in an effort to satisfy this criterion.

Appendix M1 Accuracy, Error Rates, and Predictive Values

The WVU study generally evaluated accuracy with respect to both ground truth regarding known source and an "acceptable" range of conclusions [10–12]. The authors generally reported accuracy and error rates using the following metrics: correct within rate, correct outside rate, incorrect within rate, and incorrect outside rate. This procedure differed from the one utilized in the current study, wherein accuracy and error rates were computed solely based upon ground truth, irrespective of the available features.

Because the authors of the WVU study provided a breakdown of the distribution of examiner conclusions for each comparison, we were able to compute accuracy, error, and predictive values for the results from the WVU in the same manner employed in the current study (i.e., based solely upon ground truth). Table S27 summarizes the key metrics from the current study as compared to those computed based upon the WVU study. Overall, the rates are rather consistent, with confidence intervals for the two studies overlapping across nearly all metrics. The largest differences in rates were observed with respect to accurate conclusions on nonmated pairs; as highlighted in Table S27, the current study exhibited lower true negative and correct non-association rates. This observation could be attributed to sample differences or differences in study design. In particular, some of the questioned impressions in the WVU study had two knowns provided (which differs from the current study wherein we only provided a single known); the authors of the WVU study note that in some responses, it was clear that participants could take advantage of this aspect of the WVU study design by reporting *Excl* for a given known when they identified the other known in the case, and therefore the resulting conclusions are not necessarily independent.

Metric	*Abbreviation*	**Current Study**			**WVU Study**		
		Rate	*Counts*	*C.I.*	*Rate*	*Counts*	*C.I.*
Positive predictive value	PPV	98.8%	(725/734)	[97.7%-99.4%]	100.0%	(100/100)	[96.4%-100%]
False discovery rate	FDR; (1-PPV)	1.2%	(9/734)	[0.6%-2.3%]	0.0%	(0/100)	[0.0%-3.6%]
Positive predictive association	PPA	95.1%	(1135/1194)	[93.7%-96.2%]	98.8%	(162/164)	[95.7%-99.9%]
Negative predictive value	NPV	91.2%	(1515/1661)	[89.7%-92.5%]	94.6%	(350/370)	[91.8%-96.7%]
False omission rate	FOR; (1-NPV)	8.8%	(146/1661)	[7.5%-10.3%]	5.4%	(20/370)	[3.3%-8.2%]
Negative predictive association	NPA	90.8%	(1861/2050)	[89.4%-92.0%]	93.3%	(416/446)	[90.5%-95.4%]
True positive rate (TPR)	TPR_{PRES}	30.0%	(725/2417)	[28.2%-31.9%]	28.6%	(100/350)	[23.9%-33.6%]
Correct association rate	CAR_{PRES}	17.0%	(410/2417)	[15.5%-18.5%]	17.7%	(62/350)	[13.9%-22.1%]
True positive + Correct association rate	$TPR+CAR_{PRES}$	47.0%	(1135/2417)	[45.0%-49.0%]	46.3%	(162/350)	[41.0%-51.7%]
True negative rate (TNR)	TNR_{PRES}	41.9%	(1515/3615)	[40.3%-43.5%]	72.2%	(350/485)	[67.9%-76.1%]
Correct non-association rate	CNR_{PRES}	9.6%	(346/3615)	[8.6%-10.6%]	13.6%	(66/485)	[10.7%-17.0%]
True negative + Correct non-association rate	$TNR+CNR_{PRES}$	51.5%	(1861/3615)	[49.8%-53.1%]	85.8%	(416/485)	[82.3%-88.8%]
False positive rate (FPR)	FPR_{PRES}	0.2%	(9/3615)	[0.1%-0.5%]	0.0%	(0/485)	[0.0%-0.1%]
Incorrect association rate	IAR_{PRES}	1.4%	(50/3615)	[1.0%-1.8%]	0.4%	(2/485)	[0.0%-1.5%]
False positive + Incorrect association rate	$FPR+IAR_{PRES}$	1.6%	(59/3615)	[1.2%-2.1%]	0.4%	(2/485)	[0.0%-1.5%]
False negative rate (FNR)	FNR_{PRES}	6.0%	(146/2417)	[5.1%-7.1%]	5.7%	(20/350)	[3.5%-8.7%]
Incorrect negative association rate	INR_{PRES}	1.8%	(43/2417)	[1.3%-2.4%]	2.9%	(10/350)	[1.4%-5.2%]
False negative + incorrect non-association rate	$FNR+IAR_{PRES}$	7.8%	(189/2417)	[6.8%-9.0%]	8.6%	(30/350)	[5.9%-12.0%]

Table S27. Accuracy, error rates, and predictive values for the current study compared to the recently published WVU study [10–12]. Any metrics with Clopper-Pearson confidence intervals that do not overlap between the two studies are highlighted in blue.

Appendix M2 Consensus

The WVU study evaluated examiner agreement (which we refer to as consensus) using interquartile range (IQR) [10–12]. The authors also report the mean proportion of responses that fell within the IQR as a means of communicating the overall level of examiner agreement for the study. For the current study, we report interquartile range for each QKset in Figure 7. However, several QKsets contain percentile cutoffs that split conclusions, causing a decision category to be simultaneously within and outside the IQR (e.g., if the 15th to 30th percentile of responses is *Assn*, then *Assn* would be considered both in and out of the IQR). To report the proportion of conclusions that fall within/outside this consensus range, it is thus necessary to adjust the raw IQR so that it does not split conclusion categories. This was achieved in three manners:

- Lenient IQR: expansion/dilation of the range — For any categories that are split, round the lower bound down and round the upper bound up to include the split category in the adjusted IQR. For example, if 5 responses of *Assn* fall outside the raw IQR and 5 fall within the raw IQR then all *Assn* responses are included in the adjusted IQR.
- Moderate IQR: adaptive adjustment of the range — For any categories that are split, round to the nearest possible conclusion in both directions. For example, if 7 responses of *Assn* fall outside the raw IQR and 3 fall within the raw IQR then all *Assn* responses are excluded from the adjusted IQR based upon the density of responses.
- Stringent IQR: contraction/closing of the range — For any categories that are split, round the lower bound up and round the upper bound down to exclude the split category from the adjusted IQR. For example, if 5 responses of *Assn* fall outside the raw IQR and 5 fall within the raw IQR then all *Assn* responses are excluded from the adjusted IQR.

Table S28 compares the examiner agreement results for participants in the current study versus those in the WVU study. Again, the results are extremely consistent for both studies. Based upon mean $\pm$ one standard deviation, the proportion of responses for mated trials, nonmated trials, and overall do not differ substantially, irrespective of the IQR adjustment method.

	Current Study			**WVU Study**
	Lenient	*Moderate*	*Stringent*	
Mated	84.0% $\pm$ 9.7%	82.8% $\pm$ 10.1%	81.0% $\pm$ 11.1%	79.7% $\pm$ 14.1%
Nonmated	88.6% $\pm$ 9.9%	87.4% $\pm$ 10.2%	85.4% $\pm$ 11.1%	89.8% $\pm$ 6.7%
Total	86.8% $\pm$ 10.1%	85.6% $\pm$ 10.4%	83.7% $\pm$ 11.3%	85.6% $\pm$ 11.1%

Table S28. Consensus results (via mean $\pm$ SD percentage of conclusions reported that fall inside the IQR) for the current study compared to the recently conducted WVU study [10–12].

Appendix M3 Inter-Rater Reliability

The WVU study also reports inter-rater reliability for conclusions reached by participating FFEs using Gwet's AC2 [10–12,50]. Rather than simply measuring agreement, Gwet's AC2 accounts for both chance agreement (i.e., two examiners agree by chance—such as when struggling to choose a decision category) as well as the level of disagreement (i.e., two examiners reporting *ID* and *HighAssn* disagree much less than two examiners reporting *ID* and *Excl*).

Table S29 details the inter-rater reliability results for the current study versus the WVU study. Overall, both studies found moderate to substantial agreement in the conclusions reached by participating FFEs. The variation in coefficients and the comparison types that yielded moderate versus substantial agreement may be explained by a variety of factors, including the

high true negative rate in the WVU study, different number of examiners completing each comparison, different number of total comparisons, and different conclusion scales used for collecting responses. Nonetheless, these results do indicate relatively high reliability of FFE conclusions, even using the seven-level conclusion scale.

	Current Study			WVU Study		
	Gwet AC2	*SE*	*Verbal Equivalent*	*Gwet AC2*	*SE*	*Verbal Equivalent*
Mates	0.6730	0.0255	Substantial	0.6562	0.1369	Moderate
Nonmates	0.6810	0.0223	Substantial	0.8818	0.0546	Substantial
Total	0.6170	0.0172	Moderate	0.7509	0.0875	Substantial

Table S29. Inter-rater reliability results for the current study compared to the recently conducted WVU study [10–12].

