Document Title:	**Understanding the Expert Decision Making Process in Forensic Footwear Examinations: Accuracy, Decision Rules, Predictive Value and the Conditional Probability of an Outcome**
Author(s):	**Jacqueline A. Speir**
Document Number:	**304650**
Date Received:	**April 2022**
Award Number:	**2016-DN-BX-0152**

Final Technical Report: 2016-DN-BX-0152
Project Title: Understanding the Expert Decision Making Process in Forensic Footwear Examinations: Accuracy, Decision Rules, Predictive Value and the Conditional Probability of an Outcome

Principal Investigator: Jacqueline A. Speir, Associate Professor
West Virginia University, 208 Oglebay Hall, PO Box 6121, Morgantown, WV 26506
Jacqueline.Speir@mail.wvu.edu, 304.293.9233

Submitting Official: Katie Stores, Associate VP for Research Administration, Director of Office of Sponsored Programs; Katie.Stores@mail.wvu
Recipient Organization: West Virginia University Research Corporation, 866 Chestnut Ridge Road, PO Box 6845, Morgantown, WV 26506, **DUNS:** 191510239, **EIN:** 550665758

Award Period & Amount: 10/01/2016 - 09/30/2020; $213,735.00
Reporting Period End Date: 12/29/2020 - *Final Report*

Keywords: footwear, conclusion scales, accuracy, consensus, inter-rate reliability, predictive value, chi-square

Table of Contents

1. Executive Summary

Forensic footwear examination and interpretation is a complex and distributed activity influenced by a host of competing and evolving factors that vary as a function of case attributes and examiner experience. The entire pattern recognition process and ultimate conclusion drawn by the expert decision maker with regard to source is an amalgamation of several sources of variability that are not necessarily independent, nor linearly related. To date, there are few footwear reliability studies that report on the accuracy and reproducibility of conclusions drawn by examiners when evaluating the same case materials. In an effort to address this gap, the purpose of this study was to solicit responses and conclusions from footwear examiners in the United States in order to infer accuracy, reproducibility, predictive value, and decision rules. This was accomplished by preparing and distributing simulated case materials to a total of 115 footwear experts. Of these participants, 77 completed all analyses, resulting in a total of 840 usable conclusions. These conclusions were used to compute several numerical metrics, including consensus, inter-rater reliability, accuracy, predictive value and inferred decision rule support, coverage, strength and confidence; major results were four-fold.

First, reproducibility was evaluated as a function of three metrics, including the interquartile range (IQR), consensus and inter-rater reliability (IRR). The observed community agreement in conclusions via IQR was found to equal 85.6% $\pm$ 11.1% (median of 89.3% and a 90% confidence interval between 83.5% and 87.6%). Moreover, consensus ranged from a low of 0.5105 (for comparison 003Q versus 003K1), to a maximum of 0.9733 (for comparison 007Q versus 007K1), with a mean of 0.7821 $\pm$ 0.1422 and a median of 0.7743. Likewise, IRR, as measured using the Gwet AC_2 agreement coefficient, was found to be 0.7509 with a standard error of 0.0875 and a 90% confidence interval of 0.6070 to 0.8948. After benchmarking, this was found to equate with the verbal equivalent of '*substantial*' agreement.

Second, accuracy in conclusion was evaluated and found to equal 82.8% $\pm$ 11.9% (median of 85.7% and 90% confidence interval between 80.5% and 84.9%). Using the Scientific Working Group for Shoeprint and Tire Tread Evidence (SWGTREAD) 2013 seven-point conclusion standard [1], the data was further probed to determine correct (positive) predictive value (PV) for a mate-prevalence of 31.5%, with results indicating a PPV that varied between 94.5% for exclusions, 85.0% for identifications, and between 70.1% and 65.2% for limited associations and association of class, respectively (with all other conclusions producing PVs between these extremes). After data transformation based on ground truth (and therefore reduction to a three-point conclusion standard similar to that used in other forensic pattern evidence comparisons such as fingerprints), the case study material revealed a false positive rate of 0.48%, a false negative rate of 15.6%, a (correct) positive predictive value of 98.8% and a (correct) negative predictive value of 93.3%. When adjusted for the same mate-prevalence used in the 2011 FBI fingerprint study [2] (or 62%), the comparable footwear PPV is 99.7% (versus fingerprints at 99.8%) and NPV is 79.6% (versus fingerprints at 86.6%) [2].

Next, conclusions were evaluated using the chi-square test of independence to determine the degree to which accuracy varied as a function of both examiner and case related factors. For any significant results, an adjusted Pearson's residual post-hoc analysis with a Bonferroni correction was applied [3, 4]. Interestingly, results failed to detect any dependence between accuracy and the examiner attributes of education, certification status, frequency of continuing education/training annually, case load, nor familiarity with the SWGTREAD 2013 conclusion standard [1]. When case-attributes were evaluated, some dependencies were observed. More specifically, examiners were more accurate than expected when evaluating case 004 and test impression 007K1, and less accurate than expected when evaluating test impression 003K1. In addition, experts performed more accurate than expected when reporting exclusions, but less accurate than expected when reporting limited association and association of class characteristic conclusions. This fits with intuition since these latter two conclusion categories are the least restrictive decisions that can reached using the SWGTREAD 2013 scale [1], and they can reasonably be reached for both known mates and known non-mates.

Fourth, impression feature identification and annotation were evaluated using a customized graphical user interface (GUI). Based on reports, results indicate considerable variation in feature identification/annotation (as low as 66.5% agreement) despite much higher agreement in conclusions. This implies that examiners can come to the same conclusion, but the features they identify and annotate, and the weight applied to these features, can vary considerably. This was also supported by decision rule induction using the dominance-based rough set approach (DRSA). The induced rules were evaluated as a function of strength, support, coverage and confidence, and the highest conditional probability (the probability that a set of conditions will be reported given a specific conclusion/decision) was no more than 0.46. This further suggests that although consistency in outcomes/conclusions are apparent across examiners comparing the same simulated case materials, the reasoning reported to justify these conclusions is much more variable and requires greater examination in further white-box studies.

2. Overview

2.1 Major Goals & Objectives

This work proposed four major deliverables:

- To quantify the variability in forensic footwear expert decisions via accuracy and positive predictive value;
- To identify factors that affect footwear examination and conclusions;
- To evaluate the interaction between factors and expert decisions;
- To induce and evaluate decision rules and their associated quality as a function of strength, support, certainty and lift.

To date, the following results have been realized (with a subset disseminated in three peer-reviewed publications):

- Production of 7 forensic cases involving a total of 12 comparisons, wherein each case included 1-2 known exemplars (outsole images), 2 Handiprint test impressions per known, and 1 questioned impression;
- Successfully solicited and enrolled a total of 115 participants;
- Processed and digitized all background surveys describing participant demographics;
- Received results from 77 participants (67% of enrolled);
  - Of the "lost" participants, 1 included an erroneous submission of results (blank folder and unable to reach participant), 1 case-packet was returned by USPS due to an invalid address (unable to reach participant for correct address), 8 participants withdrew (deployed, promoted, change in interest in participation, etc.), 28 were delinquent, 27 received email reminders, extended due-dates, etc., but all remained unresponsive while 1 did not provide contact information so unable to reach).
- Created a customized graphical user interface (GUI) to collect results;
- Evaluated and processed the conclusion accuracy associated with all 77 submitted results (evaluation of 77 x 12 = 924 conclusion responses);
- Processed and evaluated rule-induction using the dominance-based rough set approach (DRSA).

2.2 Research Question

Comparing and interpreting forensic footwear evidence includes an assessment of several sources of variability, including but not limited to class and subclass features, the quantity, clarity and complexity of randomly acquired characteristics (RACs), the manner of deposition and recovery of impressions at the crime scene, the possible employment of enhancement techniques, and last but not least, the training and experience of the expert examiner.

As a consequence, a degree of variation in the conclusions reached by a group of comparative scientists examining the same evidence is to be expected. Moreover, this variation is likely to be a function of the conclusion scale used during the evaluation process. The purpose of this research was to report the degree of variation observed, and to investigate the possibility of dependence between conclusions and both case and examiner factors.

2.3 Research Design

One hundred and fifteen (115) forensic footwear examiners were recruited through a variety of media, including electronic solicitation, word-of-mouth, and in-person announcements during regional and international conferences. Enrolled participants completed a background survey providing information regarding their education, experiences, job capacity, certification status, as well as details concerning the nature and frequency of training, research, teaching and professional development activities. Over the span of 19 months (February 2017 - August 2018) results were collected from 77 examiners, resulting in a cumulative response rate of 67%, with each participant performing 12 comparisons and reporting a total of 924 individual conclusions.

2.4 Data Analysis

2.4.A Participants

Participants responded to a variety of questions designed to ascertain the demographics of the expert examiners. Self-reporting revealed that 7 of the 77 participants had either never performed a comparison and/or were still in training. Since this evaluation was meant to determine the accuracy and conformity in reporting for footwear examiners actively performing casework, the results from these 7 participants were excluded when creating summary statistics.

Results indicate that the majority of participants (83%) were actively working in a crime lab at the time of participation (Table 1). In addition, 36% of all participants had completed 11-50 comparisons when they agreed to participate in this study, while another 20% had completed 51-100 comparisons, and 23% had completed more than 100 comparisons (Table 2). Participants were also asked to report the frequency at which different types of activities

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

were performed using a Likert scale. It appears that few participants collect or develop impressions (presumably at scenes), but more frequently enhance, photograph and compare impressions (presumably in laboratories), which fits with anecdotal reports within the field (Table 3).

Footwear Examiner Status	Working in Lab	Consultant	In-Training	Retired	Supervised Casework
#	64	7	5	0	1
Cases Completed	0-50	51-100	101-150	151-200	>200
#	31	13	7	3	16

Table 1: Examiner self-reported casework experience. The first row describes n = 77 participant responses, while the second row describes n = 70 participant responses (excludes the 7 examiners that self-reported an absence of casework).

Database Searches	No Response	1-10	11-50	51-100	101-200	>200
#	3	25	21	7	3	1
Examine/Compare Impressions	No Response	1-10	11-50	51-100	101-200	>200
#	3	12	25	14	13	3

Table 2: Type of work performed by n = 70 participants, including database searches and comparison of impressions (note that 10 examiners reported that they had never performed a database search).

Frequency	Very Seldom	Seldom	Occasionally	Frequent	Very Frequent	N/A
Collect Evidence	21	14	11	3	5	16
Develop Impressions	20	15	7	4	3	21
Enhance Impressions	10	13	15	16	14	2
Photograph Evidence	6	7	14	12	28	3
Database Searches	15	10	12	11	10	12
Examine/Compare Impressions	2	4	16	19	29	0

Table 3: Frequency of activities performed as assessed using a Likert scale for n = 70 participants.

Since footwear examiners may be asked to perform comparisons on multiple types of evidence, and since experts can cross-train and/or move from one discipline to another throughout their careers, Table 4 reports additional overlap with crime scene processing, firearms/toolmarks, and fingerprint analysis, with a large percentage co-listing fingerprints as an area of current occupation.

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

Discipline	Arson/ Explosives	Fingerprints	Firearms/ Toolmarks	Questioned Documents	Trace	Crime Scene/ Bloodstain	Controlled Substances
Current	6	28	14	4	16	16	2
Past	5	2	3	3	13	32	14

Table 4: Examiner (n = 70) reports of current and past forensic activities (examiners were asked to report all that applied, so row totals can eclipse 70).

Table 5 reports the number of years of footwear experience and the number of years of total forensic experience for examiners that participated in this study; 53% of respondents had 8 or more years of experience in footwear, and the majority (86%) have been in the forensic field as a whole for more than 8 years. Table 6 reports the frequency of training in the last 5 years, as well as the types of training providers. Note that the majority of participants (97%) have attended one or more training sessions beyond laboratory specific activities. In addition to employment and professional experiences, each examiner's traditional academic history was queried; Table 6 reports the highest level of education earned for each participant, with 54% possessing a Bachelor's degrees, and 40% having earned a Master's degree.

Years of Footwear Experience	<1 Year	1-2 Years	3-5 Years	6-8 Years	8+ Years
#	4	7	10	12	37
Total Years of Forensic Experience	<1 Year	1-2 Years	3-5 Years	6-8 Years	8+ Years
#	0	2	1	5	60

Table 5: Years of experience for n = 70 participants (note that 2 examiners did not give a response for "Total Years of Forensic Experience").

Training in Last 5 Years	0 Times	1-2 Times	3-4 Times	4+ Times
#	2	24	20	24
Training Provider	IAI	Conference	Private or Consultant	Vendor or Supplier
#	45	56	41	16
Education Level	Associate	Bachelor	Master	Doctorate
#	1	38	28	2

Table 6: Participant training and education for n = 70 examiners. Note that 1 examiner selected "other" for their education level, and 11 examiners selected "other" for their training provider.

Finally, Table 7 reports that half of the participants in this study use the SWGTREAD (2013) scale [1] (without modification) in their laboratory, and approximately half are certified. Moreover, 87% participated in a proficiency test in the past year, and 60% have taught courses in the field of forensic footwear.

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

Summary	# Yes	# No
Uses SWGTREAD conclusion standard	35	35
Certified	33	37
Proficiency tested in past year	61	9
Received footwear training prior to casework	68	2
Has taught courses in footwear	42	28
Has conducted research in footwear	27	42
Has published works on footwear	14	56

Table 7: Summary of n = 70 participants' backgrounds, including use of the SWGTREAD (2013) [1] conclusion standard (without modification), certification, proficiency testing, and further activities related to teaching and research (note that one examiner did not give a response regarding past research).

2.4.B Case Variety

Each case is summarized in Table 8; five of seven required the analysis of two exemplars, while the remaining two required the analysis of a single exemplar. Each was comprised of 1200 PPI digital and print imagery, collected using a flatbed Epson Expression 11000XL Graphic Arts Scanner, and printed using a Canon Pixma Pro-1. Case materials consisted of a single questioned impression, 1-2 outsole exemplars, and 2 Handiprint exemplar replicates per known shoe. The questioned impressions were created under reasoanbly natural conditions (walking at a regular pace/stride length) using a range of media (blood, dust, wax), substrates (linoleum/ceramic/vinyl tiles, paper), and processing techniques (lifting, chemical/digital enhancement). Effort was expended to create "crime scene-like" impressions of the type, variety and quality encountered by analysts during routine casework. However, it is acknowledged that the wearer creating the questioned impressions had a smaller foot size than the actual outsoles used in this study, which may have created experimental limitations.

Case	Manufacturer of Known(s)	Size & Style of Known(s)	Substrate of Unknown	Medium of Unknown	Processing of Unknown	# of Known(s)
001	Converse	All Star (9)	Ceramic Tile	Blood	Leucocrystal Violet	2
002	Nike	Lebron James (10)	Vinyl Tile	Dust	Digitally Enhanced Gel Lift	1
003	Nike	Rosherun (9)	Ceramic Tile	Blood	Leucocrystal Violet	2
004	Nike	Air Max (10.5)	Linoleum Tile	Wax	Gel Lift of Magnetic Powder	2
005	Nike	Air Max (11)	Vinyl Tile	Dust	Digitally Enhanced Gel Lift	1
006	Nike	Air Max Cage (10)	Paper	Dust	Digitally Enhanced	2
007	Under Armour	Unknown (10 & 11)	Ceramic Tile	Blood	Leucocrystal Violet	2

Table 8: Shoes, substrates, media, and processing techniques used to create simulated case materials.

2.4.C Case Analyses

Each participant received a package via USPS of all relevant case materials, including high resolution color prints, a set of blank acetates for overlay annotation, a CD containing the

electronic reporting software, a copy of the Scientific Working Group for Shoeprint and Tire Tread Evidence (SWGTREAD) 2013 Conclusion Standard [1] and an instruction document (with additional weblinks to access electronic copies of all case materials, including digital files of 1200 PPI imagery). Participants were asked to process the simulated cases as if each were routine casework, and analyze the case materials according to their training and expertise, assuming that no time had passed between collection of the questioned and test impressions (*i.e.,* the absence of any change due to continued usage/wear). After performing a routine analysis, participants were asked to respond to a series of questions using a customized software reporting interface that solicited responses regarding the similarity, dissimilarity, clarity and value of manufacturing and wear-acquired features used when reaching conclusions. The instruction packet and illustrations of the graphical user interface can be reviewed in Appendix A.2. Based on all instructions, the only anticipated deviation from typical casework was the absence of consultation and/or any type of independent verification of examiner conclusions prior to reporting.

2.4.D Evaluation of Class Characteristics

Table 9 reports the value that examiners assigned to the class characteristics of outsole design, physical size, and the size of individual or grouped tread elements when comparing a questioned impression with an exemplar. For each row in Table 9, the same feature (*e.g.,* design) if summed across all values (association, exclusion, not evaluated or insufficient) will equal n = 70 (illustrated as a series of gold-shaded cells for row 001K1). From the summary data, it is clear that the majority of examiners routinely evaluate these class characteristics in terms of association or exclusion. In fact, 99.9% ± 0.41% (mean percentage ± 1 standard deviation) compared overall design (median of 100%), 96.8% ± 2.0% compared the overall physical size (median of 97.1%), and 97.9% ± 1.9% compared individual/grouped tread size between questioned and test impressions (median of 97.9%). However, the only shoes with observable class differences (and therefore value for exclusion) exists for comparisons of appropriate questioned impressions with 003K2, 005K1, 007K1 and 007K2B. Thus, all other selections of "value for exclusion" are not fully understood.

In contrast, a limited number of examiners chose not to evaluate physical size of the outsole and/or physical size of tread features. This observation was not anticipated, since by definition, physical size refers to the "dimensions, shapes, spacing and relative positions of the footwear outsole design components" [5]. In hindsight, the structure of the reporting interface may have created confusion, but moving forward, additional study may be warranted to determine how examiners define physical size, if the unevaluated observations are a product of varying interpretations in the definition, and under what circumstances these features would not be evaluated during a comparison.

	Association			Exclusion			Not Evaluated			Insufficient		
Case	Design	Size	Tread Size	Design	Size	Tread Size	Design	Size	Tread Size	Design	Size	Tread Size
001K1	65	55	58	5	9	10	0	2	2	0	4	0
001K2	68	63	65	2	2	2	0	3	3	0	2	0
002K1	66	54	61	4	9	5	0	1	1	0	6	3
003K1	68	60	62	2	8	8	0	1	0	0	1	0
003K2	62	27	34	8	41	32	0	2	3	0	0	1
004K1	66	51	55	4	9	12	0	3	2	0	4	1
004K2	69	65	68	1	2	1	0	1	0	0	2	1
005K1	61	27	40	8	34	25	1	1	0	0	8	5
006K1	66	57	67	3	2	3	0	3	0	1	8	0
006K2	63	51	54	7	8	15	0	3	1	0	8	0
007K1	61	21	23	9	47	44	0	2	3	0	0	0
007K2	65	46	46	5	21	21	0	2	3	0	1	0

Table 9: Evaluation of class features when comparing questioned and test impressions for n = 70 examiners (note that "design" refers to the geometric pattern, "size" refers to the physical size of the outsole, and "tread size" refers to the size of individual tread elements or groups of tread elements). The same sub-column header across all columns (an example is highlighted in orange) will sum to n = 70.