Appendix N Participant Assessments of Class Characteristics

This appendix provides support for Section 5, Additional Results.

As part of the comparison process, participants were asked to assess whether the Q and K correspond in design, size, mold, and wear; these assessments were not asked if participants indicated that the Qs were *NotSuitable*. (See *Appendix C5.4* for instructions provided to participants.) These responses can be evaluated in terms of the QKset class characteristic categories (*Appendix N1*), or in terms of the participants' own conclusions (*Appendix N2*). *Appendix N3* reports the reproducibility and repeatability of these assessments.

Appendix N1 Assessments of Class Characteristics vs. Ground Truth

During comparison, participants assessed whether the Q and K correspond in design, size, mold, and wear (indicating "same," "different," or "unsure" for each). Assessments of "same" or "different" can be evaluated as correct or incorrect if they are consistent with or contradict the ground truth class characteristics known from the creation of the Qs and Ks. Assessments of "unsure" are not evaluated as correct or incorrect. Assessments that cannot be evaluated against ground truth are listed as debatable. Table S30 details how participants' assessments of class characterictics can be evaluated as correct, incorrect, or debatable with respect to ground truth. Some of the assessments (starred numbers in Table S30) require additional explanation:

- *1: For nonmated QKsets of the same make/model in which the Q and K differed in size, the mold and size were by definition different, but a given impression may or may not have had enough information to make that determination.
- *2: The questions regarding mold, size, and wear were only asked if they indicated same design and same foot.
- *3: For nonmated QKsets of the same make/model/size, the mold may or may not have been the same (ground truth not available).
- *4: For mated QKsets that were worn between the Q and K, there may or may not have been any physical changes to the outsoles in the intervening time (ground truth not available).
- *5: For nonmated QKsets of the same make/model, wear may or may not have been different (ground truth not available).

					Ground truth class characteristics						
					Mated		*Nonmated*				
							Same make and model			*Different make or model*		*Same make/model, Diff foot*
					Unworn	*Worn*	*Same size*	*½ size diff*	*1 size diff*	*Same foot*	*Diff foot*	
Participant assessments	Different Design											
	Same Design	Diff Foot										
		Same foot	Same Foot							*2		
			Size	Diff				*1	*1			
				Same				*1	*1			
			Mold	Diff			*3	*1	*1			
				Same			*3	*1	*1			
			Wear	Diff		*4	*5	*5	*5			
				Same		*4	*5	*5	*5			

Table S30. Participant assessments of class characteristics vs. ground truth. Cells highlighted orange indicate incorrect assessments and cells highlighted blue indicate "correct" assessments. Cells highlighted gray indicate debatable assessments. See text for discussion of the starred cells.

Table S31 summarizes how the participants' assessments of design, size, and mold (rows) relate to the actual similarities and differences in class characteristics between the Qs and Ks in the QKsets (columns). The values in each column indicate the percentage of trials that participants assessed as suitable that fall within the given category. For example, (3rd column, bottom row) the *Baseline Dataset* had 2,193 trials on nonmated QKsets in which the Q and K were the same make, model, foot, and size, and on which the participant assessed the Q as suitable; on 12.2% of those trials (3rd column, 3rd row), the participant assessed the Q and K as the same design and foot (correct), but different size (incorrect). Note that wear is not included in this summary,

because wear is not necessarily a contradiction for any of these columns (e.g., mated QKsets may legitimately have different wear), and therefore we cannot definitively label an assessment of wear as incorrect.

The highlighted cells indicate assessments of design, size, or mold that are contrary to the actual differences in the QKsets. Note that we do not assess "unsure" as correct or incorrect: highlighted cells are incorrect, but non-highlighted cells are not necessarily correct. Nonmated QKsets in which the Q and K were the same make, model, and size we know are the same design and size, but do not know whether they were produced using the same mold (therefore, the 1.7% in the 3rd column, 6th row is not highlighted). The next to last row totals the incorrect assessments. On nonmated QKsets, assessments of design or size were often incorrect: 14.6% of trials when the Q and K were the same make, model, and size; 47.8% when the Q and K differed by a ½ size; 34.1% when the Q and K differed by one size, 22.6% when the Q and K were of different makes or models; and 13-15% when the Q and K were from opposite feet (note the small number of trials for different feet).

		Actual QKset Type (% of trials assessed as suitable)							
				Nonmated					
		All	*Mated*	*Same make, model, and foot*			*Different make or model*		*Same make/model, Diff foot*
				Same size	*½ size diff*	*1 size diff*	*Same foot*	*Diff foot*	
Same design, opposite foot		0.8%	0.2%	0.7%	0.2%	0.7%	0.6%	10.0%	87.0%
Diff design		6.9%	0.5%	1.7%	1.9%	4.0%	68.4%	80.0%	8.7%
Unsure design		2.9%	1.6%	3.1%	0.4%	5.3%	9.0%	5.0%	-
Same design and foot	Diff size	11.9%	2.6%	12.2%	37.8%	43.0%	11.6%	5.0%	4.3%
	Unsure size	10.6%	7.3%	13.8%	13.9%	17.5%	6.0%	-	-
Same design, foot, and size	Diff mold	1.0%	0.7%	1.7%	0.6%	0.3%	0.6%	-	-
	Unsure mold	8.5%	8.8%	10.8%	7.7%	4.0%	1.1%	-	-
	Same mold	57.3%	78.4%	56.0%	37.4%	25.2%	2.6%	-	-
Incorrect assessments (sum of highlighted cells)			4.0%	14.6%	47.8%	34.1%	22.6%	15.0%	13.0%
# of trials assessed as suitable		5,862	2,378	2,193	481	302	465	20	23

Table S31. Summary of associations between participants' assessments of design, size, and mold (rows) and the actual differences in the QKsets (columns), expressed as percentages of the trials assessed as suitable. Highlighted cells indicate assessments of design, size, or mold that are incorrect (i.e. contrary to the actual differences in the QKsets). (Subset of *Baseline Dataset*: see Table S32 for additional detail.)

Fig S25 provides a different summary of the same data, but focuses on how participants' assessments of design, size, and mold were associated with the quality of the questioned impression. There is a general trend of increased uncertainty with lower-quality questioned impressions. For QKsets of the same make, model, and size (both mated and nonmated), participants made

fewer "same" assessments (blue) as quality declined; similarly, for nonmated QKsets of different make, model, or size, participants generally made fewer "different" assessments (red) as quality declined.

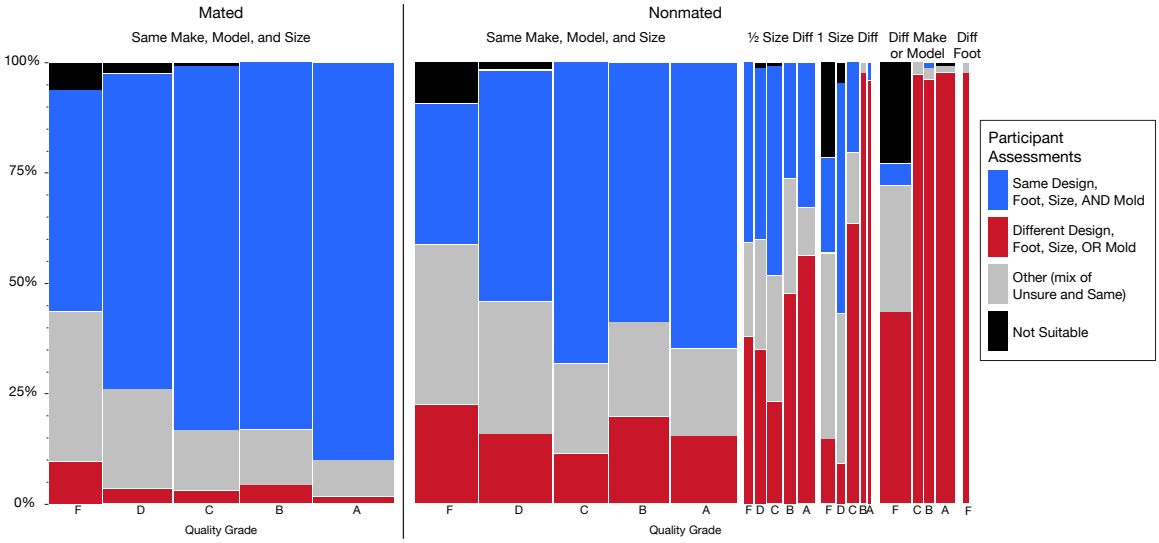

Fig S26. Summary of participants' assessments of design, size, and mold (color-coding) with respect to QKset type and quality grade. Blue indicates trials in which the participant assessed the Q and K as the same in terms of design, foot, size, and mold; red indicates trials with any differences in those assessments; gray is the residual, in which at least one assessment was unsure and the remainder were same. (*Baseline Dataset*)

Table S32 provides detail for the data summarized in Table S31, with the addition of assessments of wear.