2.4.E Features Marked

In total, 3,524 features of interest were annotated by examiners when reporting 840 conclusions. Not surprisingly, wear patterns and RACs accounted for the majority of annotations (46% and 36%, respectively and resulting in 82% combined). Table 10 reports the feature type and frequency of marking per case. The first nine (9) items in the table could be selected by the user from a pull-down menu, and included features such as stippling, mold defect, die cut variation, air bubble, foxing strip, etc. The tenth option was "other," which required the examiner to provide input (a label) for the selected feature. After reviewing these inputs, some of the items marked as "other" could be remapped to existing features for the purpose of summarization. For example, an examiner selected "other" and typed "specific wear," but for the purpose of an overall summary, this was remapped to "wear" in Table 10. After this remapping, 210 annotated features marked as "other" persisted.

	Case 1		Case 2	Case 3		Case 4		Case 5	Case 6		Case 7			
Feature	K1	K2	K2	K1	K2	K1	K2	K1	K1	K2	K1	K2A	K2B	Total
Stippling	0	0	2	80	58	21	14	2	4	3	4	1	7	196
Mold Defect	7	9	0	4	3	2	1	0	1	1	5	0	5	38
Die Cut Variation	0	0	0	0	0	0	0	0	1	2	0	0	0	3
Air Bubble	0	6	2	0	0	1	6	0	0	0	0	0	2	17
Foxing Strip	3	6	0	0	0	0	0	0	0	0	0	0	0	9
Toe/Heel Cap	0	3	0	0	0	0	0	1	0	0	0	0	0	4
Wear	230	325	41	116	48	125	74	28	158	160	159	21	129	1614
Schallamach	1	6	3	0	0	41	106	1	3	1	3	0	3	168
RAC	73	116	39	137	88	192	260	43	15	27	92	5	178	1265
Other	14	25	29	20	14	5	5	41	12	10	13	2	20	210
Total	328	496	116	357	211	387	466	116	194	204	276	29	344	3524

Table 10: Summary of all features marked per comparison, totaling 3,524 annotations (note that examiners were permitted to mark on the questioned impression only, the known impression only, or both simultaneously; regardless of which option they selected, each marking was counted as a single feature).

Figure 1 illustrates that nearly a quarter of the 210 features marked as "other" and that could not be remapped were listed as "cannot determine" (suggesting that an examiner noted a difference or similarity between the questioned and known impression, but was unable to label the feature's identity, possibly because they did not have access to the physical outsole for the known). A smaller percentage were grouped as miscellaneous (*e.g.,* "wear turning into RAC," "movement/slippage," "void," and "possible incomplete mixing of outsole material"). Finally, just over half (56%) of the features marked "other" could be categorized as class characteristics.

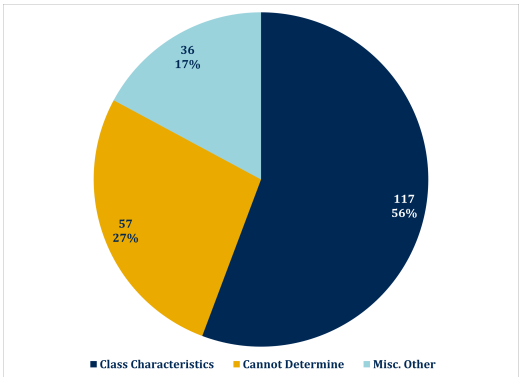

Figure 1: Frequency of qualitative description for 210 features marked as "other." Note that 56% were actually class characteristics (*e.g.,* "spacing of elements," "same physical shape and size," "design element difference," etc.).

2.4.F Interquartile Range

Tables 11 and 12 report examiner conclusions for all mated pairs, non-mated pairs, the combined dataset, and each individual comparison (questioned (Q) versus known (K)). Results are presented as both frequency/count and percentage. The gold-shaded cells in Table 12 correspond to community agreement (which is roughly defined as the interquartile range (IQR)). By definition, the interquartile range is intended to capture the middle 50% of the data. However, as applied to the categorical conclusions here, the IQR is intended to approximate the community agreement in conclusions, which means although it cannot capture less than the middle 50% of all conclusions, for comparisons with high agreement, it can extend and capture a higher degree of consensus among responses. This is illustrated in the final three columns of Tables 11 and 12, which report the number (percentage) of participants whose conclusions were within the minimum of the interquartile range. Results indicate that with 90% confidence (based on the Clopper-Pearson Exact method) [6], the community agreed upon IQR includes 75.9% - 83.2% of all responses for mated pairs, 87.4% - 92.1% for all non-mated pairs, and 83.5% - 87.6% for all combined data, with a low of 56% for 003Q versus 003K1, and a maximum of 97% for 007Q versus 007K1.

Comparison	Exclusion	Indications	Limited	Association	High Degree	Identification	IQR Count (Median %) (Mean% + SD%)	IQR % Lower	IQR % Upper
Combined	370	76	87	135	64	100	715 (89.3) (85.6 + 11.1)	83.5	87.6
Non-Mates	350	66	31	35	2	0	436 (91.4) (89.8 + 6.69)	87.4	92.1
Mates	20	10	56	100	62	100	279 (85.7) (79.7 + 14.1)	75.9	83.2

Table 11: Count (percentage) of examiners providing SWGTREAD (2013) [1] conclusions for mated (M) pairs, non-mated (NM) pairs and for all data combined, including community agreement (IQR) and its 90% confidence interval as a function of sample size.

Comparison	Exclusion	Indications	Limited	Association	High Degree	Identification	IQR Count (%)	IQR % Lower	IQR % Upper
001K1 NM	49 (70)	15 (21)	2 (3)	4 (6)	0 (0)	0 (0)	64 (91.4)	83.8	96.2
003K2 NM	66 (94)	2 (3)	0 (0)	2 (3)	0 (0)	0 (0)	66 (94.3)	87.4	98.0
004K1 NM	65 (93)	4 (6)	1 (1)	0 (0)	0 (0)	0 (0)	65 (92.9)	85.6	97.1
005K1 NM	34 (49)	13 (19)	13 (19)	9 (13)	0 (0)	0 (0)	60 (85.7)	77.0	92.0
006K2 NM	33 (47)	18 (26)	12 (17)	7 (10)	0 (0)	0 (0)	63 (90.0)	82.0	95.2
007K1 NM	68 (97)	1 (1)	1 (1)	0 (0)	0 (0)	0 (0)	68 (97.1)	91.3	99.5
007K2B NM	35 (54)	13 (20)	2 (3)	13 (20)	2 (3)	0 (0)	50 (76.9)	66.7	85.2
001K2 M	2 (3)	0 (0)	5 (7)	12 (17)	28 (40)	23 (33)	63 (90.0)	82.0	95.2
002K1 M	6 (9)	8 (11)	25 (36)	30 (43)	0 (0)	0 (0)	55 (78.6)	68.9	86.3
003K1 M	11 (16)	2 (3)	4 (6)	19 (27)	20 (29)	14 (20)	39 (55.7)	45.2	65.9
004K2 M	0 (0)	0 (0)	1 (1)	0 (0)	7 (10)	62 (89)	62 (88.6)	80.3	94.2
006K1 M	1 (1)	0 (0)	21 (30)	39 (56)	7 (10)	1 (1)	60 (85.7)	77.0	92.0

Table 12: Count (percentage) of examiners providing SWGTREAD (2013) [1] conclusions for each Q versus K comparison, and whether or not the known is a mated (M) or non-mated (NM) pair. The gold-shaded cells represent the conclusions that span the interquartile range (IQR), and the final three columns of the table report the number (percentage) of participants that reported conclusions within the interquartile range, and the 90% confidence interval for this estimate. Note that comparisons 002Q versus 002K1, 005Q versus 005K1, and 006Q versus 006K1 all had 1 response of "insufficient detail," which is not shown in this table; all other rows will sum to 70 (and result in percentages that sum to 100% barring rounding) except 007K2B which sums to 65 since 5 examiners reviewed a different impression (denoted as 007K2A) that had only limited circulation before being replaced with 007K2B (researchers felt that 007K2A was too easy owning to a patent/prominent RAC that spanned almost a full lug and therefore decommissioned this impression within a month of starting the study).

2.4.G Examiner-Specific Impact on Results

In order to determine the degree to which a single or specific examiners impacted reliability results, Figure 2 reports the frequency (percent) of examiners with 0, 1, 2, etc. responses outside of the IQR range (out of a total of 12 responses across 7 cases). Inspection indicates that 19% of all respondents were always within the IQR, while 33% were outside for a single conclusion. In contrast, and of more concern, are the examiners that are consistently outside of the IQR (*e.g.,* the 6 examiners with 4 or 5 conclusions outside of the IQR). Possible explanations for this discrepancy are numerous, but may include disparities in training, examiner inexperience, and/or a persistent variation in interpretation of the SWGREAD 2013 conclusion standard [1]. Regardless of the origin, these variations should be addressed in order to allow these analysts to self-calibrate against community norms. In addition, if the 12 comparisons in this study are considered representative of typical casework, then the 18 analysts with 3 conclusions outside of the IQR are also expected to form conclusions and opinions that are consistently different from the majority of their peers approximately 25% of the time.

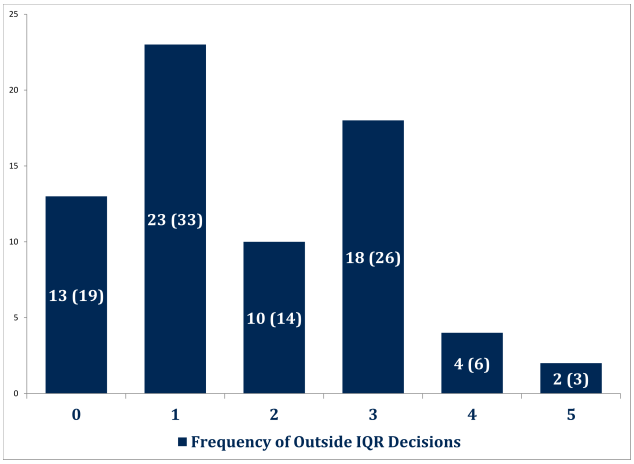

Figure 2: Frequency (percent) of examiners with 0, 1, 2, etc. responses outside of the IQR range.

2.4.H Expected Conclusions

Accuracy is defined according to the President's Council of Advisors on Science and Technology (PCAST) 2016 report [7]., or the known probability (or frequency) at which "an examiner obtains correct results both (1) for samples from the same source (true positives) and (2) samples from different sources (true negatives)" [7]. Unfortunately, this is not trivial to calculate when the conclusion standard for the community is a seven-point scale (such as the Scientific Working Group for Shoeprint and Tire Tread Evidence (SWGTREAD) 2013 conclusion standard [1]), rather than binary conclusive determinations (such as identification and exclusion, as exists within some other forensic pattern sciences). Moreover, there is little guidance on how to handle this nuance when attempting to determine the accuracy for assignment to decisions within a categorical scale that mimics a Likert scale varying from strong to weak dissociations (*i.e.,* exclusion, indications of non-association) and weak to strong associations (*i.e.,* association of class, high degree of association, identification). In addition, this complication is not alleviated by a research study with known ground truth. For example, even though ground truth was known for every simulated comparison pairing a crime scene-like questioned impression with a known test impression within this reliability study, the questioned prints were deposited and collected under natural conditions, and thus vary in both quality and clarity, as well as inherent discrimination potential (degree and type of wear, presence/absence of randomly acquired characteristics (RACs), etc.), possibly resulting in outcomes that span a range of SWGTREAD (2013) conclusion categories [1]. Thus, a contrived research paradigm with ground truth does not solve the issue of defining a reasonable or accurate accepted conclusion. Accordingly, although the research team knew which shoe created which impression, binary conclusions such as identification and exclusion were not anticipated for each and every mated pair and non-mated pair, respectively. Instead, to define an expected/accepted conclusion, each questioned/known impression combination was independently evaluated with respect to ground truth, observable features (and their associated reliability), and the SWGTREAD (2013) conclusion criteria [1]. For example, consider a known non-mated shoe without any significant characteristics of use, and that agrees in both outsole design and physical size with a questioned impression. Under this scenario (assuming no differences between impression features and outsole characteristics) a conclusion of association of class would be defined as reasonable (or acceptable) according to the SWGTREAD (2013) guidelines [1]. In other words, "...the known footwear is a possible source of the questioned impression and therefore could have produced the impression" [1], noting (importantly) that other outsoles with the same characteristics observed in the impression are also included in the population of possible sources.

As a result, the research team was presented with a difficulty not believed to be present in many other forensic reliability studies. In order to address this challenge, solutions for similar problems in other fields were considered. This revealed that consensus is typically the major study goal in subjective judgment analysis, while accuracy is relegated for idealized scenarios (*e.g.,* the accuracy of a weather forecast or a financial prediction that can be assessed by gathering additional information after a time delay). Thus, the research team approached the accuracy assessment problem using an accepted technique employed in other fields, such as the evaluation of surgical procedures or images, wherein a small number of individuals

"establish a gold standard" [8, 9]. In other words, the research team was afforded an "oracle" status, and permitted to define what would be considered accurate and inaccurate, while still allowing for some degree of opinion evolution.

To achieve this, each questioned/known impression combination was independently evaluated with respect to ground truth, observable features (and their associated reliability), and the SWGTREAD (2013) conclusion criteria [1], allowing the research team to draft an acceptable set of conclusions for each comparison. This process was repeated independently by four members of the research team (including one practitioner partner). All draft results were tabulated and through conference, discrepancies were discussed and evaluated until agreement was obtained within the team. The process of defining an acceptable range of conclusions was repeated a second time after data collection and during analysis of results, during which time the research team examined the range of responses provided by the 70 members of the forensic footwear community, and predominant categories on either side of any previously accepted range were re-evaluated after consideration of participant responses. This review resulted in two changes; first, the acceptable conclusions permitted for the comparison of 003Q with 003K2 was reduced from exclusion and indications of non-association to exclusion only. Consequently, any selection of indications of non-association for this pairwise comparison were deemed a 'failure to exclude' wherein exclusion is considered the correct answer based on the observable and reliable size differences that could be measured (varying between 3mm and 8mm) between the questioned and test impression. Second, the conclusions permitted for the comparison of 005Q with 005K1 was expanded from exclusion and indications of non-association to allow for exclusion, indications of non-association and limited association. The extension of the permitted range for this pairwise comparison was based on participants' detection of a size difference, but many comments (made by nearly 30% of respondents) indicating an inability to confirm that the differences being observed were 'reliable.' Note that these two changes do not reflect any fundamental persuasion of opinion of the research team by the community group-decisions. Rather, both are a reflection of the artificial research paradigm used in this study. More specifically, examiners were not able to prepare their own exemplars or inspect outsoles, which is typically afforded in actual casework, ergo their comments regarding reliability (*i.e.,* an examiner is detecting a size difference, but commenting on its reliability without being able to perform additional comparisons). For comparison 003K2, the size difference was large enough that the community deemed it reliable in the absence of the outsoles, while for case 005, the community noted a size difference but expressed uncertainty in its reliability (which presumably could be rectified if afforded the actual outsoles). Thus, the research team in one instance reduced, and in another instance expanded, the accepted range of conclusions in order to account for limitations in study-design.

Conversely, consensus became the focus of reproducibility. Fortunately, measuring consensus with ordinal scales is somewhat easier than assessing accuracy, but its quantification differs from both consensus estimation/group decision making and crowd ranking, where the goal of the latter is to reach consensus or agreement through discussion and opinion evolution, while the goal of the former is to quantify the degree of agreement reached by independent observers during a single round of decision making. Thus, for the purpose of this study,

consensus and dissension are considered a proxy for reproducibility, where reproducibility is defined according to the PCAST (2016) report [7], or the known probability (or frequency) at which "different examiners obtain the same result, when analyzing the same samples" [7]. As with accuracy, this metric is likewise complicated by the use of a seven-point conclusion standard. For example, if a participant can select between two binary categories (*i.e.,* agree or disagree), then if an actual ranking or agreement model exists for the decision (which is assumed to be true for expert opinions within scientific disciplines, versus, say, users' preferences in movies) then agreement should be higher for these types of binary decisions, than for experts presented with Likert scales with increasing numbers of categories (*i.e.,* strongly disagree, disagree, neutral, agree, strongly agree). With regard to the SWGTREAD (2013) conclusion standard [1], after removing insufficient detail, the remaining conclusions represent an ordinal scale, ranging from strong to weak exclusionary statements, followed by weak to strong associative statements. When presented with similar scales, Tastle and Wierman (2007) [10] illustrate that measures of agreement are poorly described by typical metrics, such as the mean, standard deviation, and entropy. As a specific example, consider a five-point Likert scale; if the mean response is near the end points of the scale (one or five) the variance must be smaller than if the mean is at the midpoint (three) [11]. Thus, a more appropriate measure of consensus (C) was sought, as illustrated in Eq. 1. This metric is bounded between zero and one, and is an estimate of the variability in responses, where $i = 1, 2, ..., n$ equals the index of the category of interest (n equals six in this study for each of the SWGTREAD (2013) conclusion categories after excluding insufficient detail [1]), X_i equals the value assigned to the category of interest, p_i equals the proportion of conclusions in the category of interest relative to the total, μ_x equals the mean score across all conclusion categories, and d_x equals the width of the conclusion categories ($d_x = X_{max} - X_{min}$ = 6 - 1 = 5) [12].

$$C = 1 + \sum_{i=1}^{n} p_i \ log_2 \left(1 - \frac{|X_i - \mu_x|}{d_x}\right) \tag{1}$$

Figure 3 reports the frequency (percentage) of expert decisions within each SWGTREAD (2013) conclusion category [1] with the expected (accepted) decision categories highlighted in green, and consensus (C) calculated according to Eq. 1 (16). In addition, the spread of decisions per comparison is illustrated via box plots that highlight the median, interquartile range (IQR) and possible outliers (1.5 × IQR) [13].