					Actual QKset Type (# of trials)						
							Nonmated					
					All	*Mated*	*Same make,model,size*	*½ size diff*	*1 size diff*	*Diff make or model*	*Diff make/model, Diff foot*	*Same make/model, Diff foot*
Participant assessments	NotSuitable				170	39	51	2	25	53	-	-
	Same design, opposite foot				48	4	16	1	2	3	2	20
	Diff design				406	12	37	9	12	318	16	2
	Unsure design				168	38	69	2	16	42	1	-
	Same design and foot	Diff size	Diff mold	Diff wear	135	13	47	26	39	10	-	-
				Unsure wear	45	7	14	8	12	4	-	-
				Same wear	2	-	1	1	-	-	-	-
			Unsure mold	Diff wear	80	8	35	18	12	7	-	-
				Unsure wear	47	7	13	17	8	2	-	-
				Same wear	1	-	-	1	-	-	-	-
			Same mold	Diff wear	253	15	104	72	37	25	-	-
				Unsure wear	109	9	43	33	17	5	1	1
				Same wear	25	3	10	6	5	1	-	-
		Unsure size	Diff mold	Diff wear	26	4	21	-	-	1	-	-
				Unsure wear	12	2	6	1	-	3	-	-
				Same wear	-	-	-	-	-	-	-	-
			Unsure mold	Diff wear	35	8	25	1	1	-	-	-
				Unsure wear	183	61	75	23	12	12	-	-
				Same wear	6	2	3	-	1	-	-	-
			Same mold	Diff wear	70	9	41	15	4	1	-	-
				Unsure wear	246	67	116	20	32	11	-	-
				Same wear	45	20	15	7	3	-	-	-
		Same size	Diff mold	Diff wear	37	5	27	2	1	2	-	-
				Unsure wear	10	1	7	1	-	1	-	-
				Same wear	13	10	3	-	-	-	-	-
			Unsure mold	Diff wear	126	19	95	10	1	1	-	-
				Unsure wear	267	110	116	26	11	4	-	-
				Same wear	107	80	26	1	-	-	-	-
			Same mold	Diff wear	496	72	375	40	8	1	-	-
				Unsure wear	1,017	409	465	86	48	9	-	-
				Same wear	1,847	1,383	388	54	20	2	-	-
	Total trials				6,032	2,417	2,244	483	327	518	20	23
	Subtotal: suitable trials				5,862	2,378	2,193	481	302	465	20	23
	Subtotal: same design and foot				5,240	2,324	2,071	469	272	102	1	1
	Subtotal: same design, foot, and size				3,920	2,089	1,502	220	89	20	-	-

Table S32. Associations between participants' assessments of design, size, mold, and wear (rows) and the actual differences in the QKsets (columns). Highlighted cells indicate assessments of design, size, or mold that are incorrect (i.e. contrary to the actual differences in the QKsets). (*Baseline Dataset*: see Table S31 for summary.)

Participants' assessments of wear had only a limited association with actual wear between the Q and K. For the 98 mated QKsets that were NOT worn between the Q and K, participants assessed wear as the same in 66% of responses, unsure in 28%, and different in 6%. For the 6 mated QKsets that WERE worn between the Q and K, participants assessed wear as the same in 51% of responses, unsure in 42%, and different in 7%. For the 3 mated QKsets for which the Q was unused but the K was worn for up to two weeks before collection, participants assessed wear as the same in 76% of responses, unsure in 12%, and different in 12%. Of the 153 "Diff wear" assessments of mated QKsets shown in Table S32, 136 (89%) were on QKsets in which the footwear was not worn between the Q and K and therefore should have had no differences in wear.

Appendix N2 Assessments of Class Characteristics vs. Conclusions

Table S33 summarizes how the participants' assessments of design, size, and mold (rows) relate to their conclusions (columns). For example, (2nd column, next to last row) out of the 725 trials on mated QKsets that resulted in *IDs,* 93.8% assessed the design, foot, size, and mold were the same for the Q and K. In Table S33 note that in a few trials, the participants' assessments appear to contradict their own conclusions (highlighted). These may be clerical errors, or possibly dissimilarities that the FFE decided fell within the confines of acceptable/expected variation. These apparent contradictions are not limited to one or two participants: for example, there were 11 trials by 9 participants with conclusions of *ID* or *HighAssn* trials that were also assessed as having the Q and K of different size or different mold. Note that none of these incorrect assessments were on erroneous *IDs* (FPs).

			Conclusions (% of trials assessed as suitable)												
		All	*ID*		*HighAssn*		*Assn*		*LimitedAssn*		*Inc*	*NonAssn*		*Excl*	
			M	*NM*	*M*	*NM*	*M*	*NM*	*M*	*NM*		*M*	*NM*	*M*	*NM*
Same design, opposite foot		0.8%	-	-	-	-	-	0.1%	-	0.3%	0.5%	-	0.3%	2.7%	2.6%
Diff design		6.9%	-	-	-	-	-	-	-	-	-	2.3%	1.4%	7.5%	25.7%
Unsure design		2.9%	-	-	-	-	0.1%	0.6%	4.2%	4.3%	34.5%	11.6%	6.4%	1.4%	1.8%
Same design and foot	Diff size	11.9%	0.3%	-	0.2%	-	0.1%	0.4%	0.6%	1.9%	1.0%	14.0%	13.9%	34.2%	37.7%
	Unsure size	10.6%	0.4%	-	1.7%	-	2.5%	4.7%	30.4%	33.0%	38.1%	32.6%	26.9%	6.2%	5.3%
Same design, foot, and size	Diff mold	1.0%	0.8%	-	0.5%	2.0%	0.3%	-	-	-	0.5%	-	1.4%	4.1%	2.4%
	Unsure mold	8.5%	4.7%	11.1%	4.4%	2.0%	12.6%	11.3%	15.4%	13.7%	5.6%	11.6%	11.0%	8.9%	4.6%
	Same mold	57.3%	93.8%	88.9%	93.2%	96.0%	84.3%	83.0%	49.4%	46.6%	19.8%	27.9%	38.7%	34.9%	19.9%
Contradictory assessments (sum of highlighted cells)			1.1%	-	0.7%	2.0%	0.4%	0.5%	0.6%	2.3%	2.0%				
# of trials assessed as suitable		5,862	725	9	410	50	682	852	312	575	197	43	346	146	1,515

Table S33. Summary of associations between participants' assessments of design, size, and mold (rows) and their conclusions (columns). Highlighted cells indicate assessments of design, size, or mold that contradict conclusions. (Subset of *Baseline Dataset*: see Table S34 for additional detail.)

Table S34 provides detail for the data summarized in Table S33, with the addition of assessments of wear. These assessments may indicate possible causes of erroneous *Excls* (FNs) or incorrect *NonAssns* (INs). The pink highlighted cells indicate incorrect assessments of design, size, or mold on mated QKsets, on trials that resulted in FNs or INs: 49% of the *Excls* on mated QKsets had incorrect assessments of design, size, or mold; 16% of the *NonAssns* on mated QKsets had incorrect assessments of design, size, or mold.

Blue highlighted cells indicate assessments of different wear on mated QKsets that might have contributed to FNs or INs:

- For the FNs, 54 responses indicated different wear (but same design, and same or unsure size and mold), of which 43 (29% of all FNs) were not worn between the Q and K and therefore appear to be incorrect assessments of wear; in total, 78% of FNs had incorrect assessments of design, size, mold, or wear.
- For the INs, 11 responses indicated different wear (but same design, and same or unsure size and mold), of which 10 (23% of all INs) were not worn between the Q and K and therefore appear to be incorrect assessments of wear; in total, 40% of INs had incorrect assessments of design, size, mold, or wear.