Note that when the IQR is used as the accepted range for accuracy, then the mean accuracy is 85.6% ± 11.1% (with a median of 89.3% and a 90% confidence interval between 83.5% and 87.6%). Conversely, when the research team is afford the right to define the range, the expert accuracy ranged from a low of 55.7% to a high of 97.1%, with a mean of 82.8% ± 11.9% (a median of 85.7% and a 90% confidence interval between 80.5% and 84.9%). The observed difference in mean accuracy is 2.8%, with a standard deviation of 16.3% based on addition in quadrature. Assuming the difference is normally distributed around a mean of zero, then the observed value using IQR differs from the expected by 2.8/16.3 = 0.17 standard deviations,

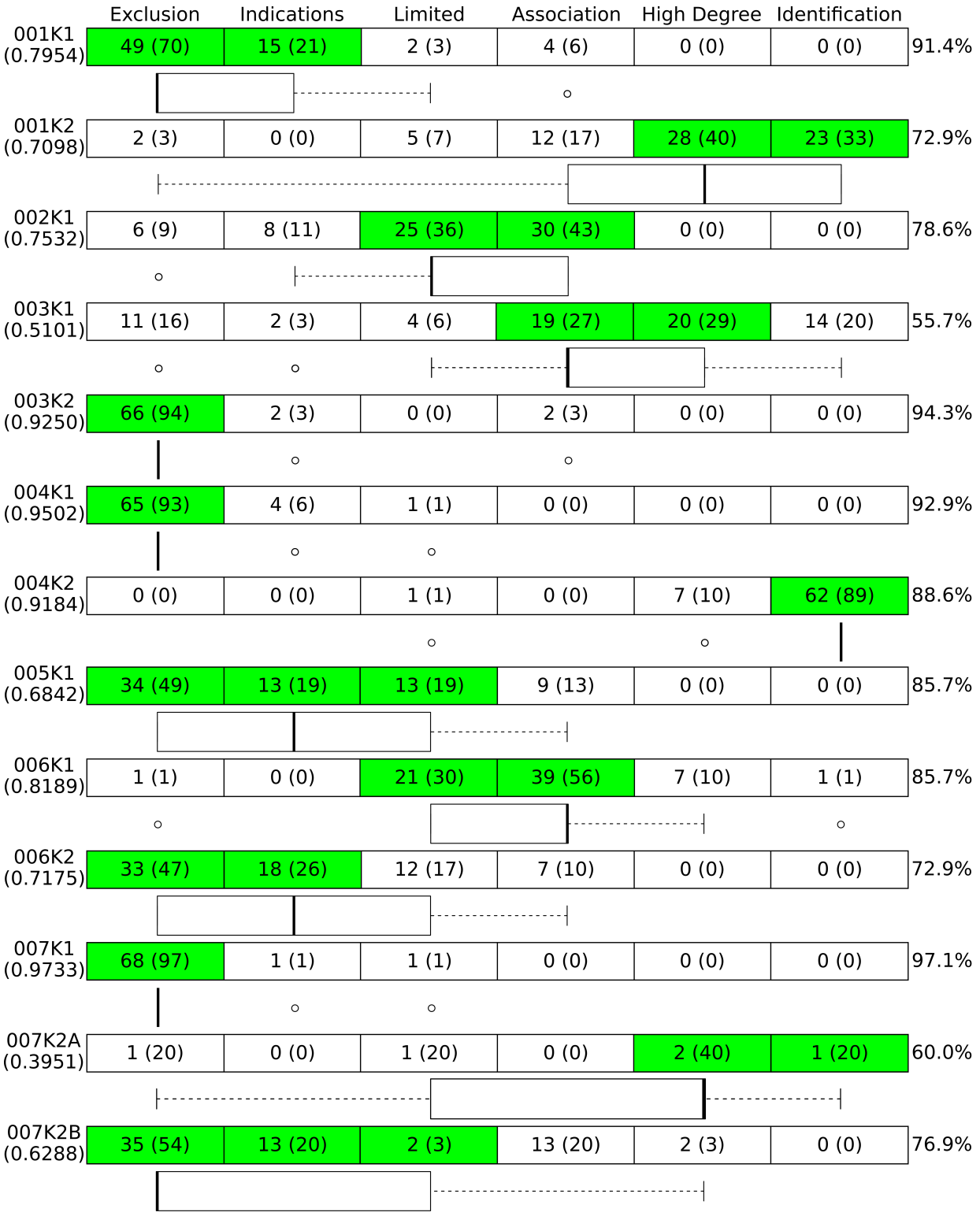

Figure 3: Range of expert conclusions for each questioned-known impression comparison, as a function of frequency (percentage), with the acceptable conclusions highlighted in green and reported as a total percentage at far right (accuracy). Consensus of examiner decisions (C) [12] is reported below the comparison number and visually illustrated in the form of a box plot detailing median (bold line), interquartile range (IQR) and if present, outliers (o) [13].

with a probability p(outside 0.2σ) of almost 85%, meaning a failure to detect any statistically significant difference in the accuracy estimates using either the IQR or the research team's defined range based on the assumption of normality [14]. In other words, if the IQR is treated as the forensic footwear community's group decision for each comparison, then on average, the research team's expected range of conclusions (as the "oracle") is not statistically different from the community's group decision.

Using consensus to evaluate the dispersion in responses (which are metrics that are independent of the number of participants), and ignoring comparison 007Q versus 007K2A based on sample size, the remaining consensus measures range from a low of 0.5105 (for comparison 003Q versus 003K1), a maximum of 0.9733 (for comparison 007Q versus 007K1), with a mean of 0.7821 $\pm$ 0.1422 and a median of 0.7743. In terms of mates and non-mates, the consensus among mated pairs equals 0.7421 $\pm$ 0.1516 (median of 0.7532), and the consensus among non-mated pairs equals 0.8106 $\pm$ 0.1396 (median of 0.7954). One less this value is a measure of dispersion, and both collectively describe the reproducibility in responses when using an ordinal conclusion scale that varies between strong to weak disassociations and weak to strong associations.

This metric considers not only the proportion of responses within a selected category, but also the distance between each category. For the purposes of computation, the distance between categories was considered constant (*i.e.,* the data and information required to transition from exclusion to indications of non-association is equal to the data and information required to transition from indications of non-association to limited association). Stated another way, the difference in temperature between 10^o and 15^o is the same as the difference between 15^o and 20^o. Although true for a relative evaluation of changes in temperature, this is unlikely to be true for the SWGTREAD (2013) conclusion standard [1] which is ordinal, but the scaling between each category is unknown. Moreover, without further study, use of an alternative set of distance weightings is equally speculative.

In addition to consensus, an assessment of reproducibility was conducted via inter-rater reliability (IRR). This metric is well-suited to quantify the degree to which a series of comparisons of questioned and test impressions are categorized the same way when analyzed by different examiners. High IRR means that examiners (raters) are interchangeable, which is desirable when the goal of the research study (and a criminal investigation/proceeding) is a measure of the similarity or dissimilarity between a questioned and test impression, regardless of the rater. Conversely, low IRR indicates that the individual rater or examiner plays a significant role in the categorization outcome [15], which in the case of forensic footwear comparisons suggests expert disagreement when presented with the same evidence, which often results in reduced clarity for the *trier-of-fact* tasked with interpreting the weight of evidence.

For nominal scales, agreement means that two raters provide identical conclusions. However, the concept of partial agreement exists when using an ordinal scale. For example, identification can be thought of as a certain conclusion, while high degree of association can be thought of as a highly probable conclusion. If some raters conclude identification and others conclude high degree of association these raters are not in total agreement, but they are also not in complete disagreement. Thus, the concept of partial agreement must be considered.

In addition to partial agreement, a measure of IRR should account for chance. Agreement by chance is not considered false, but rather a "bonus" that inadvertently inflates an agreement metric since it is not based on an underlying process. Moreover, chance agreement is higher when fewer categories are provided [15], so the number of categories present in a scale must also be accounted for.

Given the ordinal reporting scale and the need to characterize partial and chance agreement, this summary employs the weighted Gwet AC_1 coefficient (also referred to as the AC_2 coefficient) as illustrated in Eq. 2 [15]. The variable p_e denotes the percent chance agreement, while p_a denotes the percent realized agreement. The percent chance agreement is computed based on π_k which reports the fraction of examiners (raters) that compared questioned-test impression i and concluded k, across all comparisons n (or 835 since 007K2B accounts for 65 comparisons rather than 70).

$$\begin{aligned} AC_2 &= \frac{p_a - p_e}{1 - p_e} \\ p_a &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{q}\frac{r_{ik}(r_{ik}^{*}-1)}{r_i(r_i-1)} \\ p_e &= \frac{T_w}{q(q-1)}\sum_{k=1}^{q}\pi_k(1-\pi_k) \\ r_{ik}^{*} &= \sum_{l=1}^{q} w_{kl} r_{il} \\ \pi_k &= \frac{1}{n}\sum_{i=1}^{n}\frac{r_{ik}}{r_i} \end{aligned} \tag{2}$$

Conversely, the percent realized agreement is computed based on r_{ik}^{*} which describes the number of examiners that performed comparison i and reported a conclusion k, combined with any other conclusions l that are in partial agreement with k. The degree of partial agreement is a function of a weighting factor w_{kl} wherein raters are penalized less for reaching decisions in categories directly adjacent to one another and more for decisions separated by several categorical levels [15]. Although various weighting options exist (quadratic, ordinal, linear, etc.) an ordinal weighting system was employed in this analysis. The weight factor w_{kl} for two categories of interest (k and l) can be computed according to Eq. 3, where q represents the total number of categories into which conclusions can be classified ($q = 6$ for this study after removal of insufficient detail), and M_{kl} and M_{max} are combinations, or the number of combinations of 2 out of $max(k,l) - min(k,l) + 1$, and 2 out of q (or 15 for this study), respectively. Finally, T_w is the total of all weight factors (or 26.67 when using a six-level reporting structure) [15].

$$
\begin{aligned}
w_{kl} &= \begin{cases} 1 - M_{kl}/M_{max} & \text{if } k \neq l \\ 1 & \text{if } k = l \end{cases} \\
M_{kl} &= \begin{pmatrix} max(k,l) - min(k,l) + 1 \\ 2 \end{pmatrix} \\
M_{max} &= \begin{pmatrix} q \\ 2 \end{pmatrix} = \frac{6!}{2!(6-2)!} = 15 \\
T_w &= \sum_{l=1}^{q} \sum_{k=1}^{q} w_{kl} = 26.67
\end{aligned} \tag{3}
$$

Increasing numerical values of the AC_2 coefficient indicate increasing levels of examiner agreement. Moreover, with proper benchmarking, the magnitude of the coefficient can be related to a verbal scale. The purpose of benchmarking is to calibrate the verbal scale while accounting for study design (number of raters, number of comparisons, and number of response categories). To perform benchmarking, a four-step process is required [15]. First, the agreement coefficient and its standard error (SE) are computed (see Gwet (2014) [15] for SE computation). Second, the interval membership probability (IMP) is computed as illustrated in Eq. 4 (assuming a normal distribution) for each interval (a, b) in a verbal equivalent scale (poor, slight, fair, moderate, substantial and almost perfect). Third, cumulative probabilities are computed, starting from the highest benchmark level ("almost perfect"). Finally, the reported verbal equivalent for agreement for a specific study (*i.e.,* "moderate," "substantial," etc.) is found to be equal to the agreement category that contains the smallest cumulative probability exceeding 0.95 [15].

$$
IMP = P\left(\frac{AC_2 - b}{SE} \leq Z \leq \frac{AC_2 - a}{SE}\right) \tag{4}
$$

Using the Gwet AC_2 agreement coefficient [15], the footwear examiner agreement is described in Table 13. After benchmarking the computed coefficient, a verbal interpretation of footwear examiner performance maps between moderate and substantial agreement (excluding decisions of insufficient detail which are not part of an ordinal scale).

In an effort to try to place context on the IRR agreement computed for the footwear examiners, Table 14 reports a sampling of agreement coefficients collected from the literature. Note that these studies typically involve a very small number of experts, and benchmarking is not routinely performed despite its utility. However, based on just a small sampling of other studies, it does appear that the benchmarked IRR for the 70 experts in this study is relatively high.

Data Set	# Comparison Pairs	# Possible Conclusions	# Total Decisions	Gwet AC_2	SE	90% Confidence Interval Gwet AC_2		Verbal Equivalent
Six-Level Combined	12	6	832	0.7509	0.0875	0.6070	0.8948	Substantial
Six-Level Non-Mates	7	6	484	0.8818	0.0546	0.7919	0.9717	Substantial
Six-Level Mates	5	6	348	0.6562	0.1369	0.4310	0.8813	Moderate

Table 13: Inter-rater reliability analysis results for the Gwet AC2 agreement coefficient and the corresponding verbal equivalent for agreement after benchmarking. See Gwet (2014) [15] for SE computation.

Paper	Assessment Topic	# of Raters	# Categories/ # Conclusions per Case	# of Cases	Coefficient	Results	Verbal Equivalent(s)
Onate, et. al. [16]	LESS Landing Assessment	1 expert 1 novice	2 / 15	19	Fleiss' Kappa	κ=0.46 to 1.00	Moderate to Perfect
Andreasen, et. al. [17]	Medical Claim Compensation	15 experts	2 / 6	12	Fleiss' Kappa Gwet's AC_1	κ=0.41 to 0.53 AC1=0.43 to 0.54	Moderate
Gschließer, et. al. [18]	Diagnosis of Retinopathy	7 experts	6 / 1 2 / 2	52	Fleiss' Kappa	κ=0.26 to 0.55	Fair to Moderate
Acklin & Fuger [19]	Criminal Court Decisions	3 experts	3 / 3	150	Fleiss' Kappa	κ=0.24 to 0.81	Fair to Substantial
		3 experts	3 / 3	150	Krippendorff's Alpha	α=0.18 to 0.51	Poor to Moderate
		2 "experts"*	2 / 3	150	Cohen's Kappa	κ=0.35 to 1.00	Fair to Perfect
Nawrocka, et. al. [20]	Stage of Decomposition	120 novices	13 / 1 12 / 1 10 / 1	12	Krippendorff's Alpha	α=0.81 to 0.85	N/A
Lee, et. al. [21]	Automobile Color	6 novices	18 / 1	1000	Fleiss' Kappa	κ=0.22 to 0.98	Fair to Very Good

Table 14: Examples of IRR coefficients for experts and novices performing various design tasks. *Agreement was computed between two "decisions," which were a judge's verdict versus a pooled consensus decision from either two or three experts. Also note that Fleiss' Kappa is an extension of Cohen's Kappa for more than two raters, and that none of these studies benchmarked their verbal scales.

2.4.I Predictive Value & Error Rates

The benefit of a seven-point conclusion standard within the footwear community is the ability to succinctly describe the population of shoes that could have contributed to the questioned impression. From this "population" vantage-point, the scale is "U-shaped;" at either extreme (*i.e.,* exclusion and identification) the population is exact, while the internal categories permit wider associations or disassociations between a given shoe, and any other shoe of the same make, model, size, etc. Unfortunately, this increased degree of freedom in expression complicates any computation of accuracy since there are endless situations when, for example, an association of class is a valid conclusion for a known non-mated shoe *i.e.,* "...the known footwear is a possible source of the questioned impression... (and) other footwear with the same class characteristics observed in the impression are included in the population of possible sources") [1].

In order to allow for a direct comparison of error rates from this study and those reported in other forensic pattern evidence fields based on a three-point standard, a data transformation was required. This was achieved by remapping the SWGTREAD (2013) conclusion standard [1] into the Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST) 2013 three-level reporting scale used by fingerprint analysts [22] (*i.e.,* exclusion, inconclusive and individualization). The explicit purpose of this transformation was to allow for a first order comparison against the 2011 FBI fingerprint black box study [2]. To achieve this reformulation, footwear experts' decisions of exclusion and indications of non-association were re-assigned as exclusions, reports of limited association and association of class were re-classified as inconclusive, and finally, outcomes of high degree of association and identification were re-categorized as individualization. Moreover, ground truth (mates and non-mates) were used to create an appropriate confusion matrix. Although the authors acknowledge that this breakdown is an over-simplification of the conclusions that can be drawn in the field of forensic footwear evidence, and that many may assert that these groupings are problematic (*e.g.,* that high degree of association is not the same conclusion as identification/individualization), this segmentation nonetheless allowed for a reasonable comparison of expert error rates and predictive values among the fields of footwear and fingerprint evidence.

Table 15 reports the confusion matrix based on this data transformation; in total, two false positives occurred resulting in a false positive rate (FPR) of approximately 0.5% (FPR = 2/418). As a note for comparison, this same metric was reported as 0.1% (or six false inclusions) for past studies in fingerprint analysis [2]. In this study, both instances of false positives were committed for case 007K2B; two examiners reached high degree of association, reporting agreement of class characteristics and wear. However, this is invalid; wear differences are apparent, and this comparison included a possible manufacturing anomaly wherein the heel portion of the outsoles for the known non-mate and questioned impression appear to be affixed to the midsole in slightly rotated (mismatched) orientations, further precluding agreement when the questioned and known impressions are overlaid.

Confusion Matrix of Binary Footwear Conclusions		Examiner Conclusion		
		Identification (Positive)	Exclusion (Negative)	Total
True Conclusion	Identification (Positive)	162	30	192
	Exclusion (Negative)	2	416	418
	Total	164	446	610

Table 15: Confusion matrix of 610 forensic footwear decisions, reclassified as binary conclusions using ground truth, for direct comparison with the 2011 FBI fingerprint black box study [2]. Note than any examiner conclusions of either insufficient detail, limited association, or association of class have been excluded from analysis as these were reclassified as inconclusive outcomes.

The false negative rate (FNR) was higher than the false positive rate (approximately 16%), with 30 decisions (out of 192 possible identifications) falsely excluding a known mated shoe; this is approximately double the error rate observed for fingerprint analysts at 7.5% [2]. In total, 23 different analysts committed these 30 errors, with five examiners committing the error twice and one analyst providing three false negatives, meaning 43% of these errors were committed by six experts (5 experts $\times$ 2 errors + 1 expert $\times$ 3 errors = 13/30) which is less than 9% of all participants. As a general observation, it appears that many of the false

negatives or incorrect eliminations resulted from an improper characterization of impression size wherein a size difference (physical size and/or size and spacing of outsole elements) was reported when one did not exist.

After considering the observed error rates for this study, computation of predictive values (posterior probabilities) was conducted. Positive predictive value reports the percentage (or probability) of strong inclusionary decisions that are true mates, while negative predictive value describes the percentage (or probability) of strong exclusionary decisions that are true non-mates [2]. This evaluation is important because ground truth is not known in casework for which error rates are desired, thus an understanding of this "likelihood of correctness" in research studies represents a useful alternative. Based on 610 comparisons, the correct predictive value equals 98.8%, and the negative predictive value equals 93.3% (when 31% of comparisons are conducted on known mates and 69% on known non-mates) (Table 16).

Ground Truth	Conclusion	CWR	90% Confidence CWR		COR	90% Confidence COR		WP	OP	CPV	90% Confidence CPV	
Mates	Strong Inclusion	0.8438	0.7941	0.8852	0.9952	0.9850	0.9991	0.3148	0.6852	0.9878	0.9621	0.9962
Non-Mates	Strong Exclusion	0.9952	0.9850	0.9991	0.8436	0.7941	0.8852	0.6853	0.3148	0.9327	0.9132	0.9481

Table 16: Computed error rates and predictive values based on ground truth across 70 forensic footwear experts performing 610 comparisons. Key: CWR = correct within rate, COR = correct outside rate, WP = within prevalence, OP = outside prevalence and CPV = correct predictive value.

Across all possible mate-prevalences, Figure 4 is a plot of PPV and NPV based on ground truth for the 610 conclusive comparisons conducted in this study based on standard error computations using the Clopper-Pearson (exact method) [6] and Standard logit confidence intervals for predictive values [23]. As a point of comparison, the positive and negative predictive value for fingerprint examiners at a 62% mate-prevalence was previously found to be 99.8% and 86.6%, respectively [2]. This is illustrated in Figure 4 by the solid vertical line. At this same prevalence, the corresponding footwear performance values are 99.7% and 79.6%, with 90% confidence intervals between 98.9% to 99.9%, and 74.8% to 83.7%, respectively.