Participant assessments					*Conclusions (# trials)*												
					All	*ID*		*HighAssn*		*Assn*		*LimitedAssn*		*Inc*	*NonAssn*		*Excl*	
						M	*NM*	*M*	*NM*	*M*	*NM*	*M*	*NM*		*M*	*NM*	*M*	*NM*
	NotSuitable				170	-	-	-	-	-	-	-	-	-	-	-	-	-
	Same design, opposite foot				48	-	-	-	-	-	1	-	2	1	-	1	4	39
	Diff design				406	-	-	-	-	-	-	-	-	-	1	5	11	389
	Unsure design				168	-	-	-	-	1	5	13	25	68	5	22	2	27
	Same design and foot	Diff size	Diff mold	Diff wear	135	-	-	-	-	-	1	1	1	-	3	2	9	118
				Unsure wear	45	-	-	-	-	-	-	-	1	-	-	1	7	36
				Same wear	2	-	-	-	-	-	-	-	-	-	-	-	-	2
			Unsure mold	Diff wear	80	-	-	-	-	-	-	-	2	-	-	7	8	63
				Unsure wear	47	-	-	-	-	-	-	-	-	1	-	13	7	26
				Same wear	1	-	-	-	-	-	-	-	-	-	-	1	-	-
			Same mold	Diff wear	253	-	-	-	-	-	1	1	1	-	1	9	13	227
				Unsure wear	109	1	-	-	-	1	1	-	2	1	1	14	6	82
				Same wear	25	1	-	1	-	-	-	-	4	-	1	1	-	17
		Unsure size	Diff mold	Diff wear	26	-	-	-	-	-	-	-	-	-	-	2	4	20
				Unsure wear	12	-	-	-	-	-	-	-	-	1	1	2	-	8
				Same wear	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			Unsure mold	Diff wear	35	-	-	-	-	2	-	1	-	3	3	11	1	14
				Unsure wear	183	-	-	-	-	3	6	37	63	42	5	21	1	5
				Same wear	6	-	-	-	-	1	2	-	1	-	-	1	1	-
			Same mold	Diff wear	70	-	-	1	-	1	-	4	11	4	-	15	2	32
				Unsure wear	246	-	-	1	-	5	25	47	107	23	5	32	-	1
				Same wear	45	3	-	5	-	5	7	6	8	2	-	9	-	-
		Same size	Diff mold	Diff wear	37	-	-	-	-	-	-	-	-	-	-	2	5	30
				Unsure wear	10	-	-	-	-	-	-	-	-	1	-	2	1	6
				Same wear	13	6	-	2	1	2	-	-	-	-	-	1	-	1
			Unsure mold	Diff wear	126	-	-	-	-	2	7	2	9	2	3	21	12	68
				Unsure wear	267	2	-	3	-	55	69	42	67	9	2	15	1	2
				Same wear	107	32	1	15	1	29	20	4	3	-	-	2	-	-
			Same mold	Diff wear	496	2	-	1	3	11	43	13	37	9	5	69	39	264
				Unsure wear	1,017	17	1	43	2	228	342	108	183	25	4	43	4	17
				Same wear	1,847	661	7	338	43	336	322	33	48	5	3	22	8	21
	Total trials				6,032	725	9	410	50	682	852	312	575	197	43	346	146	1,515
	Subtotal: suitable trials				5,862	725	9	410	50	682	852	312	575	197	43	346	146	1,515
	Subtotal: same design and foot				5,240	725	9	410	50	681	846	299	548	128	37	318	129	1,060
	Subtotal: same design, foot, and size				3,920	720	9	402	50	663	803	202	347	51	17	177	70	409

Table S34. Associations between participants' assessments of design, size, mold, and wear (rows) and their conclusions (columns). Yellow highlighted cells indicate assessments of design, size, or mold that appear to contradict their conclusions. Pink highlighted cells indicate incorrect assessments of design,size, or mold on mated QKsets that resulted in FNs or INs. Blue highlighted cells indicate assessments of different wear on mated QKsets that might have contributed to FNs or INs. (*Baseline Dataset*: see Table S33 for summary.)

Fig S26 shows associations between participants' assessments of wear and conclusions. Assessments of wear are as would be expected, increasingly notably for *NonAssn* and *Excl*.

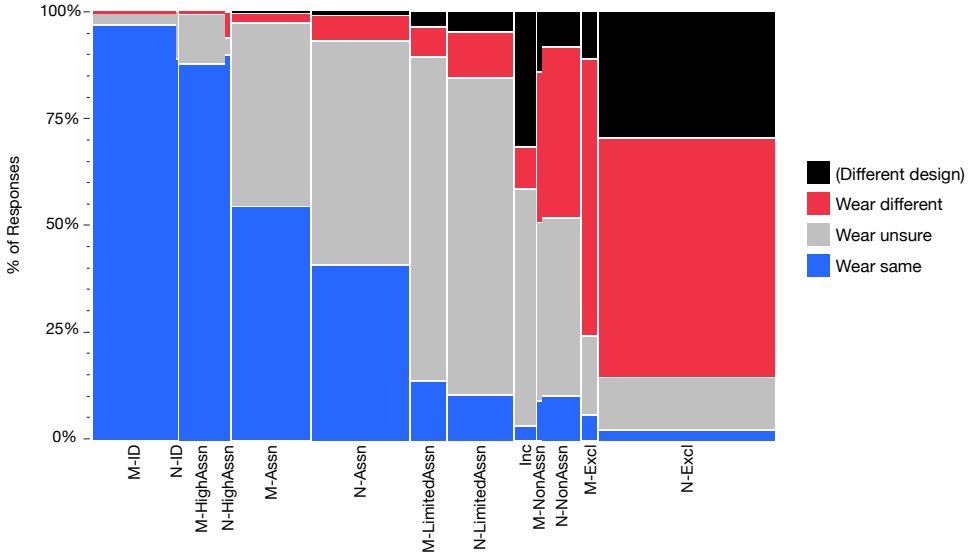

Fig S27. Associations between wear assessments and conclusions. (*Baseline Dataset*)

Appendix N3 Reproducibility and Repeatability of Assessments of Class Characteristics

Table S35 summarizes the reproducibility of participants' assessments of design, mold, size, and wear, and shows how disagreements on these assessments are associated with disagreements on conclusions. For example (fifth row) considering all pairwise combinations of trials, participants differed in their assessment of design, mold, size, or wear on 67% of pairs of trials; if limited to pairs of trials in which the participants reached different conclusions, that rises to 78%; if limited to pairs of trials in which the participants differed by three or more conclusion categories, that rises to 91%. Participants usually agreed with each other on assessments of design, but often disagreed regarding mold, size, or wear. Participants often disagreed with each other regarding difficulty, but usually did not differ by more than one difficulty category.

	% of pairs of trials			
	All	*Same conclusion*	*Different conclusion*	*Conclusion delta 3 or more*
Disagreed on Design	12%	8%	14%	10%
Disagreed on Mold	37%	30%	42%	48%
Disagreed on Size	30%	19%	39%	48%
Disagreed on Wear	46%	32%	58%	76%
Disagreed on Design, Mold, Size, or Wear	67%	53%	78%	91%
Disagreed on Difficulty	59%	57%	60%	60%
Disagreed on Difficulty (>±1 category)	13%	10%	14%	14%

Table S35. Reproducibility of assessments among participants. Rows indicate the extent to which pairwise combinations of responses from the same QKsets did not have the same assessments. Columns indicate all pairs of trials, or pairs of trials that reached the same/different conclusions, or pairs of trials in which the conclusions differed by 3 categories or more (see *Appendix H2* for explanation of delta). (*Reproducibility Dataset*)

Table S36 summarizes the repeatability of these assessments. For example, comparing first vs. second responses on repeated trials, participants changed their assessments of design on 5% of trials, but that rises to 14% when participants changed their conclusions three categories or more. Examiners often changed assessments of Design, Mold, Size, or Wear (49% of repeated

trials; 66% of trials resulting in changed conclusions). Examiners also often changed assessments of difficulty (48%), but usually did not differ by more than one difficulty category.

	% of repeated QKsets			
	All	*Same conclusion*	*Changed conclusions*	*Conclusion delta 3 or more*
# distinct QKsets	578	347	231	80
Changed Design	5%	3%	7%	14%
Changed Mold	21%	16%	29%	41%
Changed Size	18%	11%	28%	40%
Changed Wear	34%	23%	50%	75%
Changed Design, Mold, Size, or Wear	49%	38%	66%	86%
Changed Difficulty	48%	46%	51%	46%
Changed Difficulty (>±1 category)	6%	5%	9%	13%

Table S36. Repeatability of assessments on QKsets that were assigned twice to the same participants. Columns indicate all repeated QKsets, repeated QKsets resulting in changed conclusions, and repeated QKsets resulting in conclusions changed 3 or more categories (e.g. ID to LimitedAssn is a 3-category delta). Percentages indicate the portion of repeated QKsets that changed a given assessment (e.g. examiner assessments of wear changed in 33.6% of all repeats, in 49.8% of repeats that resulted in changed conclusions, and in 75.0% of repeats that resulted in changed conclusions that differed by 3+ categories).

Appendix O Minor Results

This appendix reports a variety of factors that generally had relatively minor results when compared to those reported in the previous appendices.

Appendix O1 Limitations

As part of the comparison process, participants were asked to indicate any limitations associated with the QKset: "Please indicate any limitations that kept you from making a more definitive conclusion OR that were a notable source of difficulty in making the comparison. Check all that apply. Leave blank if not applicable."

Table S37 summarizes the limitations selected by participants. Overall, 67% of trials listed at least one limitation. Only 32% of trials with definitive conclusions indicated limitations, compared to 85% of probable conclusions, 91% of class associations, and 97% of neutral responses.

The quality/clarity of the questioned impression was the most-used limitation (cited in 45% of trials), and had the most striking association with conclusions: note that most class associations and neutral responses indicated this as a limitation. Of the 162 distinct Q images, all but six indicated quality/clarity was a limitation: 64 of the Q images had more than 50% of responses indicate quality limitations, and 13 had more than 90% of responses indicate quality limitations. Most of the other types of limitations showed trends that were similar to quality/clarity.