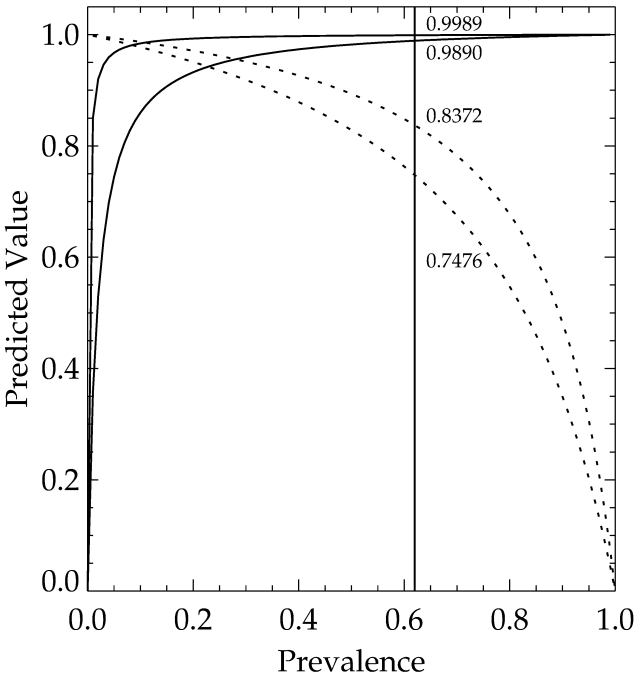

Figure 4: Plot of 90% confidence intervals for positive (solid lines) and negative (dashed lines) predictive value as a function of mate-prevalence.

2.4.J Possible Factor Dependencies

In an effort to better understand the factors that impact a footwear examiner's decision accuracy, all final conclusions were evaluated using the chi-square test of independence to determine the degree to which accuracy varied as a function of both examiner and case related factors. For any significant results, an adjusted Pearson's residual post-hoc analysis with a Bonferroni correction was applied [3, 4]. The examiner attributes under evaluation were education level, certification status, frequency of footwear examination/comparison, frequency of continuing education/training annually, and use of the SWGTREAD 2013 conclusion scale [1] for casework. In addition, case variables were considered, including individual comparisons, questioned impression clarity, case difficulty, number of knowns provided, presence of known match, ground truth, and final SWGTREAD conclusion [1].

Examiner/Laboratory Attributes

When evaluating accuracy, there was no evidence to suggest that performance was dependent upon education level (global p-value = 0.0966, fail to reject the null hypothesis of independence), as evidenced by the relatively consistent ratios between within and outside of range decisions across all degree types (Figure 5).

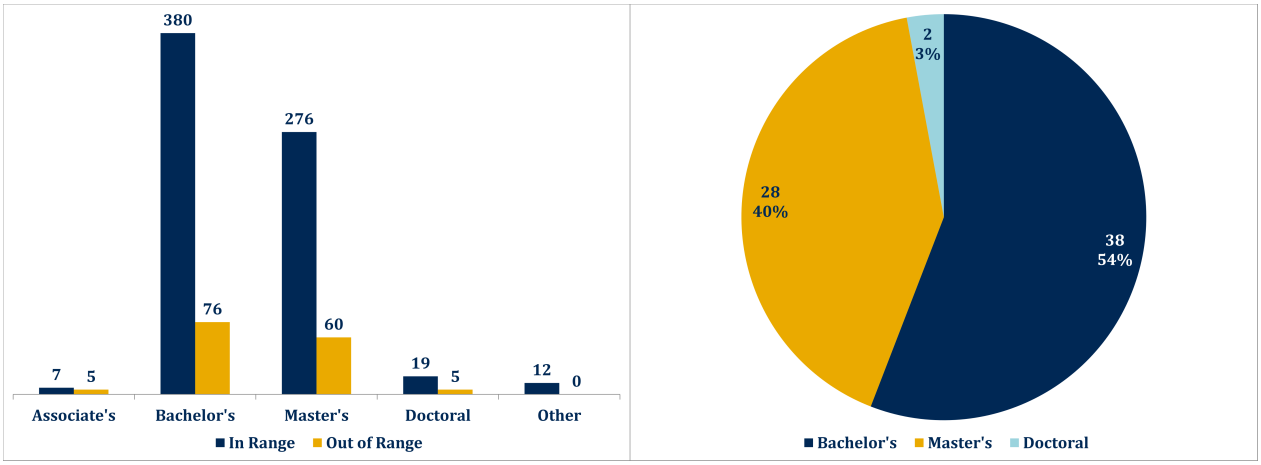

Figure 5: Accuracy of expert conclusions (within accepted range or outside of accepted range) as a function of whether the participant's highest level of education (left) and frequency of education level for 68 participants (right). Note: the remaining two experts have either an Associate's or a Licentiate Degree (reported as Other).

In addition to education, certification is another qualification that footwear experts can obtain. Although certification is not mandatory, after accumulating a specific amount of experience and training, and subsequently passing a written and practical examination, examiners can pursue this option. When considering the certification status of participants in this study, just under half of the analysts were certified forensic footwear examiners (Figure 6). Similar to education level, a dependence was not detected between global accuracy and certification (global p-value = 0.6921, fail to reject the null hypothesis of independence) as illustrated by the consistent number of in/out of range conclusions across all experts (Figure 6).

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

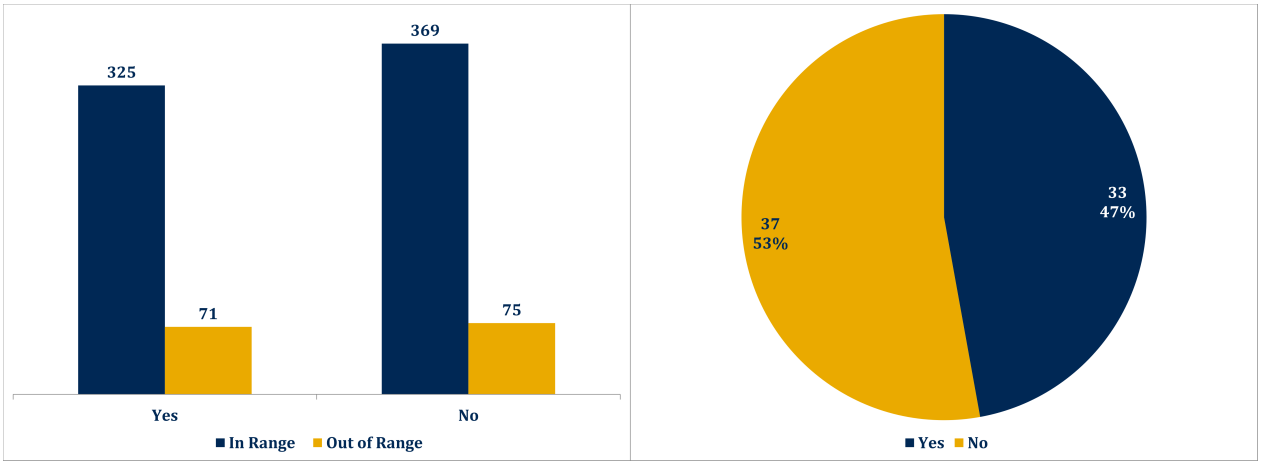

Figure 6: Accuracy of expert conclusions (within or outside of accepted range) as a function of whether the participant is a certified footwear examiner (left) and frequency of certification all 70 participants (right).

As part of certification, footwear analysts are required to participate in a minimum amount of continuing education/training courses, specifically for re-certification every five years [24], although again, this is not mandatory in order to conduct casework. When considering the frequency of training within the year prior to participating in this study, the vast majority of analysts participated in some form of training (with only 3% indicating that they did not do any continuing education during this time period), as exhibited in Figure 7. However, the amount of training attended likewise did not have a detected impact on expert performance ($p = 0.4838$).

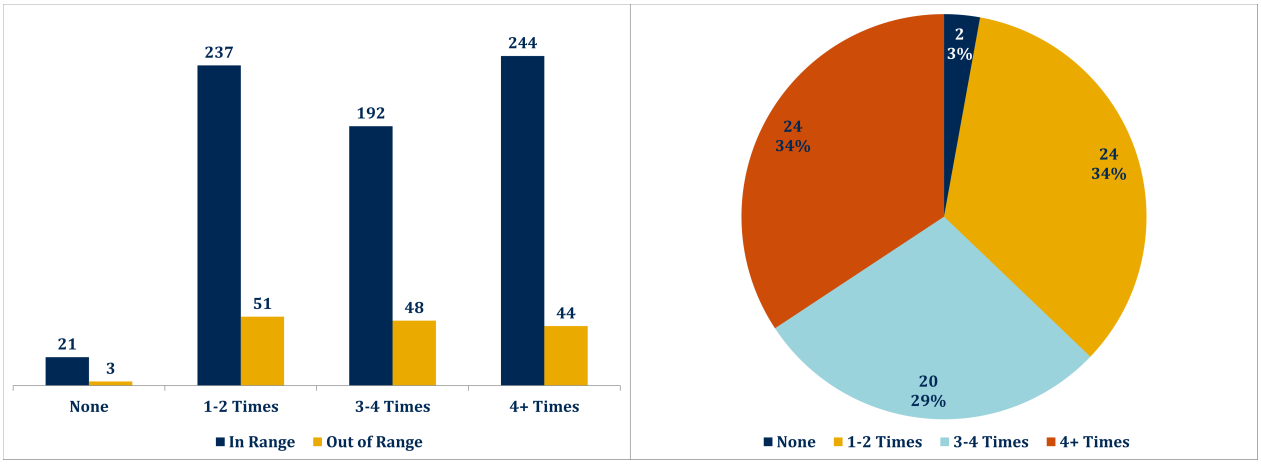

Figure 7: Accuracy of expert conclusions (within or outside of accepted range) as a function of the number of times they attend training annually (left) and frequency of certification all 70 participants (right).

Frequently, pattern evidence analysts in crime laboratories split their time between multiple disciplines, depending on caseloads. As such, it was of interest to collect background information from each participant to determine how frequently he/she conducts footwear examinations and comparisons, the task under study in this research. Over half (about 68%) of the experts in this study analyze footwear evidence in this capacity at least frequently

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

(with 27% responding frequent and 41% selecting very frequent), as detailed in Figure 8. Of the remaining participants, another 23% do comparisons occasionally and only 9% either seldom or very seldom. Chi-square analysis failed to reject the null hypothesis of independence (global p-value = 0.8626), indicating that there is no evidence that accuracy varies with frequency of footwear evidence analysis, as illustrated in Figure 8.

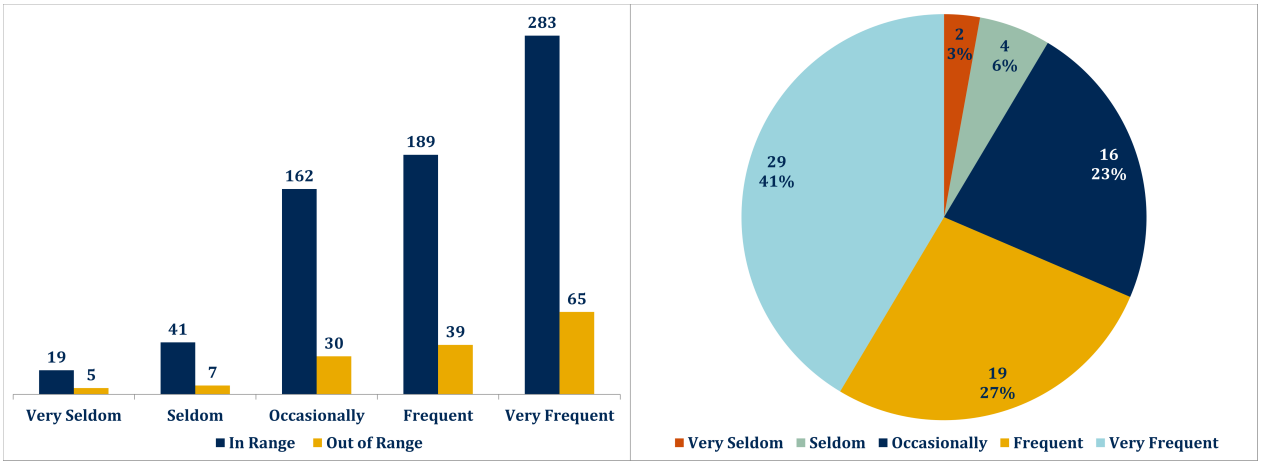

Figure 8: Accuracy of expert conclusions (within accepted range or outside of accepted range) as a function of how frequently the participant conducts forensic footwear examinations and comparisons in their current job capacity (left) and frequency of examination/comparison for all 70 participants (right).

For this study, all examiners were required to reach a conclusion based upon the seven-level SWGTREAD 2013 scale of conclusions [1]; however, not all laboratories use this standard. More specifically, of all participants, only half use the SWGTREAD scale regularly for casework, while the remainder use a different criteria (*e.g.,* different number of levels, terminology/articulation, and/or observation requirements) for reaching conclusions (Figure 9). Upon assessment of accuracy, there is no evidence that use (or lack thereof) that regular use of the SWGTREAD 2013 conclusion scale [1] for casework influenced expert performance (global p-value = 0.8555, fail to reject the null hypothesis of independence), as indicated by the near identical ratio between in/out of acceptable range conclusions in either scenario, shown in Figure 9.

Case/Comparison Attributes

After analysis of examiner attributes and their potential impact on accuracy, of which there was no evidence that any of the tested variables yielded differences in performance, an evaluation of eight case related factors was conducted: (i.) overall case, (ii.) individual comparison, (iii.) self-reported questioned impression clarity, (iv.) self-reported case difficulty, (v.) number of known shoes provided, (vi.) presence or absence of true source footwear, (vii.) ground truth, and finally, (viii.) reported SWGTREAD conclusion [1].

Owing to the fact that a variety of cases were presented to examiners (with various media, substrates, enhancement methods, outsoles, etc.), it was important to evaluate whether overall case or individual comparison of those provided had a significant impact on accuracy.

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

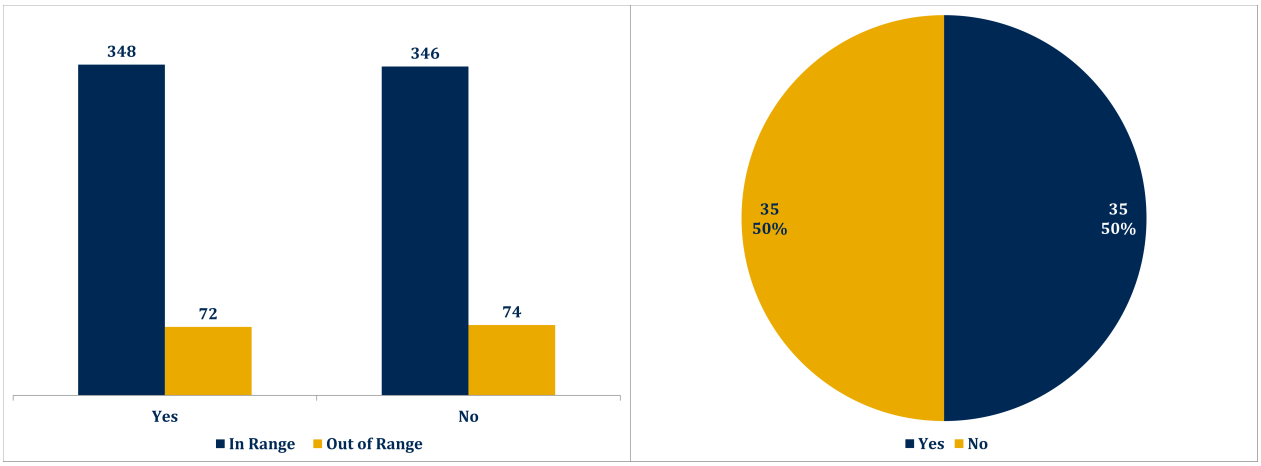

Figure 9: Accuracy of expert conclusions (within accepted range or outside of accepted range) as a function of SWGTREAD use within the analyst's laboratory (left) and frequency of SWGTREAD 2013 [1] use for all 70 participants (right).

Each participant was asked to conduct an examination of seven total cases; of these cases, five had two known shoes (resulting in 10 comparisons) and two had one known shoe (accounting for two additional examinations), as detailed in Figure 10. When evaluating performance as a function of case, the global p-value was significant ($p = 0.0106$, reject the null hypothesis of independence), thus providing evidence that accuracy varies with case (Figure 10).

The post-hoc analysis failed to detect evidence to reject the null hypothesis of independence for any of the cases; however, it showed that examiner accuracy was nearly dependent for case 003 and 004 ($p = 0.0078$ and 0.0063, respectively compared against a Bonferroni adjusted α of 0.0036).

For case 004, higher accuracy was observed than would be expected if the two variables (accuracy and case presented) were independent. The questioned impression for this case (a wet residue impression on tile, enhanced with magnetic powder and lifted with a white gelatin lifter) exhibited relatively high quality and contained several clearly patent Schallamach patterns and RACs that could be used for comparison. These facts likely made this a relatively straight forward case. Conversely, examiners were less accurate than would be expected for case 003. The questioned impression in this case was a blood print on ceramic tile, enhanced with leuco-crystal violet. However, the crime-scene print is a partial impression, which is not uncommon, but with a very smooth medial edge. For analysts potentially unfamiliar with the specific type of shoe, and therefore lacking knowledge about design and tread element size/spacing between sizes, this smooth edge may have appeared to be the true perimeter. Therefore, it is postulated that some experts used the edge to measure physical size of the outsole, and when compared against the true source shoe (003K1), erroneously reported a size difference and reached an exclusionary decision, thereby falling outside the range of expected conclusions.

Subsequently, an assessment of performance for individual comparisons was conducted in an attempt to determine if individual knowns exhibited variation in performance (Figure 10),

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

which indicated that accuracy varied as a function of single comparisons ($p = 9.637e^{-12}$, reject the null hypothesis of independence). Unsurprisingly, accuracy varied for the questioned-known comparison 003K1 ($p = 3.930e^{-10}$), in the same direction as previously stated for the case-level evaluation (and 003K2 nearly so).

Interestingly, however, an additional comparison (007K1) resulted in dependence that did not exhibit significance for overall case ($p = 0.0009$). For known shoe 007K1, analysts were more accurate than would be expected for the comparison with the questioned impression (blood on ceramic tile, enhanced with leuco-crystal violet). More specifically, this suspect footwear was not the source of the evidence and actually was a full size larger than the shoe that created the questioned impression, and likely as a direct result, examiners were highly accurate in excluding this shoe as the potential source.