One notable association between limitations and conclusions was the relation of *HighAssn* to RAC limitations: 45% indicated "Insufficient number of corresponding RACs" and 47% indicated "Lack of clarity of RACs"; 68% of *HighAssns* indicated one or both RAC limitations.

	Percent of trials								
	All	*ID*	*HighAssn*	*Assn*	*Limited Assn*	*Inc*	*Not Suitable*	*NonAssn*	*Excl*
Any limitations	67%	27%	86%	88%	96%	95%	100%	84%	34%
Quality/clarity of the questioned impression	45%	9%	41%	58%	79%	72%	90%	54%	20%
Insufficient quantity/area of outsole reproduced in the questioned impression	26%	5%	25%	25%	56%	55%	77%	37%	9%
Distortion/movement in the questioned impression	17%	3%	8%	15%	38%	39%	48%	27%	8%
Background/substrate interference in the questioned impression	34%	11%	30%	47%	51%	42%	60%	46%	18%
Images/photographs of the questioned impression	3%	2%	4%	3%	4%	5%	9%	6%	2%
Images/photographs of the outsole of the known item of footwear	2%	5%	7%	2%	1%	4%	0%	3%	1%
Images/transparencies of the test impressions from the known item of footwear	2%	2%	5%	2%	2%	5%	1%	4%	2%
Insufficient number of corresponding RACs	14%	0%	45%	19%	20%	14%	11%	14%	3%
Lack of clarity of RACs in the questioned impression	20%	5%	47%	33%	27%	23%	11%	27%	4%

Table S37. Summary of limitations. (*Baseline Dataset*)

Fig S27 shows that the number of limitations was strongly associated with conclusions: if a participant listed no limitations, 82% of responses were definitive, and the proportion of non-definitive conclusions increased with the number of limitations. Error

rates were not notably associated with number of limitations: the rates of incorrect conclusions echoed the rates of correct conclusions.

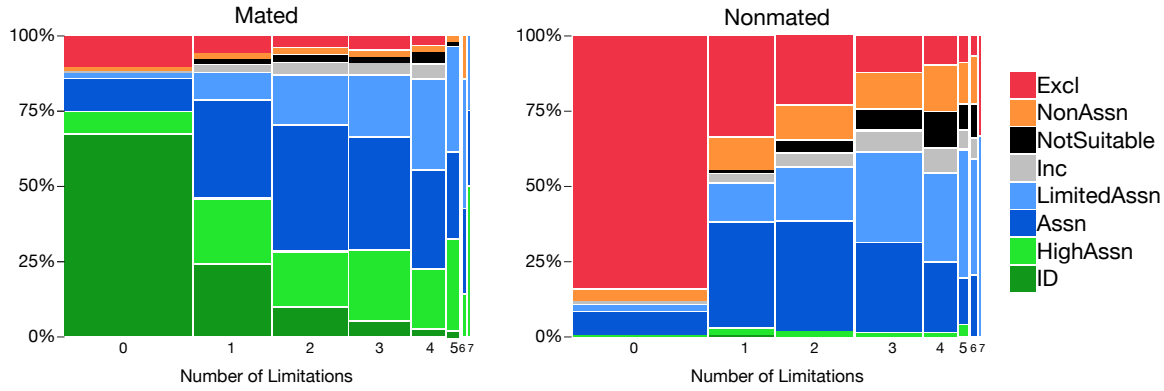

Fig S28. Conclusions by number of limitations. (*Baseline Dataset*)

Appendix O2 Effects of Collection Attributes for Questioned Impressions

Deposition: We had four types of deposition in the QKsets included in the study: walking (76% of QKsets), jumping (13%), kicking (7%), and running (4%). These were notably associated with quality: mean quality was notably different for walking (16.1, C), jumping (14.8, D), running (13.8, D), and kicking (11.1, F). This is reflected in the accuracy rates: TPR for walking was 32%, jumping 31%, running 5%, and kicking 5%. Examiners almost always excluded nonmated QKsets in which the Q and K were of different make/model, except when the Q was from a kick: TNR was 98% when the deposition was not a kick, but 35% for kicks; this is to be expected given that all kicked QKsets of different make/model were quality F.

Substrate: The types of substrates used are listed (with counts) in *Appendix C4.2*. The only notable result with respect to substrate was that the highest FNRs were for one QKset on cloth (QK213) and one QKset on plastic sheeting (QK083), as discussed in in Section 4.3 (main paper) and *Appendix E3*. This study did not have enough samples to make a clear statement regarding such substrates, but future studies may consider evaluating the effects of such malleable substrates in examinations.

Matrix: The Q impressions were collected using five types of matrices: residue (33% of QKsets), soil or dust (24%), blood (13%), mud (4%), and black powder (2%, used for the unused Q impressions). Matrix was notably associated with quality: mean quality for residue was 17.1 (B), soil was 15.5 (C), blood was 13.5 (D), and mud was 11.8 (F). Even when controlling for quality, the conclusions on mud trials had notably lower rates of definitive and probable conclusions than the other matrices.

Overlapping impressions: Of the 269 QKsets, 69 of the Qs included overlapping/superimposed impressions. These were not found to be associated with erroneous conclusions: in fact, the overlapping Qs were associated with higher rates of accuracy (higher TNR and TPR) than non-overlapping Qs. This unintuitive result may be explained in part because the overlapping Qs were of somewhat higher quality than the non-overlapping Qs (mean quality 16.2 vs 15.2): when the study team selected the superimpositions to be included in the study, we only included Qs from different types of footwear (e.g. boot vs running shoe) so that it would be clear which Q was intended for comparison, apparently in the process including more high-quality Qs.

Partial Qs: Of the 269 QKsets, 141 of the Qs were partial (as opposed to full heel-toe). We found no notable association between the partial vs. full extent of the Qs with conclusions or accuracy rates.

Appendix O3 RACs

During comparison, participants were asked "If you observed any randomly acquired characteristics (RACs) that CORRESPOND between the known item of footwear and the questioned impression, click here to mark them. Only mark RACs if they are present in BOTH the questioned impression AND images of the known." The resulting summary counts (Table S38) are generally as might be expected: *IDs* always have RACs marked, as do most *HighAssns*. The few RACs marked in *Excls* and *NonAssns* may have been the result of participants changing their minds, or not following instructions.

		Trials	*Trials with RACs marked*	*Number of RACs marked*			
				Min	*Median*	*Mean*	*Max*
Mated	ID	725	725	1	6	7.6	50
	HighAssn	410	368	0	2	2.7	34
	Assn	682	74	0	0	0.2	3
	LimitedAssn	312	9	0	0	0.1	8
	Inc	60	5	0	0	0.2	3
	NonAssn	43	1	0	0	0.0	1
	Excl	146	1	0	0	0.0	5
Nonmated	ID	9	9	3	16	14.0	28
	HighAssn	50	36	0	1.5	3.1	18
	Assn	852	19	0	0	0.0	4
	LimitedAssn	575	8	0	0	0.0	1
	Inc	137	2	0	0	0.1	12
	NonAssn	346	1	0	0	0.0	2
	Excl	1515	1	0	0	0.0	1

Table S38. Number of RACs marked by conclusion (*Baseline Dataset*).

Appendix O4 Use of Printed Materials

In Comparison question #10 participants were asked "Did you use the printed photographs/transparencies in making this comparison? Please indicate if you used the printed photographs/transparencies provided in the envelope at all when conducting your comparison in this set."

Participants used the printed materials in 5,617 of the 6,032 in the *Baseline Dataset*:

- 46 participants ***always*** used the printed photos (31 of whom completed all 100 comparisons)
- 70 participants used the photos on all but 1-5 QKsets.
- Only 3 participants never used the printed photos (none of whom completed all 100 comparisons)

There was no clear association between QKset and photo usage: photos were used almost all the time for all QKsets.

Use of printed photos had no notable association with accuracy of results.

Appendix O5 Typicality and Overall Difficulty of the Study Samples

In Comparison question #12 participants were asked "Was the questioned impression in this comparison set typical of impressions encountered by you in operational casework?"

- Yes: 84% of trials ("typical" in the discussion below)
- Yes, but it is considered unusual and encountered infrequently (11% of trials)
- No: 6% of trials ("not typical")

Most QKsets were considered typical:

- 18 of the 269 QKsets were rated typical by all participants
- All but 2 QKsets were rated typical by a majority of participants (and those were rated typical by 48% of participants)

The individual trials flagged as not typical did not have an effect on error rates, but they did have a higher proportion of *NotSuitable* and *Limited Association* responses than for trials flagged as typical.

Flagging trials as typical or not typical was not associated with error: very high and very low error rates were associated with participants who flagged many trials as not typical, and with participants who flagged every trial as typical.