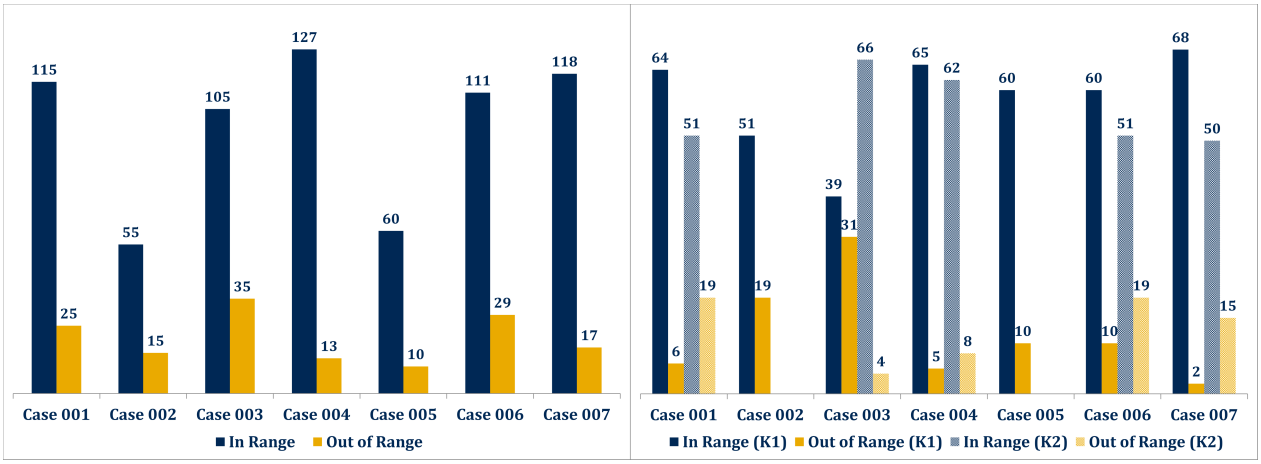

Figure 10: Accuracy of expert conclusions (within accepted range or outside of accepted range) as a function of the overall study case (left) as well as individual comparison (right).

For each comparison, participants were asked to rate the clarity of the presented questioned impression as either unsuitable, low, moderate, or high. Overall, there was roughly a 1:2:1 split between clarity ratings; approximately 25% of questioned comparisons were deemed low clarity, 52% were classified as moderate, and another 22% were considered high quality impression, as illustrated in Figure 11. Notably, one comparison for case 005Q versus 005K1 was excluded from the chi-square analysis because it was rated as unsuitable, but the expert continued analysis thereafter. As exhibited by the relatively consistent ratio of in versus out of range conclusions for each category, a global analysis of clarity revealed that there was no evidence to suggest that performance varied as a function of questioned impression quality ($p = 0.5769$, fail to reject the null hypothesis of independence).

In addition to reporting perceived clarity of the questioned impression, each analyst was asked to report the overall difficulty (easy, moderate, challenging) of each comparison. Similar to the breakdown observed for the clarity ratings, difficulty again exhibited an approximately 1:2:1 split, with roughly 22% of comparisons deemed easy, 56% classified as moderate, and the remaining 22% rated as challenging, as illustrated in Figure 12. Likewise, there is no evidence to suggest that accuracy was dependent upon self-reported difficulty globally ($p = 0.1397$, fail

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

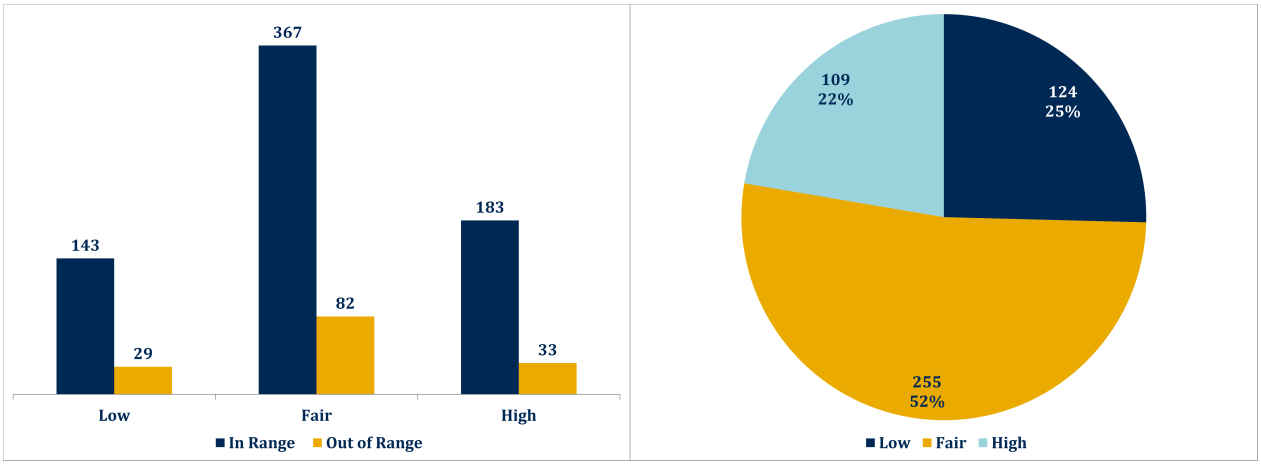

Figure 11: Accuracy of expert conclusions (within accepted range or outside of accepted range) as a function of the participant's self-reported questioned impression clarity (low, moderate, or high) (left) and frequency of reported clarities (right). Note: one comparison was excluded from this analysis because the examiner deemed the impression of unsuitable clarity; the final conclusion was outside of the acceptable range.

to reject the null hypothesis of independence), as evidenced by the consistent split between accurate and inaccurate decisions (Figure 12).

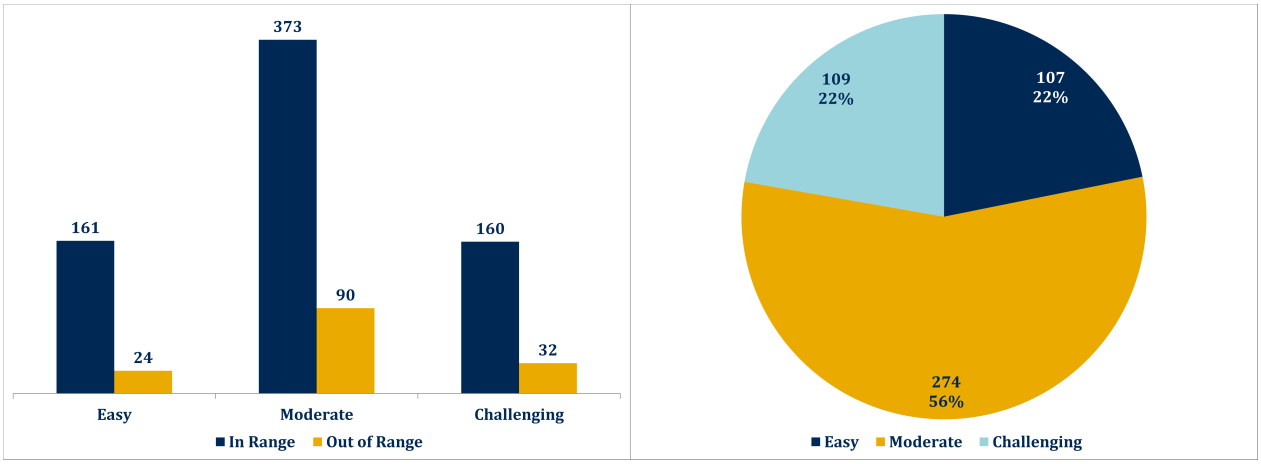

Figure 12: Accuracy of expert conclusions (within accepted range or outside of accepted range) as a function of the participant's self-reported case difficulty (easy, moderate, or challenging) (left) and frequency of reported difficulties (right).

During the design of this footwear black box study, several decisions were made regarding the cases to be presented to experts for examination, such as the number of known items per case (one or two) and the presence of a true source/known mate (KM) as one of the suspect shoes provided (closed- or open-set design). In an attempt to characterize performance across multiple conditions, this study included five cases with two known shoes (of which one did not contain KM footwear) and two cases with one known shoe (of which one did not contain KM footwear). Ultimately, chi-square analysis did not reveal a relationship between performance and the number of exemplars presented in a given case ($p = 0.8706$, fail to reject the null hypothesis of independence), as shown by the roughly 4:1 ratio of in versus out of range results

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

irrespective of the number of provided knowns (Figure 13). Similarly, when considering the presence of a KM in a given case, there was no evidence detected to suggest that accuracy was impacted by this factor ($p = 0.0754$, reject the null hypothesis). These results should be considered in combination when discussing the design of black box studies. More specifically, the 2016 PCAST Report on Ensuring Scientific Validity of Feature-Comparison Methods noted that such studies must employ open-set designs (*i.e.,* a KM is not provided in every case) in order to properly estimate error rates [7]. When a black box study has a closed set design, then there is a KM in every case. A consequence is a possible underestimate of the false positive rate, which is typically considered the "worst" of all possible errors in the pattern disciplines since it incorrectly links an individual to a piece of evidence. Suppose, for example, that a case contains two knowns for comparison and the examiner is able to reach an exclusion on the first comparison; in a closed-set, the analyst could simply call an identification on the other known without ever evaluating it and be correct. Interestingly, the results from this research do not indicate analysts perform differently for open- versus closed-set cases. However, this outcome may be confounded by additional study design factors, such as the fact that both of the open-set cases (005 and 007) exhibited size differences, which is arguably a straightforward means for reaching exclusions (and high levels of accuracy). Thus, additional research is merited.

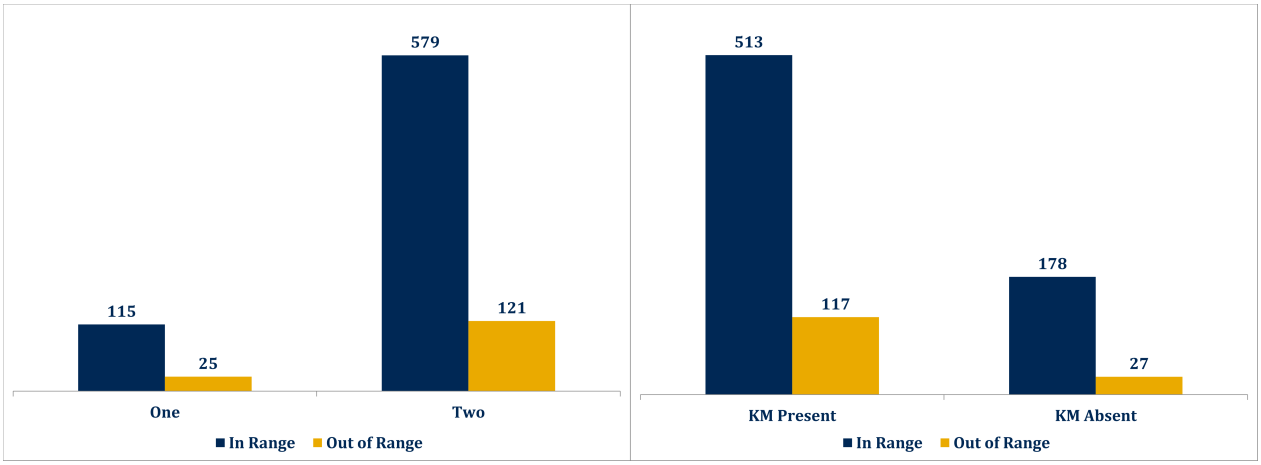

Figure 13: Accuracy of expert conclusions (within accepted range or outside of accepted range) as a function of the number of knowns provided for comparison in the case (left) as well as whether a known match was included as one of the knowns for the case (right).

Lastly, performance was evaluated as a function of both ground truth (source or non-source known footwear) and SWGTREAD 2013 conclusion [1], wherein it is expected that for KM shoes (true source) experts reach an associative decision and for KNM footwear (non-source or known non-mate) decisions are a function of observable non-associations (Figure 14). In total, 7 of 12 suspect shoes provided to examiners for comparison with the questioned impressions were KNMs. When considering ground truth, the global chi-square revealed a dependence with performance ($p = 1.753e^{-05}$, reject null hypothesis of independence), as evidenced by the varied ratio for within versus outside range conclusions (Figure 14, exhibiting an in to out of range ratio of approximately 3:1 for KMs and 7:1 for KNMs). Notably, experts exhibited higher decision accuracy than expected for non-source known footwear. Of the

seven non-source shoes provided for comparison, three exhibited a measurable size difference and one (case 004) contained patent RACs and Schallamach patterns that could be used for elimination. Thus, there are confounding factors that may explain why participants performed better than anticipated in this study when examining known non-mated pairs.

Likewise, evidence suggested that accuracy varied as a function of SWGTREAD conclusion ($p = 8.957e^{-16}$), namely for exclusion, limited association, and association of class characteristics ($p = 4.660e^{-15}, 3.403e^{-04}, 2.280e^{-09}$, respectively as compared against a Bonferroni adjusted α of 0.0042). Again, experts performed better than expected for exclusion outcomes ($p = 4.668e^{-15}$). Conversely, examiners were less accurate than expected for limited association and association of class characteristic conclusions ($p = 3.403e^{-04}, 2.280e^{-09}$, respectively). Given that these two conclusions are the least certain decisions on the SWGTREAD 2013 scale [1], and they can reasonably be reached for both KMs and KNMs, it is reasonable to expect that performance on these levels would be lower owing to the lowered certainty and increased variability in selecting one of the two conclusions.

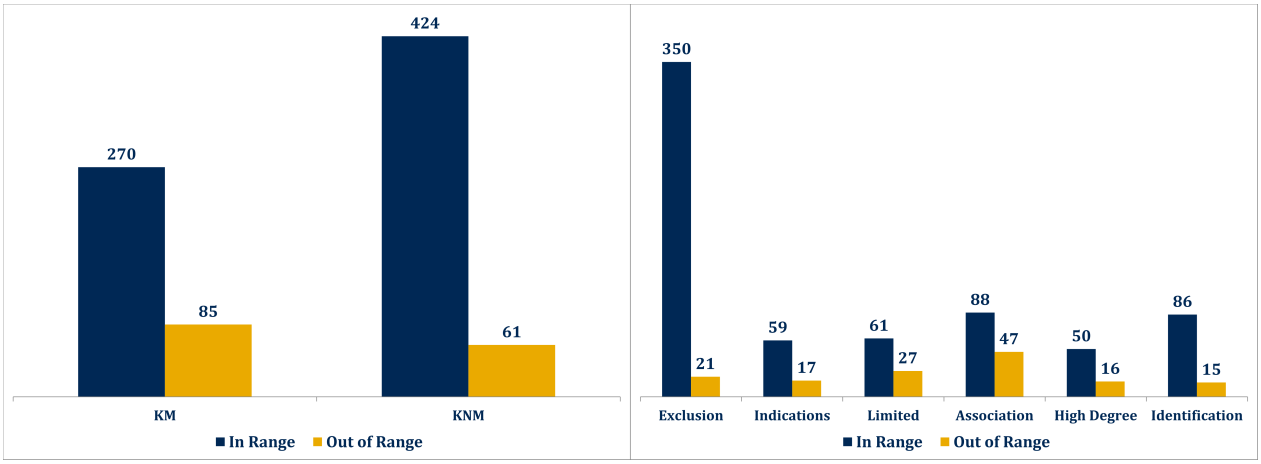

Figure 14: Accuracy of expert conclusions (within accepted range or outside of accepted range) as a function of the ground truth for a comparison (known match (KM) or known non-match (KNM)) (left) as well as the SWGTREAD 2013 conclusion [1] reached for a comparison (right).

2.4.K Dominance-Based Rough Set Approach (DRSA)

It is hypothesized that the forensic footwear comparison and decision process is influenced by several factors. This hypothesis suggests that examiner conclusions and interpretations are appropriately studied using a *rough set technique*, which is a method that attempts to derive meaningful decision rules from an information/decision table, providing explanations of embedded trends using an intuitive and comprehensible form of *"if x, then y"* statements. Fortunately, rough set theory does not rely on strict model assumptions such as normality or *a priori* information, other than that the input data is representative of the real world [25]. However, the classical rough set approach (CRSA) is primarily used with categorical data, and this approach does not take into account scaled attributes or those that are preference-ordered [25, 26] such as a footwear examiner's preference for high quality and complete

(versus low quality and partial) impressions. Conversely, scaled or preference attributes (also referred to as *criteria*) can be handled by the *dominance-based rough set approach* (DRSA).

It is further hypothesized that the examiner abides by several guidelines when making exemplary decisions, however, the process is complicated by numerous factors that may or may not be present at any given time depending on the details of the case, making it extremely difficult to directly state any given single decision rule let alone a dominating rule. For this reason, the idea of inferring a preference model is very attractive, as the expert need only answer questions (such as rating the degree of correspondence in wear, the similarity in class features, etc.), and provide exemplary decisions (*e.g.,* "identification," "exclusion," etc.) and through the rough set approach, information available regarding the expert's findings can be used to produce a preferential model that enables an understanding of the decision maker's reason(s) for his or her choice(s).

Formally, each questioned-test impression comparison in the hands of a single examiner becomes an object or case that can be described in an information table. Each case has numerous characteristics, which can be categorized as regular attributes and criteria. Attributes are features without a preference-ordered range of values, whereas criteria have ranked values [27]. The information table is a 4-tuple $S = \langle U, Q, V, f \rangle$, where $U = \{x_1, x_2, \ldots, x_k\}$ is the finite set of objects and $Q = \{q_1, q_2, q_3, \ldots, q_p\}$ is the finite set of features. The set Q is usually divided into set C of conditional features and set D of decision attributes. If Q includes the decision attribute, then the information table is called a *decision table*. Conditional attributes in C refer to the features considered to reach each outcome, and set D contains the decision attribute d (*i.e.,* $D = \{d\}$), and $d = \{Cl_1, Cl_2, \ldots, Cl_t\}$. This means that a decision maker should assign one and only one decision to a case.

Indiscernibility Relation: Regular attributes that do not involve ranked values can be divided into two categories: qualitative and quantitative attributes. For qualitative features such as expert certification status, the similarity between cases can be assessed using an indiscernibility relation denoted by I where $P^{=}$ is a subset of qualitative attributes from C, x and y are objects in the dataset, and q_i denotes an attribute within $P^{=}$. Eq. 5 indicates that two objects (x and y) are deemed the same or *indiscernible* when they have the same value with respect to a specific variable q_i in a feature set $P^{=}$.

$$xIy \Leftrightarrow [f(x,q) = f(y,q)] \text{ for all } q_i \text{ in } P^{=} \tag{5}$$

Similarity Relation: The concept of indiscernibility can be extended to account for situations in which *small differences* in the available information are deemed *meaningless*. For example, is having four randomly acquired characteristics in agreement significantly better than having three? Perhaps, but only if the fourth is complex in geometry, whereas little value may be gained from a fourth characteristic (versus three) when the fourth is both small in size and simple in shape. Thus, establishing an association between two instances using a similarity relation (R) is very useful, whereby objects are treated as analogous when the quantitative descriptor differs less than a given percentage (ε) [27, 28]. According to Eq. 6, q_i is the numerical measure by which the instances are compared and is part of set $P^{\sim}$ which is typically a different/separate subset of attributes from $P^{=}$ in C.

$$yRx \Leftrightarrow \frac{|f(y,q) - f(x,q)|}{f(y,q)} \leqslant \varepsilon \text{ where } 0 < \varepsilon < 1 \text{ for all } q_i \text{ in } P^{\sim} \tag{6}$$

The statement yRx implies directionality, where y is the subject and x is the referent (*i.e.,* yRx means *"y is similar to x"*), and is not equivalent to xRy because the converse is not necessarily true. Therefore, $R^{-1}(x)$ (Eq. 9) can be considered as a group of objects *to which* x *is similar*, where x is the subject and y is the referent; Eq. 6 can then be modified to reflect xRy as shown in Eq. 7.

$$xRy \Leftrightarrow \frac{|f(x,q) - f(y,q)|}{f(x,q)} \leqslant \varepsilon \text{ where } 0 < \varepsilon < 1 \text{ for all } q_i \text{ in } P^{\sim} \tag{7}$$

Thus, each object can be represented by two similarity classes, where $R(x)$ represents a set of objects similar to x and $R^{-1}(x)$ contains a set of objects to which x is similar as shown in Eq. 8 and Eq. 9, respectively. Note that if ε is a small number, overlap and symmetry within and between $R^{-1}(x)$ and $R(x)$ may be found.

$$R(x) = \{y \in U : yRx\} \tag{8}$$
$$R^{-1}(x) = \{y \in U : xRy\} \tag{9}$$

An example calculation for a similarity relation can be defined according to Eq. 10 where $\epsilon_j(y_j) = \alpha_j y_j + \beta_j$ such that $C_j(x,y)$ equals 0 when there is no evidence of similarity, and 1 when there is evidence of similarity [29].

$$C_j(x,y) = \begin{cases} 1 & \text{if } |x_j - y_j| \leq \epsilon_j(y_j) \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Dominance Relation: Using indiscernibility and similarity relations to determine the association between decisions would be inappropriate if there are ordinal attributes included in the dataset, as neither relation can adequately bring to light the differences in the decision maker's preference. Therefore, it would be more applicable to work with the dominance relation, D, to manage criteria. Let $P^{>}$ be the appropriate subset of criteria in C and the dominance relation D is defined for each pair of cases x and y according to Eq. 11:

$$xDy \Leftrightarrow f(x,q) \succeq f(y,q) \text{ for all } q_i \text{ in } P^{>} \tag{11}$$

In this expression, object x dominates object y, which means x is *at least as good as* y for all criteria q_i in $P^{>}$. This may also mean that x and y have similar or identical description on all considered attributes, and therefore, x should be assigned to a class that is *not less certain/similar* than y; otherwise, the pair (x,y) is said to be *inconsistent* with the dominance principle. Following this definition, xDy does not necessarily imply that yDx, as it is possible for y to have less positive evaluations than x on particular criteria.

2.4.L Comprehensive Relation

The indiscernibility, similarity and dominance relations can be combined into four comprehensive relations indicated by the notation S_P [27]. Eq. 12 is interpreted as follows: for any pair of objects $x, y \in U,\ x$ is indiscernible from y, y is similar to x, and x dominates y. Eq. 13 is similar to Eq. 12 with the only difference that x is similar to y when quantitative attributes are considered. On the other hand, Eq. 14 applies to instances such that y is similar to and dominates x, whereas Eq. 15 describes cases where x is similar to y and y dominates x.

$$xS_Py \Leftrightarrow [xIy \text{ for each } q \in P^{=},\ yRx \text{ for each } q \in P^{\sim},\ \text{and } xDy \text{ for each } q \in P^{>}] \quad (12)$$

$$xS^*_Py \Leftrightarrow [xIy \text{ for each } q \in P^{=},\ xRy \text{ for each } q \in P^{\sim},\ \text{and } xDy \text{ for each } q \in P^{>}] \quad (13)$$

$$yS_Px \Leftrightarrow [xIy \text{ for each } q \in P^{=},\ yRx \text{ for each } q \in P^{\sim},\ \text{and } yDx \text{ for each } q \in P^{>}] \quad (14)$$

$$yS^*_Px \Leftrightarrow [xIy \text{ for each } q \in P^{=},\ xRy \text{ for each } q \in P^{\sim},\ \text{and } yDx \text{ for each } q \in P^{>}] \quad (15)$$

These overall relations then make up corresponding information granules which describe a combination of attributes that can be used to group similar objects and reduce redundant information. $D_P^{U^-}(x)$ and $D_P^{L^-}(x)$ are called P-dominated sets whereas $D_P^{L^+}(x)$ and $D_P^{U^+}(x)$ are known as P-dominating sets [27].

$$D_P^{U^-}(x) = y \in U : xS_Py \quad (16) \qquad D_P^{L^+}(x) = y \in U : yS_Px \quad (18)$$

$$D_P^{L^-}(x) = y \in U : xS^*_Py \quad (17) \qquad D_P^{U^+}(x) = y \in U : yS^*_Px \quad (19)$$

For any subset of attributes $P \subseteq C$, an instance x is said to belong to $Cl_t^{\geqslant}$ **with certainty** (Figure 15) if $x \in Cl_t^{\geqslant}$ with respect to P, and for each $y \in U$ dominating x on $P^{>}$ and indiscernible from x on $P^{=}$, and x is similar to y on $P^{\sim}$, y must also belong to $Cl_t^{\geqslant}$ (*i.e.,* $D_P^{L^+}(x) \subseteq Cl_t^{\geqslant}$). Therefore, the observation x constitutes what is termed the P-lower approximation, denoted as $\underline{P}(Cl_t^{\geqslant})$, as illustrated in Eq. 20 [27].

$$\underline{P}(Cl_t^{\geqslant}) = \{x \in U : D_P^{L^+}(x) \subseteq Cl_t^{\geqslant}\} \quad (20)$$

However, when an observation $x \in U$ is classified to an upward union of classes $Cl_t^{\geqslant}$ for $t = 2, \ldots, n$, this can create an inconsistency in the dominance principle if either one of the following two conditions holds [27].

1. x belongs to class Cl_t or better ($Cl_t^{\geqslant}$) although there is another case y that has better conditional evaluation but belongs to a class worse than $Cl_t^{\geqslant}$ (*i.e.,* $x \in Cl_t^{\geqslant}$, but $D_P^{L^+}(x) \cap Cl_{t-1}^{\leqslant} \neq \varnothing$); or
2. x belongs to a worse class than Cl_t even though there is another instance y that has better conditional evaluation but belongs to class $Cl_t^{\geqslant}$ or better (*i.e.,* $x \notin Cl_t^{\geqslant}$, but $D_P^{L^-}(x) \cap Cl_t^{\geqslant} \neq \varnothing$).

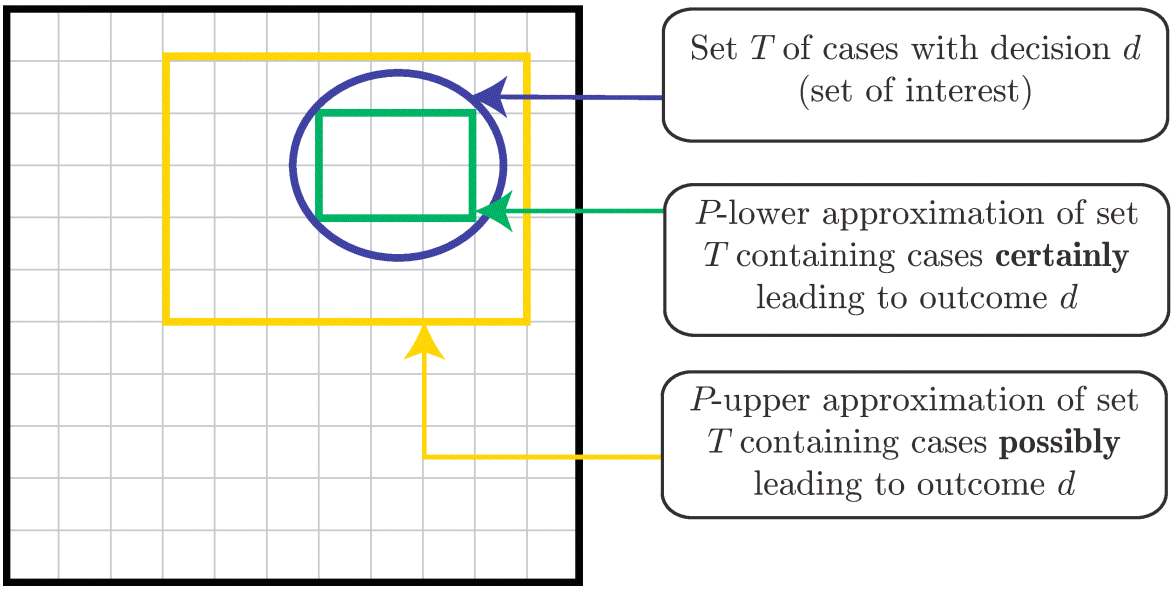

Figure 15: Defining lower and upper approximations of set T with respect to P. Note that the yellow set includes cases with the same description that *possibly* lead to other outcomes, in addition to the ones sharing the same outcome. Image courtesy of Madonna Nobel.

Should either one of the two conditions hold, it is said that x can be classified into $Cl_t^{\geqslant}$ with some ambiguity for $P \subseteq C$ (*i.e.,* the P-upper approximation, denoted as $\overline{P}(Cl_t^{\geqslant})$, comprises the collection of observations that can **possibly** (Figure 15) be included in an upward union of classes, $Cl_t^{\geqslant}$), shown in Eq. 21.

$$\overline{P}(Cl_t^{\geqslant}) = \{x \in U : D_P^{U^-}(x) \cap Cl_t^{\geqslant} \neq \varnothing\} = \bigcup_{x \in Cl_t^{\geqslant}} D_P^{U^+}(x) \text{ for } t = 1, \ldots, n \tag{21}$$

This means that in addition to the set of objects in the lower approximation, $y \in U$ *could belong* to $Cl_t^{\geqslant}$ if there is at least one $x \in Cl_t^{\geqslant}$ such that y dominates x on $P^{>}$, is indiscernible from x on $P^{=}$ and is similar to x on $P^{\sim}$ (*i.e.,* $y \in D_P^{U^+}(x)$) [27]. Analogously, the P-lower and P-upper approximations of $Cl_t^{\leqslant}$ can be defined according to Eq. 22 and 23 [27].

$$\underline{P}(Cl_t^{\leqslant}) = \{x \in U : D_P^{L^-}(x) \subseteq Cl_t^{\leqslant}\} \tag{22}$$

$$\overline{P}(Cl_t^{\leqslant}) = \{x \in U : D_P^{U^+}(x) \cap Cl_t^{\leqslant} \neq \varnothing\} = \bigcup_{x \in Cl_t^{\leqslant}} D_P^{U^-}(x) \text{ for } t = 1, \ldots, n \tag{23}$$

All the cases *possibly* belonging to upward and downward unions of classes (denoted as $Cl_t^{\geqslant}$ and $Cl_t^{\leqslant}$, respectively) form the P-boundary regions of $Cl_t^{\geqslant}$ and $Cl_t^{\leqslant}$ [27].

$$Bn_P(Cl_t^{\geqslant}) = \overline{P}(Cl_t^{\geqslant}) - \underline{P}(Cl_t^{\geqslant}) \quad (24) \qquad Bn_P(Cl_t^{\leqslant}) = \overline{P}(Cl_t^{\leqslant}) - \underline{P}(Cl_t^{\leqslant}) \quad (25)$$

The approximations of upward and downward unions of classes serve to induce a generalized description of the cases in an information table in terms of *"if ..., then ..."* statements, known as decision rules. It is important to note that these statements are not intended to imply causation, but serve as a semantic and comprehensive way of relating the features of a case to the conclusion assigned by the expert [30]. However, once the rule has been formed, important insight and quantitative metrics of accuracy and consistency can be determined.

Decision rules induced under a hypothesis that cases belonging to $\underline{P}(Cl_t^{\geqslant})$ suggest a *certain* assignment of the case to *"at least class $Cl_t^{\geqslant}$ or more certain/similar"* [27]. Similarly, rules made under a hypothesis that cases belonging to $\overline{P}(Cl_t^{\geqslant})$ will *possibly* be classified to *"at least class $Cl_t^{\leqslant}$ or more certain/similar"* [27]. On the other hand, if a rule is generated under a hypothesis that observations belong to the intersection $\overline{P}(Cl_s^{\leqslant}) \cap \overline{P}(Cl_t^{\geqslant})$, then the result would be classified as *approximately* belonging to a class between Cl_s and Cl_t with $s < t$ [27]. Note that these rules account for some inconsistent cases that belong to the boundary regions of each class. In summary, there are a maximum of five types of decision rules that can be induced from the decision table [27]:

1. *Certain* $D_{\geqslant}$-decision rules, providing descriptive profiles of cases belonging to $\underline{P}(Cl_t^{\geqslant})$;
2. *Possible* $D_{\geqslant}$-decision rules, providing descriptive profiles of cases belonging to $\overline{P}(Cl_t^{\geqslant})$;
3. *Certain* $D_{\leqslant}$-decision rules, providing descriptive profiles of cases belonging to $\underline{P}(Cl_t^{\leqslant})$;
4. *Possible* $D_{\leqslant}$-decision rules, providing descriptive profiles of cases belonging to $\overline{P}(Cl_t^{\leqslant})$; and
5. *Approximate* $D_{\geqslant\leqslant}$-decision rules, providing descriptive profile of cases belonging to the boundaries.

The left hand side (LHS) of each rule describes dominance, indiscernibility, and/or similarity of a subset of quantitative attributes. Conversely, the right hand side (RHS) reports the class assignment [27]. Therefore, a $D_{\geqslant}$-type decision rule, which is generated from the lower approximation $\underline{P}(Cl_t^{\geqslant})$ would have the form shown in Eq. 26, where $r_q \in V_q$ is a particular value for a certain feature, $\wedge$ is a logical connector for *and*, $\sim$ implies a unidirectional similarity relationship, and $\Rightarrow$ represents *then* [30] (in this example, q_a is a criterion, q_b is a qualitative attribute and q_c a quantitative feature).

$$\text{If } (f(x,q_a) \geqslant r_{qa}) \wedge (f(x,q_b) = r_{qb}) \wedge (f(x,q_c) \sim r_{qc}) \Rightarrow x \in Cl_t^{\geqslant} \quad (26)$$

An example of a decision rule that may be induced in this form is illustrated in Eq. 27, with this fictitious example translating as follows: *"If a crime scene impression is at most of medium quality, the examiner's laboratory uses the SWGTREAD scale, and there are about 3 corresponding RACs, then the examiner will conclude the case at most as having a high degree of association."*

$$\text{If } f(x,q_a) \leqslant \text{Medium} \wedge f(x,q_b) = \text{SWGTREAD} \wedge f(x,q) \sim 3, \Rightarrow x \in Cl_6^{\leqslant} \quad (27)$$

2.4.M Quality of Decision Rules

The quality of a decision rule for a case x can be assessed using several quantitative measures referred to as the *support* ($supp_x$), *strength* (σ_x), *certainty factor* (cer_x), and *lift factor* or *coverage* (cov_x), all of which are defined in Equations (28) through (31) [30, 31] where $|\cdot|$ represents the cardinality of a set, C refers to the condition part of the rule, and D refers to the decision part of the rule.

The *support* of a rule is essentially the number of cases that match both the condition and decision parts of the rule. This variable forms the numerator for the remaining computations. *Strength* refers to the ratio of the support to the number of cases considered in the decision table [25, 27]. Extremes of this value can signify *unusual and interesting* rules for further investigation. The *certainty* of a rule, also known as the *confidence ratio*, is interpreted as a *conditional probability* $p(D(x)|C(x))$ that a case x will lead to a particular conclusion $D(x)$ given its description by a given condition $C(x)$. In other words, the confidence ratio is a *measure of certainty* to which the condition implies the decision. Finally, the *lift factor* or *coverage* is a *conditional probability* $p(C(x)|D(x))$ that a case x matches the condition part of the rule, given that it has a particular conclusion $D(x)$.

$$supp_x(C,D) = |C(x) \cap D(x)| \quad (28) \qquad cer_x(C,D) = \frac{supp_x(C,D)}{|C(x)|} \quad (30)$$

$$\sigma_x(C,D) = \frac{supp_x(C,D)}{|U|} \quad (29) \qquad cov_x(C,D) = \frac{supp_x(C,D)}{|D(x)|} \quad (31)$$

In the context of this work, each case study was evaluated by multiple examiners (the decision makers), and each examiner's decisions (DM_p) is inferred to be a function of observed features (q_1, q_2, q_3, etc.). Each q_x can be either an attribute or a criterion, collected through prompting during during the comparison process. The experts' conclusions form the corresponding decision attribute set, preferentially ordered according to the SWGTREAD (2013) conclusion standard [1] from most similar (identification) to least similar (exclusion).

2.4.N DRSA Implementation for Footwear Examinations

Although the theoretical underpinnings of DRSA exist within the literature, its coded implementation invariably includes liberties. Initially, it was hoped that the footwear dataset could be analyzed using existing and open-source code. Unfortunately, the complexity of this dataset (as well as the combination of variables) could not be analyzed in this manner. Thus, an `R-Shiny` DRSA graphical user interface (GUI) was contracted. The resulting implementation is the intellectual property of Dr. Endre Palatinus. Validation efforts were conducted by the WVU research group, using nine (9) published examples (see Appendix A.3 for details).

An information/decision table constructed based on 70 examiners each forming 12 decisions ($70 \times 12 = 840$ decisions). Clearly the decision rules that result from any DRSA analysis are a function of the input variables, and any number of information tables can be constructed in order to attempt to probe the dataset; one such example is illustrated in Table 17. This dataset is based on case features, while excluding examiner-attributes (such as certification, use/familiarity with the SWGTREAD 2013 conclusion standard [1], degree of education, etc.). More specifically, the case information associated with each decision included the dominance values of the outsole design, physical size, physical size of the tread elements, number of RACs marked, the average value of the similarity of the marked features, and the percent similarity of the actual features marked, forming an 840×6 information table.

Variable	Details	Type	α	β	Options
OBJECT	Name	X	0	0	NA
UNAME	Uder ID	Miscellaneous	0	0	NA
IMG	Image number	Miscellaneous	0	0	NA
Q3OD	Value of outsole design	Dominance	0	0	0-Not evaluated; 1-Insufficient detail; 2-Value for exclusion; 5-Value for association
Q3PSO	Value of physical outsole size	Dominance	0	0	0-Not evaluated; 1-Insufficient detail; 2-Value for exclusion; 5-Value for association
Q3PSD	Value of physical size of tread elements	Dominance	0	0	0-Not evaluated; 1-Insufficient detail; 2-Value for exclusion; 5-Value for association
RACS	Number of RACs marked	Dominance	0	0	Number of RACs marked
STRENGTH	Strength of marked feature(s) (average)	Dominance	0	0	2-Supports exclusion; 3-Supports indications of non-association; 4-Supports limited association; 5-Supports association of class; 6-Supports high degree of association; 7-Supports identification
SIMILARITY	Percent similarity between the feature on the known versus the feature on the questioned (average)	Dominance	0	0	Percentage
QF1	SWGTREAD conclusion for comparison	Decision	0	0	1-Lacks sufficient detail; 2-Exclusion; 3-Indications of non-association; 4-Limited association; 5-Association of class; 6-High degree of association; 7-Identification

Table 17: Table of variables for input to DRSA.