In the post-test survey (which was completed by 67 participants), participants indicated if the comparisons in this study were

- much easier than operational casework (2/67, or 3%)
- easier than operational casework (6/67, 9%)
- comparable to operational casework (52/67, 78%)
- harder than operational casework (7/67, 10%)
- much harder than operational casework (0)

There was no association between participants' error rates and their assessments of overall difficulty: participants with high error rates indicated a range of responses from much easier through harder than operational casework.

Appendix O6 Orientation

During comparison, participants were asked "Rotate the questioned impression until it is oriented with the toe pointing up or select one of the options below (A. Questioned impression is already oriented with the toe pointing up; B. Unsure)."

None of the results with respect to orientation were of particular note.

Of the 162 distinct Qs, the majority of responses indicated:

- 146 were upright (majority replied "already oriented with the toe pointing up" or 0° ±22.5°) — 131 were unanimously assessed as upright;
- 8 were upside down (majority indicated 180° ±22.5°);
- 4 were sideways (majority indicated other angles);
- 3 were undefined (majority replied "unsure");
- 1 had no majority response.

Every Q received at least a few upright responses, even when a supermajority of responses indicated the image was upside down or sideways. We assume some of these may have been administrative errors or lack of attention to detail. However, almost all participants had at least some of these anomalous responses, no participants had more than 9% of responses as such anomalies, and these anomalies were not associated with high participant error rates.

Associations of participants' assessments of orientation with their responses:

- Qs assessed as undefined disproportionately resulted in *NotSuitable* responses.
- Qs assessed as upright had higher rates of definitive conclusions and lower rates of class associations than Qs assessed as either sideways or upside down.

Appendix P Participant Background Questionnaire and Post-test Survey

Appendix P1 Participant Background Questionnaire Results

A background questionnaire was required of all participants prior to starting the study. All 84 participants responded to the background questionnaire. All percentages are based on 84 participants. Percentages may not add to 100% due to rounding.

	Count	*Percent*
1. How old are you?		
Under 30 years	3	4%
30-39 years	25	30%
40-49 years	36	43%
50-59 years	16	19%
60+ years	4	5%
2. Highest level of education completed.		
High School Diploma (or equivalent)	3	4%
Associate Degree	7	8%
Bachelor's Degree	41	49%
Master's Degree	28	33%
Doctoral Degree	5	6%
2a. Disciplines		
Criminology/Criminal Justice	11	13%
Forensic Science	50	60%
Natural Science (Biology, Chemistry, Physics)	45	54%
Social Science (Political Science, Psychology, Sociology)	4	5%
Other	7	8%
Question 2a was not asked of participants for whom high school was the highest level of education.		
3. After completing FFE training, provide the number of years of experience you have as an FFE.		
<1 year	5	6%
1-4 years	20	24%
5-10 years	25	30%
11-15 years	16	19%
16+ years	18	21%
4. Select the statement which best describes the frequency with which you perform footwear examinations?		
I conduct footwear examinations daily	0	0%
I conduct footwear examinations a few times a week	8	10%
I conduct footwear examinations a few times a month.	28	33%
I conduct footwear examinations a few times a year.	48	57%
I no longer conduct footwear examinations, but I previously did.	0	0%
4a. What percentage of your typical work week is spent conducting footwear examinations?		
1-25%	2	2%
26-50%	5	6%
51-75%	0	0%
76-100%	1	1%
Question 4a was only asked of participants who indicated in question 4 that they conduct footwear examinations daily or weekly.		
5. Have you testified in court as an expert in footwear examination?		
Yes	51	61%
No	33	39%

	Count	*Percent*
5a. How many times?		
1-9	37	44%
10-19	8	10%
20+	6	7%
Question 5a was only asked of participants who indicated in question 5 that they have testified in court.		
6. What other forensic examinations are you currently qualified to perform (or have been qualified to perform in the past) and have performed in operational casework? (Check all that apply. Leave blank if not applicable)		
Chemistry	15	18%
Crime scene processing	53	63%
DNA	2	2%
Fingerprints	42	50%
Firearms	5	6%
Handwriting	2	2%
Questioned documents (excluding handwriting)	3	4%
Tire impressions	62	74%
Toolmarks	10	12%
Toxicology	4	5%
Trace evidence	35	42%
Participants indicated 0-6 forensic disciplines (mean 2.8, median 3); only one participant indicated none.		
Combinations with ten or more participants:		
o *Crime scene processing OR Fingerprints OR Tire impressions*	80	95%
o *Crime scene processing AND Tire impressions*	45	54%
o *Tire impressions AND Trace evidence*	32	38%
o *Crime scene processing AND Fingerprints*	27	32%
o *Crime scene processing AND Fingerprints AND Tire impressions*	23	27%
o *Chemistry AND Tire impressions*	13	15%
7. Select the following statement which best describes your most comprehensive FFE training.		
I completed a formal program of instruction for 1 year or more	26	31%
I completed a formal program of instruction for 6-12 months	31	37%
I received informal, on-the-job training.	7	8%
I attended/completed courses and/or workshops.	20	24%
I did not receive training.	0	0%
7a. Who was primarily responsible for providing your <u>formal</u> program of instruction in footwear examination?		
Current employer	40	48%
Past employer	3	4%
An agency or company other than your current or past employer	14	17%
Question 7a was only asked of the participants who indicated they completed a formal program of instruction in Question 7.		
7aa. Select the category which best describes the agency or company responsible for providing your <u>formal</u> program of instruction in footwear examination?		
U.S. local agency	1	1%
U.S. state agency	4	5%
U.S. federal agency	3	4%
U.S. private company	5	6%
International government agency	4	5%
International private company	0	0%
Question 7aa was only askedof the participants who indicated their training was not provided by their current employer in Question 7a.		
8. Are you a certified footwear examiner (by an external certifying body)?		
Yes, currently	23	27%
No, but I was previously	3	4%
No, and I never was.	58	69%
8a. Provide the certifying body for any current *or* past footwear certifications. Check *all* that apply.		
International Association for Identification (IAI)	19	23%
Canadian Identification Society (CIS)	1	1%
Other	4	5%
Question 8a was only asked of the 23 participants who replied Yes to Question 8.		
One participant indicated IAI and included "QPS certified course recommended by IAI" under "Other." No other participants indicated more than one category.		
The remaining 3 responses entered under "Other" were AAFSAB, Chinese Association for Identification, and College of Policing.		
9. When did you last conduct a proficiency test in footwear examination?		
Within the past year	56	67%
Within the past 2 years	11	13%
Within the past 5 years	7	8%
More than 5 years	2	2%
Never	8	10%

	Count	*Percent*
9a. Who prepared your most recent proficiency test?		
Collaborative Testing Services (CTS)	49	58%
Employer (internal)	10	12%
Forensic Assurance (FA)	5	6%
Forensic ITC services	0	0%
Ron Smith and Associates (RS&A)	6	7%
Other	6	7%

Question 9a was not asked of participants who indicated that they had never conducted a proficiency test in Question 9.

9b. Has it ever been brought to your attention that you failed a proficiency test in footwear examination?		
Yes	0	0%
No	76	90%
No response	8	10%

Question 9b was only asked was not asked of participants who indicated that they had never conducted a proficiency test in Question 9.

10. After being qualified by your current employer or a previous employer to conduct footwear examinations, was it ever brought to your attention that you made an erroneous identification after the associated laboratory report was issued?		
Yes	0	0%
No	84	100%

11. Using the sliders below, provide an estimate as to the frequency of the types of known footwear received in casework for comparison over the last 12 months. (The sum of all responses shall equal 100%.)

	Athletic	*Boot*	*Other*
Minimum	30	0	0
Quartile 1	60	10	0
Average	72	20	8
Median	70	20	10
Quartile 3	80	30	10
Maximum	100	60	40

Nine participants indicated 100% athletic.

12. Provide an estimate as to the frequency of the conclusions you provided during comparison case work over the last 12 months.

	Frequently	*Infrequently*	*Never*
Identification	7 (8%)	53 (63%)	24 (29%)
High degree of association (or probably made)	23 (27%)	46 (55%)	15 (18%)
Association of class characteristics (or could have made)	77 (92%)	5 (6%)	2 (2%)
Limited association of class characteristics (or inconclusive)	52 (62%)	21 (25%)	11 (13%)
Indications of non-association (or probably did not make)	11 (13%)	31 (37%)	42 (50%)
Exclusion (or elimination)	52 (62%)	29 (35%)	3 (4%)
Lacks sufficient detail (or unsuitable)	34 (40%)	40 (48%)	10 (12%)

Notable combinations:

- *One participant replied "frequently" to all of these conclusions.*
- *Five participants frequently ID and frequently exclude.*
- *15 participants never ID but frequently exclude.*

13. Using the sliders below, provide an estimate as to the frequency of two-dimensional (2D) versus three-dimensional (3D) impressions encountered in casework over the last 12 months. (The sum of all responses shall equal 100.)