Table 18 reports the resulting decision rules as a function of the input factors described in Table 17. The first five (5) rules report the conditions that are associated with decisions of high degree of association or identification. All include a large number of RACs, and/or a high similarity in RACs between the questioned and known impressions. For example, the fourth rule has the highest coverage or $p(C(x)|D(x)$, and can be verbalized as follows: *"If any RAC is marked with a value for 'identification' and has a similarity between the known source and questioned impression of 57% or greater, it will be concluded as class 6 or greater (high degree of association or identification)."* This particular rule has a support of 76 (the number of cases that match both the condition and decision), and a strength of approximately 10% (the support versus the total number of cases or 76/840). Furthermore, the computed certainty equals 1.0 (or a probability of 1.0 that a case will lead to a decision of class 6 or greater (high degree of association or identification) given the condition). Meanwhile, the coverage equals 0.46, or an approximate probability of 0.5 that the LHS conditions will be present given a decision of class 6 or more similar. Interestingly, it is noted that of the 840 evaluations, examiners *accurately* reached high degree of association or identification a total of 133 times (or $76/133 = 0.57$). When evaluating these five (5) rules, not surprisingly, they

show adherence to the SWGTREAD 2013 conclusion standard [1]. In other words, for an examiner to conclude high degree of association or identification, one or more RACs with reasonable similarity to the known source must be present.

LHS	RHS	Support	Certainty	Coverage	Strength
STRENGTH >= 7 AND RACS >= 17	Cl7>=	1	1	0.0099	0.0012
SIMILARITY >= 83 AND RACS >= 6	Cl7>=	13	1	0.1287	0.0155
RACS >= 10 AND SIMILARITY >= 39	Cl7>=	4	1	0.0396	0.0048
SIMILARITY >= 57 AND STRENGTH >= 7	Cl6>=	76	1	0.4551	0.0905
STRENGTH >= 6 AND RACS >= 5 AND Q3PSO >= 5	Cl6>=	30	1	0.1796	0.0357
SIMILARITY >= 91	Cl5>=	67	1	0.2219	0.0798
SIMILARITY >= 38 AND STRENGTH >= 4	Cl5>=	126	1	0.4172	0.1500
STRENGTH >= 6 AND Q3PSO >= 5 AND RACS >= 4 AND Q3PSD >= 5	Cl5>=	50	1	0.1656	0.0595
SIMILARITY >= 24 AND STRENGTH >= 4	Cl4>=	127	1	0.3256	0.1512
SIMILARITY >= 16 AND RACS >= 6	Cl4>=	23	1	0.0590	0.0274
SIMILARITY >= 87	Cl3>=	80	1	0.1717	0.0952
SIMILARITY >= 13 AND STRENGTH >= 4	Cl3>=	130	1	0.2790	0.1548
SIMILARITY >= 8	Cl2>=	136	1	0.1625	0.1619
RACS >= 4	Cl2>=	133	1	0.1589	0.1583
STRENGTH >= 6	Cl2>=	136	1	0.1625	0.1619
STRENGTH >= 4 AND Q3OD >= 5	Cl2>=	166	1	0.1983	0.1976
Q3PSD <= 0 AND Q3OD <= 2	Cl3<=	2	1	0.0044	0.0024
Q3PSO <= 0 AND Q3OD <= 2	Cl3<=	2	1	0.0044	0.0024
Q3OD <= 1	Cl4<=	2	1	0.0037	0.0024
Q3PSD <= 0 AND Q3PSO <= 2	Cl4<=	13	1	0.0242	0.0155
Q3PSD <= 1 AND Q3OD <= 2	Cl4<=	2	1	0.0037	0.0024
Q3PSO <= 0 AND Q3PSD <= 2	Cl4<=	13	1	0.0242	0.0155
Q3PSD <= 1 AND Q3PSO <= 2 AND SIMILARITY <= 0	Cl4<=	20	1	0.0372	0.0238
Q3PSO <= 1 AND Q3OD <= 2	Cl4<=	4	1	0.0074	0.0048
Q3PSO <= 1 AND Q3PSD <= 2 AND SIMILARITY <= 0	Cl4<=	22	1	0.0409	0.0262
Q3PSD <= 0	Cl5<=	18	1	0.0267	0.0214
Q3PSD <= 2 AND SIMILARITY <= 44	Cl5<=	205	1	0.3046	0.2440
Q3OD <= 2 AND STRENGTH <= 5	Cl5<=	49	1	0.0728	0.0583

Table 18: Output metrics for DRSA based on input Table 17 using `R-Shiny` GUI (Palatinus, 2020).

The second to last rule has the highest support (205). In words, this rule indicates that *If the examiner selects value for exclusion for the physical size of tread elements and any marked RAC has a similarity equal to or less than 44% between the questioned and known source, it will be concluded to belong to at most class 5 (association of class) or less similar*." The resulting certainty is 1.0 and the coverage is 0.3, however, this rule is a little surprising since the condition would intuitively imply a conclusion of indications of non-association or less. In evaluating the raw data, it was determined that there were only eight (8) decisions of class 5 and value for physical tread size of exclusion, and all but one of these decisions was made by the exact same examiner, on nearly all cases (001K2, 002K1, 003K1, 005K1, 006K1, 006K2, 007K2). Instead, decisions of exclusion (class 2) were reached for 152 of these comparisons, while decisions of indications of non-association (class 3) for 14 comparisons, and limited association (class 4) for only 3 comparisons. Thus, once erroneous conditions or decisions are removed (the single examiner that reported association of class for at least one shoe *in every single case*), the resulting rules again show adherence to the SWGTREAD 2013 conclusion standard [1]. Thus, the upper limit in a dominance condition can highly influence the users

interpretation of the resulting rule. This is problematic since DRSA assumes that there are consistent exemplary (but latent) rules guide examiners, and when a single examiner's reasoning differs, interpretations become challenging. Thus, analysis of this dataset using DRSA would benefit from extensive data preprocessing (either using the rules *a posteriori* to seek out input inconsistencies as illustrated here, or *a priori* to remove what appear to be erroneous or illogical conditions or decisions). Given the fact that this dataset has 840 total decisions, 1,000 participant attributes and 3,500 impression features, this volume of data-preprocessing is an on-going exercise.

2.5 Expected Applicability

2.5.A Size, Scope & Context

To date, this is the largest footwear reliability study conducted in the United States. It is also believed to be the largest footwear reliability study conducted in the United States, or abroad, as of December 2020. Historically, there are three past studies of interest. One of the earliest reliability studies was conducted internationally by Majamaa and Ytti [32] in 1996. This study consisted of 6 simulated crime scene impressions, sent to 34 laboratories, with a 97% completion rate (responses received from 33 analysts) and using a five-level reporting scale. The second was conducted by Shor and Weisner [33] in 1999, involving 2 actual case impressions (ground truth not available), evaluated by 20 experts from 7 different laboratories, across 6 different countries (plus 3 examiners from the respective authors' own laboratory), with cases selected due to their difficulty (the questioned impressions were deemed ambiguous and controversial by experts), and each analyst was permitted to use his or her own conclusion scale when providing results. The next most recent study by Hammer et al. [34] was conducted in 2013. This study targeted International Association for Identification (IAI) certified examiners (certified as of July 2008) located in North America, each using a seven-category conclusion scale, with a total of 40 participants. Thus, this work includes nearly double the number of participants in past studies.

Using published data, Table 19 was formulated in order to compare this study with past efforts. The mean agreement for the Majamaa and Ytti [32] study was 83.8% $\pm$ 12.4% (1 standard deviation). The equivalent value for the Shor and Weisner [33] and Hammer et al. [34] studies were 78.3% and 94.3% $\pm$ 7.36%, respectively. Recall that the mean agreement for the West Virginia University (WVU) study is 85.6% $\pm$ 11.1%. Based on study design, these values match intuition; the Shor and Weisner [33] study was limited to questioned impressions deemed very challenging, and resulting in the lowest IQR. The Majamaa and Ytti [32] study was performed internationally, with limited instructions on how to interpret a prescribed scale, and with a fixed impression type (electrostatic dust lifts from paper) leading to a moderate IQR. Conversely, the Hammer et al. [34] study was limited to certified examiners in North America, and excluded class comparisons and the need for feature identification, resulting in the highest IQR.

Majamaa and Ytti (1996)	IQR Conclusions	# in IQR	Total	% in IQR
Case 1	Inconclusive → Possible	31	33	93.9
Case 2	Possible → Very Probable	31	33	93.9
Case 3	Probable → Very Probable	20	33	60.6
Case 4	Very Probable → Identification	28	33	84.8
Case 5	Identification	27	33	81.8
Case 6	Identification	27	33	81.8
Shor and Weisner (1999)	IQR Conclusions	# in IQR	Total	% in IQR
Case 1	Possible → Highly Probable	18	23	78.3
Case 2	Possible → Highly Probable	18	23	78.3
Hammer et al. (2013)	IQR Conclusions	# in IQR	Total	% in IQR
Case 1	Identification	40	40	100
Case 2	Could Have Made	39	40	97.5
Case 3	Probable	33	40	82.5
Case 4	Could Have Made	39	40	97.5
Case 5	Probable → Identification	40	40	100
Case 6	Could Have Made	35	40	87.5
West Virginia University (2019)	IQR Conclusions	# in IQR	Total	% in IQR
Case 1 K1	Exclusion → Indications	64	70	91.4
Case 1 K2	Association → Identification	63	70	90.0
Case 2 K1	Limited → Association	55	70	78.6
Case 3 K1	Association → High Degree	39	70	55.7
Case 3 K2	Exclusion	66	70	94.3
Case 4 K1	Exclusion	65	70	92.9
Case 4 K2	Identification	62	70	88.6
Case 5 K1	Exclusion → Limited	60	70	85.7
Case 6 K1	Limited → Association	60	70	85.7
Case 6 K2	Exclusion → Limited	63	70	90.0
Case 7 K1	Exclusion	68	70	97.1
Case 7 K2B	Exclusion → Limited	50	65	76.9

Table 19: IQR for former and current West Virginia University (WVU) study.

2.5.B Comparison to 2011 FBI Fingerprint Reliability Data

The second major contribution of this work is its comparison with the existing 2011 FBI fingerprint study [2]. Using ground truth (three-point conclusion standard), Figure 4 plots the 90% confidence intervals for the positive (solid lines) and negative (dashed lines) predictive value as a function of mate-prevalence. The vertical line and reported performance metrics highlight the PPV and NPV at a mate-prevalence of 62%, which corresponds to mate distribution used in the 2011 FBI fingerprint study [2]. When corrected for the same mate-prevalence the comparable footwear PPV is 99.7% (versus fingerprints at 99.8%) and NPV is 79.6% (versus fingerprints at 86.6%) [2].

Based on the actual mate prevalence used in this footwear study (31.5%) the correct predictive value varied from 94.5% for exclusions, 85.0% for identifications, and between 70.1% and 65.2% for limited associations and association of class, respectively (with all other conclusions producing PVs between these extremes). Moreover, after data transformation based on ground truth, the case study materials showed a false positive rate of 0.48%, a false negative rate of 15.6%, a (correct) positive predictive value of 98.8% and a (correct) negative predictive value of 93.3%.

The two observed false positives, resulting in a false positive rate (FPR) of approximately 0.5% (FPR = 2/418), can be compared to the same metric reported at 0.1% for fingerprint experts [2]. Similarly, the false negative rate (FNR) was found to be higher than the false positive rate (approximately 16%), with 30 decisions (out of 192 possible identifications) falsely excluding a known mated shoe. Again, as a point of comparison, this was approximately double the error rate observed for fingerprint analysts at 7.5% [2]. In total, 23 different analysts committed these 30 errors, with five examiners committing the error twice and one analyst providing three false negatives, meaning 43% of these errors were committed by six experts (5 experts $\times$ 2 errors + 1 expert $\times$ 3 errors = 13/30) which is less than 9% of all participants. As a general observation, it appears that many of the false negatives or incorrect eliminations resulted from an improper characterization of impression size wherein a size difference (physical size and/or size and spacing of outsole elements) was reported when one did not exist.

2.5.C Considerations Moving Forward

The third major contribution of this work is an examination of conclusion scales, and how this impacts reported reliability rates. While exploring these concepts, this work discussed and reported both consensus and inter-rater reliability as measures of reproducibility. Conclusion scales, training in their usage, and how they are interpreted by the *trier-of-fact* are the subject of major scientific and philosophical considerations within the field of forensic science. This work reported data using seven- four- and a three-point scale, and further discussed concepts related to interval versus ordinal categorical labels. Clearly, much additional work is needed in this area, but it is hoped that the preliminary work here will be considered within this larger context.

Lastly, footwear examiner conclusions were subjected to a dominance-based rough set approach for rule induction. Results to date indicate that the rules are very strongly dominated by data input, suggesting the need for massive data pre-processing. In addition, the resulting metrics (especially coverage) resulted in relatively low probabilities, suggesting that it is difficult to ascertain the probability of a set of conditions given a decision category. This was further supported by the construction of similarity and difference maps. In total, 3,524 features of interest were annotated by examiners across all comparisons. Thus, annotation maps were constructed in order to evaluate the frequency of features marked by examiners that reported a final SWGTREAD (2013) [1] conclusion both within (right) and outside (left) of the expected interquartile range. Inspection of each annotation map describes the frequency

of features marked by different examiners, wherein each examiner was randomly assigned a number between 1 and 70. For any location on the outsole with at least 7 marked features (10% of the 70 respondents), instead of reproducing all examiner numbers associated with the feature, an actual frequency map was plotted. Thus, the maps reveal features marked by several examiners with the total number of marks revealed by the frequency color code, while individual numbers reveal features marked by a limited or fewer number of examiners. The purpose of each map is to allow for inspection of the features deemed relevant to comparison among both groups of examiners (those reporting a conclusion in agreement with the community, and those reporting a conclusion that is inconsistent with the community IQR). In some cases, examiners reaching non-IQR conclusions marked the same features as those reaching IQR conclusions, but since these examiners arrived at non-IQR conclusions, the implication is that the total number and/or weight attributed to marked features differs. An example map is illustrated in Figure 16 for 001Q versus 001K1. Inspection of this reveals that although a total of 91% of the respondents reached a conclusion within the IQR and matching ground-truth, only 81% (57/70) justified their conclusions using comments and annotation (1 examiner came to a valid IQR conclusion, but for the wrong reason (erroneously reported a size difference), conversely, 4 examiners outside of the IQR did not note or elected not to mark any differences, and 2 examiners did not place sufficient weight on the observed differences that exist between 001Q and 001K1). Moreover, this is only considering features marked; differences and similarities in the value attributed to each feature is not assessed in these maps, but clearly any differences can only increase the variation in conditions that give rise to the same decision. Thus, this dataset is not well-formulated to determine rules and form predictions between conditions and decisions, suggesting that much additional work is needed in this area.

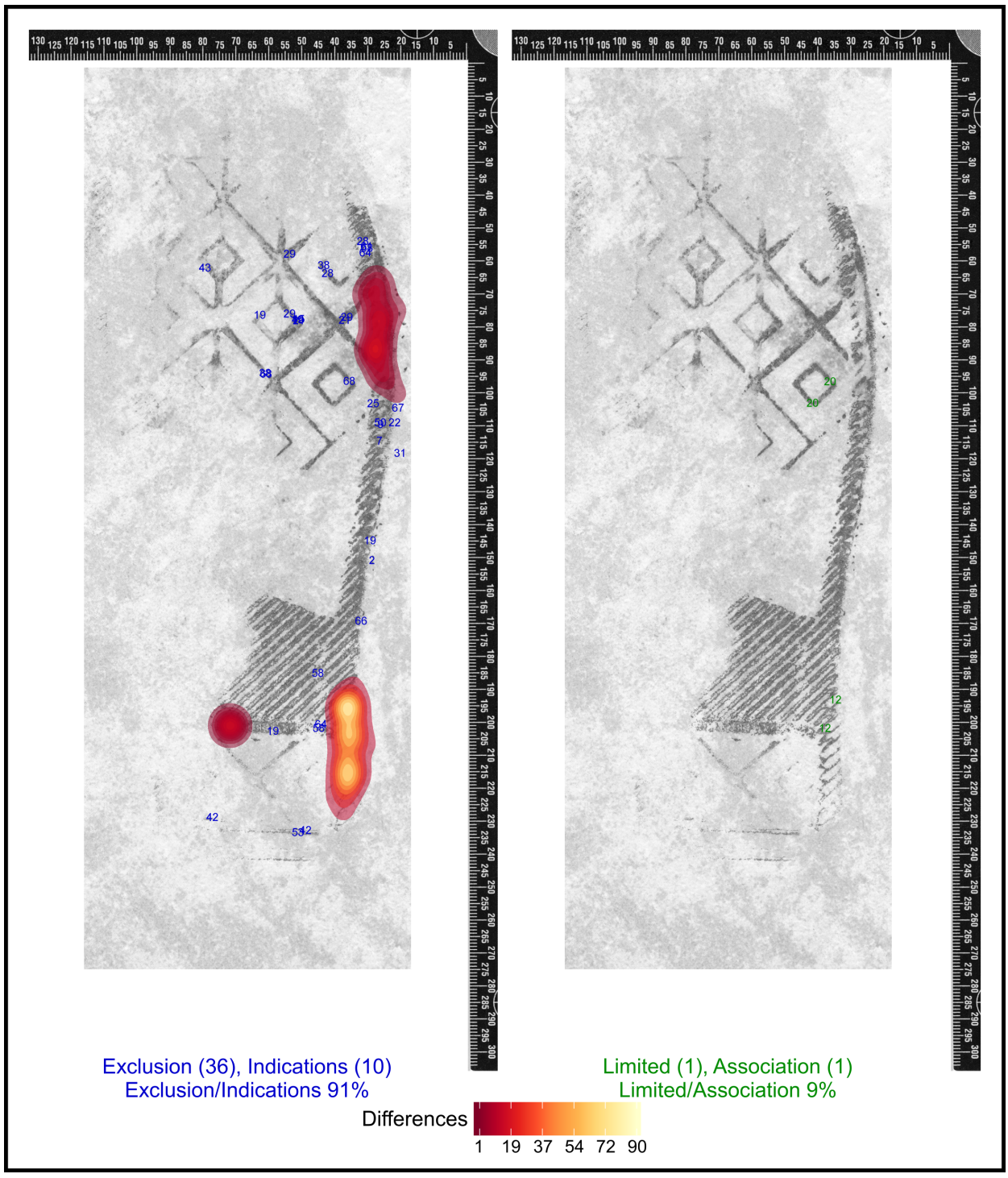

Figure 16: Difference map (red-yellow) for 001Q versus 001K1 (non-mated pair).