	2D	*3D*
Minimum	0	0
Quartile 1	60	10
Average	74	26
Median	80	20
Quartile 3	90	40
Maximum	100	100

Two participants indicated 100% 3D; 16 participants indicated 100% 2D.

	Count	*Percent*

14. Provide an estimate as to the frequency of matrices associated with 2D impressions encountered in casework over the last 12 months.

	Frequently	*Infrequently*	*Never*
Blood	36 (43%)	32 (38%)	16 (19%)
Dirt/soil	64 (76%)	19 (23%)	1 (1%)
Dust	55 (65%)	25 (30%)	4 (5%)
Grease/oil	3 (4%)	40 (48%)	41 (49%)
Paint	0 (0%)	11 (13%)	73 (87%)

15. Provide an estimate as to the frequency of substrates associated with impressions (to include both 2D and 3D impressions) encountered in casework over the last 12 months.

	Frequently	*Infrequently*	*Never*
Asphalt/concrete	14 (17%)	41 (49%)	29 (35%)
Cardboard/paper	26 (31%)	42 (50%)	16 (19%)
Carpet	6 (7%)	37 (44%)	41 (49%)
Clothing	2 (2%)	43 (51%)	39 (46%)
Countertop/tabletop	34 (40%)	28 (33%)	22 (26%)
Door	36 (43%)	39 (46%)	9 (11%)
Dirt/soil	62 (74%)	15 (18%)	7 (8%)
Hardwood/laminate flooring	46 (55%)	30 (36%)	8 (10%)
Skin	1 (1%)	25 (30%)	58 (69%)
Snow	27 (32%)	13 (15%)	44 (52%)
Tile	41 (49%)	29 (35%)	14 (17%)
Wood	17 (20%)	34 (40%)	33 (39%)

16. How do you conduct footwear examinations?		
Primarily digital (on computer)	10	12%
Primarily physical (printed photographs and physical test impressions)	32	38%
Combination of digital and physical	42	50%
17. Select the category which best describes your current employer.		
U.S. local agency	26	31%
U.S. state agency	29	35%
U.S. federal agency	2	2%
U.S. private company	1	1%
International government agency	25	30%
International private company	0	0%
Academic Institution	1	1%
Currently Unemployed or retired	0	0%
17a. Is your current employer accredited in a category of testing that includes footwear examination?		
Yes	60	71%
No	16	19%
Unsure	8	10%
17b. By whom is your current employer accredited?		
A2LA	0	0%
ANSI-ASQ National Accreditation Board (ANAB) or ASCLD/LAB	53	63%
Unsure	7	8%
Question 17b was only asked of participants who responded "yes" to Question 17a.		
17c. Under which standard is your current employer accredited?		
Both ISO 17020 and ISO 17025	6	7%
ISO 17020	2	2%
ISO 17025	48	57%
Unsure	4	5%
Question 17c was only asked of participants who responded "yes" to Question 17a.		
18. Select the statement which best describes the FFEs working at your current employer.		
I'm the only FFE currently working.	11	13%
There are other FFEs in addition to me currently working	73	87%
18a. How many additional FFEs are currently working?		
1	21	25%
2	18	21%
3	4	5%
4	6	7%
5+	24	29%
Question 18a was only asked of the participants who indicated there are other FFEs at their employer (Question 18).		
Of the 24 participants who replied "5+", 13 work for an International government agency, and 9 work for a U.S. state agency.		
19. What conclusions do you use to report comparison findings in casework?		
SWGTREAD 2013 range of conclusions (1. Identification, 2. High degree of association, 3. Association of class characteristics, 4. Limited association of class characteristics, 5. Indications of non-association, 6. Exclusion, 7. Lacks sufficient detail)	56	67%
SWGTREAD 2006 range of conclusions (1. Identification, 2. Probably made, 3. Could have made, 4. Inconclusive, 5. Probably did not make, 6. Elimination, 7. Unsuitable)	11	13%
Other	17	20%

	Count	Percent
A total of 21 participants (25%) use "inconclusive" as a category (11 SWGTREAD 2003 responses and 10 of the "Other" responses).		
The "other" responses indicated the following number of levels in the conclusion scale (not counting "not suitable"; both SWGTREAD scales have 6 levels):		
o *4 levels*	5	6%
o *5 levels*	4	5%
o *6 levels*	2	2%
o *7 levels*	2	2%
o *9 levels*	1	1%
o *Ambiguous*	3	4%
Responses received explaining "other", sorted by number of levels in conclusion scale (omitting one blank response):		
o *inconclusive, excluded, limited support, moderate support, moderately strong support, strong support, very strong support, extremely strong support, conclusive*		
o *Category 1: Source Identity/Source Attribution, Category 2A: Association with distinct characteristics, Category 2B: Association with conventional characteristics, Category 2C: Association with limitations, Category 3: Inconclusive, Category 4: Dissimilar/Non-Association, Category 5: Elimination/Exclusion; will also report unsuitable/lack sufficient detail*		
o *Categories: 1. Source Identity/Source Attribution, 2A. Association with distinct characteristics, 2B. Association with conventional characteristics, 2C. Association with limitations, 3. Inconclusive, 4. Dissimilar/Non-Association, 5. Elimination/Exclusion; also will report out when unsuitable/lack sufficient detail*		
o *optional A or B.*		
o *1. Identification 2. Very probably made 3. Probably made 4. Possibly made 5. Inconclusive 6. Exclusion*		
o *inconclusive, exclusion, Association of class characteristics, Higher degree of association, Identification*		
o *Identification, Higher Degree of Association, Association of Class Characteristics, Exclusion, Inconclusive*		
o *Identification, High Degree of Association, Could Have Made, Limited Association, Exclusion*		
o *5 categories of assocation: 1. Identification, 2. Association which has three subgroups similar to 2, 3, and 4 from SWGTREAD 2013. 3. Inconclusive, 4. Dissimilar/Non-Association, 5. Exclusion, and unsuitable for comparison*		
o *Unsuitable, Elimination, Could have made, Identification*		
o *identification, positive class characteristics, inconclusive, elimination, unsuitable*		
o *1. Identification, 2. Could have made, 3. No determination could be made (Inconclusive), 4. Could not have made (Exclusion), 5. Unsuitable*		
o *1. Identification 2. Could have made 3. Inconclusive 4. Elimination 5. Not suitable*		
o *1, 2, 3, 6, 7*		
o *Use SWGTREAD 2013 range AND traditional "could have" etc.*		
o *RCMP which is †a slightly abbreviated version of swgtread*		
20. Do you use likelihood ratios or other probability measures in reporting conclusions?		
Yes: we use likelihood ratios or other probability measures in addition to (or in support of) the conclusion scale	3	4%
Yes: we use likelihood ratios or other probability measures instead of a conclusion scale	0	0%
No	81	96%
21. Does your current employer permit the comparison of two questioned footwear impressions to determine if they were made by a common unknown source (in cases when footwear is not recovered)?		
Yes	53	63%
No	31	37%
21a. How often do you perform this type of comparison?		
Frequently	5	6%
Infrequently	30	36%
Never	18	21%
Question 21a was only asked of participants who replied "yes" to Question 21.		
The study design included a followup question ("21b. Does your current employer permit an FFE to effect an identification conclusion when conducting this type of comparison?") but that question was inadvertently not implemented in the software.		
22. Does your current employer require the use of a second FFE to ensure that the data and documentation support the primary FFE's conclusions?		
Yes	82	98%
No	2	2%
22a. Does your current employer require blind verification of conclusions? (Blind verification is performed by a second FFE who does not know the primary FFE's conclusions.)		
Yes	26	31%
No	56	67%
Question 22a was only asked of participants who replied "yes" to Question 22.		

Appendix P2 Post-Test Survey

The post-test survey was completed by 67 participants, 53 of whom completed the study. Percentages are based on 67 participants. Percentages may not add to 100% due to rounding.

	Count	Percent
1. Rate your overall assessment of quality of the known outsole images (KA through KE) provided in the comparison sets in this study.		
Exceptional	49	73%
Acceptable	18	27%

Unacceptable	0	0%
2. Rate your overall assessment of quality of the known test impressions (KF through KG) provided in the comparison sets in this study.		
Exceptional	46	69%
Acceptable	21	31%
Unacceptable	0	0%
3. Overall, did your inability to examine the original (physical) known items of footwear prevent you from making more definitive conclusions?		
Yes	7	10%
Sometimes	38	57%
No	22	33%
4. Overall, how did the difficulty of the comparisons you performed in this study correspond to the comparisons you've performed in operational casework?		
The comparisons in this study were much easier than operational casework.	2	3%
The comparisons in this study were easier than operational casework.	6	9%
The comparisons in this study were comparable to operational casework.	52	78%
The comparisons in this study were harder than operational casework.	7	10%
The comparisons in this study were much harder than operational casework.	0	0%
5. Select the statement below which best describes the process you used to examine/analyze the items of evidence prior to conducting the comparisons for the majority of the sets in this study.		
I examined/analyzed the questioned impression and the known item of footwear together and made no efforts to prioritize the examination/analysis of one item over the other.	4	6%
I thoroughly examined/analyzed the questioned impression (including marking potential RACs) prior to examining/analyzing the known item of footwear.	21	31%
I examined/analyzed the questioned impression (without marking potential RACs) prior to examining/analyzing the known item of footwear.	39	58%
I thoroughly examined/analyzed the known item of footwear (including marking RACs) prior to examining/analyzing the questioned impression.	0	0%
I examined/analyzed the known item of footwear (without marking RACs) prior to examining/analyzing the questioned impression.	0	0%
The process I used to examine/analyze the items of evidence varied from set to set.	3	4%

6. Comments (optional)

38 participants made comments in the post-test survey.

- At some point during this exercise I came to realize how much more an examiner can see on screen than when only using hard copies. I've begun to use on-screen images in all my case work.
- Different styles and makes of shoes should have been used, there was not enough variety. More running shoes should have represented, most offenders are wearing these types of shoes instead of work boots
- Excellent job! I wish PTs were made this well.
- For some comparisons, a more definitive conclusion would have been possible with information regarding the time between the offence and when the suspect footwear was seized. For example, some differences observed could be explained if it was one week or one month but not if it was the same day. Thank you.
- Great job with this study! The only thing is that the transparencies need to be cut down (length wise) so that they will fit into the manila envelopes better.
- Great quality of images etc. I found that I struggled most with decisions in which the impressions shared class characteristics, but the questioned impressions lacked detail - I sometimes couldn't decide between association of class, limited association of class (even though they may seem to be the same size), or inconclusive. Sometimes it can be hard to tell what is/are actually wear/RACs in a questioned impression - or a mold variation in the known - and this can sometimes make the decision difficult (in case work and this study).
- Having the actual item of footwear potentially could change some degrees of association where there were minor size or wear discrepancies. Having the ability to make my own set of standards would have been helpful.
- I compared the impressions in a stepwise fashion by starting with the Q general design, then the K general design. If similar, then proceeding to the Q size vs the K. If similar, then the Q wear and RACs vs the K exemplar and verified on the outsole. The study was very thorough. At times, working without the ability to obtain a second opinion was stressful. I would have asked for co-workers input on several of these sets as I realize I may have focused on one aspect of the comparison and missed a less obvious aspect, perhaps sliding up or down on the conclusion scale a bit as a result.
- I felt the time frame of 2 weeks for wear and for RACs to have been produced to be difficult to weigh in for EX purposes. I also felt the lack of like to like test impressions to inhibit some determinations of giving weight in either direction for the comparisons.
- I found as I worked through the packages I understood better the "reverse" terms attached to images. For example the questioned original stated as much, but the enhanced image of the questioned original (maintaining orientation) said "reversed".. and leading in the beginning confusion on my part because in some cases i preferred the enhanced and its title would mix me up when comparing it to the known. Hope that makes sense. Also i found that the image used to mark the RAC's were not always the best image selected to show their locations. ver all great experience, well put together packages and examples. Thank you for the experience.
- I have all my footwear examinations re-examined (verified) by another examiner prior to reporting a conclusion. This is a vital step in the process that is not part of this study.
- I liked the interface for the study. It was easy to enter in my results. This would be great to have for casework if there were a way to document possible RACs prior to exam.
- I really enjoyed this study and look forward to seeing the end result.
- I thought this study was well put-together. I am looking forward to reading about the results. Thank you!
- I was impressed at the quality of all the digital images. These comparisons were also a great learning tool as I could compare the differences that could be observed with the various documentation techniques. For example, a gel lift vs. black powdered photograph, etc. I think in a couple cases, I was more conservative in my conclusion because I didn't have a narrower time frame between the collection of the questioned impression and the collection of the

known footwear. From the construction of each case, the website, the email support, and everything else that went into this endeavor, I say congratulations on a job very well done - it was a pleasure to participate in this!

- I would have liked to receive at least 2 transparencies for each method to see reproducibility of any randomly acquired characteristics
- Ideally, a time frame reference would have been helpful. Knowing the time between the 'offence' date and documenting the Q, and the date of the collection of the K shoes would assist in addressing some of the difference observed in wear and RAC's. In casework, it is also beneficial at times to complete further test impressions when some differences can't be immediately resolved. AT times, test impressions are also made on similar materials to those found at the scene in order to address any observed differences.
- In casework, one would verify RACs present by examining the actual shoe and seeing that it is reproducible in the test impressions. In this study we did not have the known shoes, and therefore, it was treated like a proficiency test where the RACs were "verified" through the photographs and test impressions.
- in my normal casework, I trace all the visible detail in the questioned impression before I do any comparisons, and I do not use transparency impressions of the test impressions.in this study I only traced the more difficult impressions or ones with visible RACs.
- It differed from real cases in that there were many more RACs present in the questioned impressions than encountered in casework.
- It is my opinion that a wider variety of outsole patterns should have been used in the research. The majority of the 100 impressions I examined came from 2 outsole designs.
- It was hard, in most cases, to mark RACs on the known shoe as it was sometimes difficult to see them. Also, it was challenging to not consult other examiners, as I would in normal casework, especially on difficult comparisons. Also, not being able to accurately evaluate when the knowns were collected limited my ability to evaluate differences I observed between RACs and wear.
- It would have been beneficial to have additional test impressions of various amounts of pressure
- It's a good experience and I learned a lot from this study too.
- Not only did the lack of physical examination of the shoes prevent/affect some conclusions but in some sets also the inability to examine the impression itself and the lack of test impressions created with a similar substrate to the questioned impression. Also, while I have selected 'exceptional' with regards to the outsole photos - some cases needed more photos there should always be photos of the known item with different side lighting.
- Outsole exemplars would have been preferred on clear transparency film vs the somewhat opaque film that was received for each comparison set. Also, more athletic footwear comparisions would have been desired and less lugged boot-type.
- Overall, I think the materials provided were of good quality. There were times when I wished I had additional test impressions and/or the shoes. There were also times when I didn't quite feel like the responses regarding wear correspondence covered "some general wear correspondence", for example.
- Regarding my answer above, 5.C, I did not mark "potential" RACs, but I did mentally note the "evident potential" RACs. This is because every questioned impression has hundreds of "potential" RACs, therefore noting/marking them all would be unrealistic. Regarding question 3 above, not only would the original footwear assist in the comparisons, but also an image of the footwear when new, or an actual new shoe/boot from the same mold (or at least the same model/size). This would help determine/evaluate RACs and wear. I have a comment regarding question 3 from the study itself, 'Do the questioned impression and the known item of footwear correspond in outsole design?'. I take the term, 'correspond in outsole design', to mean that they correspond in design between every comparable class characteristic, and dimensions (not just similar). Therefore, it would be impossible to answer with 3B, 'the opposite foot'. Also, when I selected 2A (identification) I was perplexed as to why question 3, 4 and 5 were still being asked. And again, when I selected 2G (exclusion) I thought question 3 was redundant. In actual case work, for myself a written footwear comparison report is part of my overall A.C.E. process, to help focus, comprehend and deduce. Whereas the Black Box study does not include this step. However, I do appreciate not having to complete 100 comprehensive reports. I found the study to be a great educational and practical experience. The practice of conducting 100 comparisons (A.C.E.) is beneficial experience. Also, I now appreciate the immense enhancement ability of the GLscan. I also learned the value of increasing image quality from our currently allotted 2MB to the study's often 6MB, 8MB, 10MB or more. The study's test impressions were exceptional; particularly the 'hand-rolled' impressions, with which I am not familiar. Thank you for allowing me to participate in this study. Yours truly, Anonymous.
- Some of the questions were a little confusing. I would have loved to explain more but understood that you needed a set question/answer for the results. A contact number would have made things simpler as well. Great job with the whole project.
- Thank you for allowing me to participate in this study.
- Thank you for conducting this study. It was a great opportunity and I look forward to reading your findings.
- Thank you for providing this study. I believe the results will be truly fascinating and I would like to ensure I receive a copy of the results. It would be interesting if different jurisdictions have the same/similar results. I do, however, now loathe the sight of the EMS hiker and New Balance outsole. Fortunately I rarely see either in real life. :-)
- The acetates were too long to go in the envelope. I hope another study will included a copy of the unknown that can be marked prior to comparison, or would that be in a white box type study? Glad to participate. Things like this are what is needed.
- The extension of the due dates really helped me finish one more set.
- The study was a great experience and I hope that more can be done to further shoe/tire impression analysis.
- Very interesting study!
- Very interesting!!!! Thank's
- When marking the observed RACs I would have preferred to mark them on the impression or test impression. Sometimes the known that was selected did not show the tiny cut, due to the lighting that was used on that photo.