3. Accomplishments & Findings

The following major accomplishments and findings resulted from this work:

- Created 7 cases and 12 comparisons (consisting of 7 questioned impressions, 12 outsole knowns, and 24 Handiprint exemplars).
- Obtained IRB approval (1602021821), and enrolled and processed background surveys for 115 participants (footwear examiners).
- Created an interactive and customized graphical user interface to collect participant responses.
- Mailed packets and collected responses over the span of 19 months.
- Fully processed results across 77 participants (including 924 conclusions, 1,000 participant attributes and 3,500 impression features).
- Performed numerical and statistical evaluations on the demographics and conclusions provided from 70 participants.
- Evaluated feature identification and annotation using the customized reporting interface (results indicate considerable variation in feature identification/annotation (as low as 66.5% agreement)).
- Evaluated consensus, which ranged from a low of 0.5105 (for comparison 003Q versus 003K1), a maximum of 0.9733 (for comparison 007Q versus 007K1), with a mean of 0.7821 ± 0.1422 and a median of 0.7743.
- Evaluated inter-rater reliability (IRR) using the Gwet AC_2 agreement coefficient. The combined data set IRR was found to be 0.7509 with a standard error of 0.0875 and a 90% confidence interval of 0.6070 to 0.8948. After benchmarking, this equates with the verbal description of '*substantial*' agreement (the only higher category is referred to as '*almost perfect*.'
- Observed community agreement in conclusions via the interquartile range (IQR), which was found to equal $85.6\% \pm 11.1\%$ (median of 89.3% and a 90% confidence interval between 83.5% and 87.6%) across all comparisons.
- Compared and contrasted the observed agreement with those obtained from previous published research.
- Computed accuracy in conclusions of $82.8\% \pm 11.9\%$ (median of 85.7% and 90% confidence interval between 80.5% and 84.9%) across all comparisons.
- Compared the community agreement via IQR with accuracy and failed to detect a statistically significant difference (assuming any difference that does exist is normally distributed with a mean of zero).

- Performed chi-square tests to evaluate accuracy versus a host of examiner and case attributes as summarized below:

Dependent Tests	p-value	Independent tests	p-value
Case Difficulty v. Clarity	1.919 x 10^{-18}	Difficulty v. Accuracy	0.1397
Case v. Accuracy	0.0106	Clarity v. Accuracy	0.5769
Comparison v. Accuracy	9.637 x 10^{-12}	Education v. Accuracy	0.0966
Ground Truth v. Accuracy	1.753 x 10^{-5}	Certification v. Accuracy	0.6921
Conclusion v. Accuracy	8.957 x 10^{-16}	Cont. Education v. Accuracy	0.4838
-	-	Comparison Frequency v. Accuracy	0.8626
-	-	SWGTREAD use v. Accuracy	0.8555
-	-	# of Known Shoes v. Accuracy	0.8706
-	-	KM Provided v. Accuracy	0.0754

Table 20: Summary of global chi-square tests.

Global Test	Bonferroni Adjusted p-value
Case Difficulty v. Clarity	0.00556 (9 pairwise tests)
Case v. Accuracy	0.003571 (14 pairwise tests)
Comparison v. Accuracy	0.0020833 (24 pairwise tests)
Ground Truth v. Accuracy	0.0125 (4 pairwise tests)
Conclusion v. Accuracy	0.0041667 (12 pairwise tests)

Table 21: Summary of post-hoc chi-square tests.

- Computed correct (positive) predictive value (PV). For this dataset and mate-prevalence (31.5%), results indicate the correct predictive value varies from 94.5% for exclusions, 85.0% for identifications, and between 70.1% and 65.2% for limited associations and association of class, respectively (with all other conclusions producing PVs between these extremes).

- After data transformation based on ground truth, the case study materials show a false positive rate of 0.48%, a false negative rate of 15.6%, a (correct) positive predictive value of 98.8% and a (correct) negative predictive value of 93.3%.

- When corrected for the same mate-prevalence (62% in the 2011 FBI fingerprint study [2]) the comparable footwear PPV is 99.7% (versus fingerprints at 99.8%) and NPV is 79.6% (versus fingerprints at 86.6%) [2].

4. Artifacts & Dissemination

Major results were disseminated across a three-part series of publications in the Journal of Forensic Sciences. A fourth publication is in preparation, intended to summarize all chi-square observations of independence, and to serve as an 'interpretation piece' for both the footwear community and individual practitioners. Additional data-processing for DRSA analysis is ongoing and intended for future publication (but a manuscript is not yet in preparation). The researchers also intend to host an IAI workshop, but COVID-19 restrictions have significantly limited the opportunity to participate in 'in-person' activities, so discussions are underway to determine if/how the work might be transitioned into a virtual workshop.

Peer-Reviewed Publications

Speir, J., Richetelli, N., Hammer, L. Forensic Footwear Reliability: Part I—Participant Demographics and Examiner Agreement. Journal of Forensic Sciences, Vol. 65, No. 6, 2020, pp. 1852-1870.

Richetelli, N., Hammer, L., Speir, J. Forensic Footwear Reliability: Part II—Range of Conclusions, Accuracy, and Consensus. Journal of Forensic Sciences, Vol. 65, No. 6, 2020, pp. 1871-1882.

Richetelli, N., Hammer, L., Speir, J. Forensic Footwear Reliability: Part III—Positive Predictive Value, Error Rates, and Inter-Rater Reliability. Journal of Forensic Sciences, Vol. 65, No. 6, 2020, pp. 1883-1893.

5. Participants & Collaborating Organizations

A total of 115 footwear examiners in the United States engaged with this research. A total of 77 fully completed the study, and primary results are a function of 70 participants with case-comparison experience. In addition to the anonymous participants, the following five (5) senior personnel were responsible for the totality of the work:

Jacqueline A. Speir, Associate Professor
Principal Investigator
West Virginia University
208 Oglebay Hall, PO Box 6121
Morgantown, WV 26506-6121
P: 304.293.9233, F: 304.293.2663
E: Jacqueline.Speir@mail.wvu.edu

Nicole Richetelli
Graduate Research Assistant
West Virginia University
208 Oglebay Hall, PO Box 6121
Morgantown, WV 26506-6121
P: 304.293.4982, E: nrichete@mix.wvu.edu

Madonna A. Nobel
Graduate Research Assistant
West Virginia University
208 Oglebay Hall, PO Box 6121
Morgantown, WV 26506-6121
P: 304.293.4982, E: mnobel@mix.wvu.edu

Lesley Hammer, Forensic Scientist
External Collaborator & Consultant
10601 Prospect Drive
Anchorage, Alaska 99507
P: 907.242.0229, E: hammer.forensics@gmail.com

Endre Palatinus, Data Scientist
External Contractor
8-Max-Reger-Strass
Saarbrucken, Saarland, Germany, 66125
P: +49 157 30694233, E: palatinuse@gmail.com

A. Appendices

A.1 References

[1] SWGTREAD. Range of conclusions for footwear and tire impression examinations, 2013.

[2] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, and S.E. Fienberg. Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences of the United States of America*, 108(19):7733–7738, 2011.

[3] A. Agresti. *Categorical Data Analysis*. Wiley, 3rd edition, 2013.

[4] D. Sharpe. Your chi-square test is statistically significant: Now what? *Practical Assessment, Research & Evaluation*, 20(8), 2015.

[5] SWGTREAD. Standard for terminology used for forensic footwear and tire impression evidence. *Scientific Working Group for Shoeprint and Tire Tread Evidence*, 2013.

[6] H. Tobi, van den Berg P.B., and L.T.W. de Jon-van den Berg. Small proportions: what to report for confidence intervals? *Pharmacoepidemiology & Drug Safety*, 14(4):239–247, 2005.

[7] J. Holdren and S. Lander. Forensic science in criminal courts: Ensuring scientific validity of feature comparison methods. Technical report, President's Council of Advisors on Science and Technology (PCAST), 2016.

[8] P. Pittayapat, P. Thevissen, S. Fieuws, R. Jacobs, and Willems G. Forensic oral imaging quality of hand-held dental x-ray devices: comparison of two image receptors and two devices. *Forensic Science International*, 194(1-3):20–27, 2010.

[9] Carolyn Chen, Lee White, Timothy Kowalewski, Rajesh Aggarwal, Chris Lintott, Bryan Comstock, Katie Kuksenok, Cecilia Aragon, Daniel Holst, and Thomas Lendvay. Crowd-sourced assessment of technical skills: A novel method to evaluate surgical performance. *Journal of Surgical Research*, 187:65–71, 2014.

[10] W. J. Tastle and M. J. Wierman. Consensus and dissention: A measure of ordinal dispersion. *International Journal of Approximate Reasoning*, 45(3):531–545, 2007.

[11] Y. Akiyama, J. Nolan, M. Darrah, M.A. Rahem, and L. Wang. A method for measuring consensus within groups: An index of disagreement via conditional probability. *Information Sciences*, 345:116–128, 2016.

[12] W. J. Tastle and M. J. Wierman. An information theoretic measure for the evaluation of ordinal scale data. *Behavior Research Methods*, 38(3):487–494, 2006.

[13] R. J. Freund, W. J. Wilson, and D. L. Mohr. *Statistical Methods*. Academic Press, 3rd edition, 2010.

[14] John R. Taylor. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books, 2nd edition, 1997.

[15] K.L. Gwet. *Handbook of Inter-Rater Reliability, Four Edition, The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics, LLC, 2014.

[16] J. Onate, N. Cortes, C. Welch, and B. Van Lunen. Expert versus novice interrater reliability and criterion validity of the landing error scoring system. *Journal of Sport Rehabilitation*, 19(1):41–56, 2010.

[17] S. Andreasen, B. Backe, S. Lydersen, K. Øvrebø, and P. Øian. The consistency of experts' evaluation of obstetric claims for compensation. *BJOG: An International Journal of Obstetrics and Gynaecology*, 122(7):948–953, 2014.

[18] A. Gschließer, E. Stifter, T. Neumayer, E. Moser, A. Papp, N. Pircher, G. Dorner, S. Egger, N. Vukojevic, I. Oberacher-Velten, and U. Schmidt-Erfurth. Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity. *American Journal of Ophthalmology*, 160(3):553–560, 2015.

[19] M. Acklin and K. Fuger. Assessing field reliability of forensic decision making in criminal court. *Journal of Forensic Psychology Practice*, 16(2), 2016.

[20] M. Nawrocka, K. Frątczak, and S. Matuszewski. Inter-rater reliability of total body score-a scale for quantification of corpse decomposition. *Journal of Foensic Sciences*, 61(3):798–802, 2016.

[21] K. Lee, AA. Abdul Fatah, N. Mohd Norizan, Z. Jefrey, F.H. Md Nawi, W.F.K. Wan Nor, and et al. Inter-rater reliability of vehicle color perception for forensic intelligence. *PLOS ONE*, 14(6), 2019.

[22] SWGFAST. Standards for examining friction ridge impressions and resulting conclusions (latent/tenprint), 2013.

[23] N. Mercaldo, K. Lau, and X. Zhou. Confidence intervals for predictive values with an emphasis to case-control studies. *Statistics in Medicine*, 26(10):2170–2183, 2007.

[24] International Association for Identification (IAI). Footwear certification process, requirements & qualifications. `https://theiai.org/footwear_requirements.php`.

[25] Z. Pawlak. Rough sets. *International Journal of Computer & Information Sciences*, 11:341–356, 1982.

[26] S. Greco, B. Matarazz, and R. Słowiński. Rought approximation of a preference relation by dominance relations. *European Journal of Operational Research*, 117:63–83, 1999.

[27] S. Greco, B. Matarazz, and R. Słowiński. Rough sets methodology for sorting problems in presence of multiple attributes and criteria. *European Journal of Operational Research*, 138:247–259, 2002.

[28] R. Słowiński and Vanderpooten D. A generalized definition of rough approximations based on similarity. *IEEE Transactions on Knowledge and Data Engineering*, 12:331–336, 2000.

[29] Roman Slowinski, Instytut Informatyki, and Daniel Vanderpooten. Similarity relation as a basis for rough approximations, 1995.

[30] Z. Pawlak. Rough sets and intelligent data analysis. *Information Sciences*, 147:1–12.

[31] J.J. Liou. A novel decision rules approach for customer relationship management of the airline market. *Expert Systems with Applications*, 36:4374–4381, 2009.

[32] H. Majamaa and A. Ytti. Survey of the conclusions drawn of similar footwear cases in various crime laboratories. *Forensic Science International*, 82:109–120, 1996.

[33] Y. Shor and S. Weisner. A survey on the conclusions drawn on the same footwear marks obtained in actual cases by several experts throughout the world. *Journal of Forensic Sciences*, 44(2):380–384, 1999.

[34] Lesley Hammer, Kate Duffy, Jim Fraser, and Niamh Nic Daéid. A study of the variability in footwear impression comparison conclusions. *Journal of Forensic Identification*, 63(2):205–218, 2013.

A.2 Graphical User Interface Instructions

Footwear Examination Black Box Study

Thank you for agreeing to take part in this study aimed at understanding how experts examine and interpret footwear impression evidence. The information gathered during the course of this research will provide the forensic footwear community with greater insight regarding the comparative decision-making process. If you have any questions or comments, please contact **Jacqueline Speir** at **(304) 293-9233** or `Jacqueline.Speir@mail.wvu.edu`.

Submission deadline: Please submit case analysis results by **June 25, 2017**. The submission process is electronic (as described in Section 7, page 22 of this document), and your answers will remain completely anonymous.

Contents

1 Packet contents

1. Table 1 lists the items included in this study packet.

No.	Item	Quantity
1.	Copy of signed IRB consent form	1
2.	Letter Re: NIJ funding & IRB form changes	1
3.	New IRB consent form (for your records only)	1
4.	SWGTREAD conclusion scale	1
5.	CD-RW containing graphical user interface (GUI)	1
6.	Crime scene sample reproductions	7
7.	Outsole reproductions	12
8.	Acetate sheets	12
9.	Handiprint reproductions	24

Table 1: Instructions and case items included in this study.

2. Table 2 is the labeling key used in reference to the crime scene impression, Handiprint and outsole reproductions; the definition applies to parts in bold. Note that there are two Handiprint replicas for each known footwear to allow the examiner to estimate the range of variability in test impression reproductions.

No.	Label	Definition
1.	**001**	Case number
2.	001**Q**	Crime scene impression
3.	001**K1**/001**K2**	Known 1 / Known 2 (Outsoles)
4.	001**K1-1**/001**K1-2**	Known 1, Handiprint replica 1 / Known 1, Handiprint replica 2

Table 2: Labeling key for the case items.

3. Table 3 lists the types of substrates and media for each crime scene impression.

No.	Item	Substrate	Medium	Processing
1.	001Q	Ceramic tile	Blood	Leucocrystal Violet (LCV)
2.	002Q	Vinyl tile	Dust	Digital enhancement of gel lift
3.	003Q	Ceramic tile	Blood	Leucocrystal Violet (LCV)
4.	004Q	Linoleum tile	Wax	Magnetic powder and gel lift
5.	005Q	Vinyl tile	Dust	Digital enhancement of gel lift
6.	006Q	Paper	Dust	Digital enhancement
7.	007Q	Ceramic tile	Blood	Leucocrystal Violet (LCV)

Table 3: Substrates and media for crime scene impressions.

4. Table 4 shows the manufacturing details for each known footwear/shoe.

No.	Item(s)	Manufacturer	Style	Size	Additional Details
1.	001K1, 001K2	Converse	All Star	9	-
2.	002K1	Nike	Lebron James	10	-
3.	003K1, 003K2	Nike	Rosherun	9	Microcellular material
4.	004K1, 004K2	Nike	Air Max	10.5	-
5.	005K1	Nike	Air Max	11	-
6.	006K1, 006K2	Nike	Air Max Cage	10	-
7.	007K1	Under Armour	-	11	-
8.	007K2	Under Armour	-	10	-

Table 4: Manufacturing details for each known footwear/shoe.

5. For your convenience, digital copies of the case study reproductions are also available online at `https://tr.im/Case_Images_Download`. These copies, compressed into a .zip file, may take a minimum of 3 minutes to download and 5 minutes to extract, due to the large file size.
   (a) If you are signed in to Google Drive, right-click on the file and select **Download**.
   (b) If you are not signed in to Google Drive or do not have a Google account, left-click the **Download** icon in the top left corner (Fig. 1).
6. A CD-RW labeled with the application name ("*ExaminerReport*") and your username is enclosed. This CD contains the reporting application through which you will be able to elaborate on your opinion regarding the cases and indicate specific features on the crime scene impression and/or known footwear that helped shape your conclusions. Alternatively, the reporting application is available for download at `https://tr.im/GUI_Application_Download`. Note that it is recommended that this application be downloaded (from the CD or online) and used on a single computer.

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

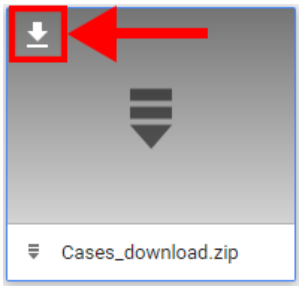

Fig. 1

2 Conducting the examination of a case study

1. Please conduct the examination of each case study as you would for normal casework according to the guidelines set by your respective laboratory and/or SWGTREAD.

2. You may annotate any of the provided materials at your discretion in order to facilitate the examination process. Acetate sheets are also provided to assist with annotation, should you wish to use them. However, please note that these sheets are not of sufficient quality for use with printers or copiers.

3. It is not necessary to document your findings in a notebook; however, making detailed notes would be helpful and is recommended so that you will be able to report and elaborate on your findings through the application.

4. *It is recommended that you complete the examination of one case and report your findings immediately thereafter using the reporting interface before proceeding to another case. The interface is designed specifically for reporting purposes only, and tailored to the goals of this study. The application should not be used as part of the examination process.*

    - The reporting application will *only* display the first Handiprint replica for each known footwear. However, please base your conclusions on the comparison between the crime scene impression and the known footwear images in totality (including the outsole and both test impression replicas).

5. The images of the outsoles and gelatin lifts have been reversed to match the orientation of the Handiprint. In other words, all prints from a *left shoe* have been oriented to look like a *left shoe*, and vice versa, for prints from a right shoe. *More specifically, the medial arch of a left shoe faces away from the ruler whereas the medial arch of a right shoe faces the ruler.* ***Therefore, the left or right orientation of a shoe will not be a reason to eliminate a known or report inconclusively.***

6. While the physical shoes will not be provided due to logistical constraints, please report your SWGTREAD conclusions based upon your confidence and experience. If your laboratory protocol requires you to report a "less certain conclusion" because the shoe is not in hand, please note this and/or explain in the reporting interface.

7. All shoes were collected immediately after the "crime" was committed; as such, ***wear differences cannot be attributed to additional usage.***

8. Please refrain from discussing and sharing the contents of the study with a fellow colleague. While the study does not involve sensitive information, the contents of your study packet have been designed specifically for you. In addition, we would like to avoid contextual bias, if possible.

9. If you have questions at any time, please do not hesitate to contact **Jacqueline Speir** at **(304) 293-9233** or `Jacqueline.Speir@mail.wvu.edu`.

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

3 Acquiring the reporting application

3.1 Retrieving the application using the CD

Note: If you do not have a CD drive, please proceed to Subsection 3.2 on page 5.

1. Insert the CD provided into the tray on your laptop or desktop. Again, please note that it is recommended that you use the application on one machine. If you are using more than one computer, please contact us before submitting the data.

2. If you are a Windows user, please follow steps (2.(a)-2.(f)). If you are a Mac user, please skip to step (3.).

   (a) Open Windows Explorer to navigate to your CD drive.

   (b) Copy the "*WINDOWS.zip*" file to your desktop.

   (c) Extract the folder in the .zip file onto your desktop (Fig. 2). Note that this process may take approximately 5 minutes to complete.

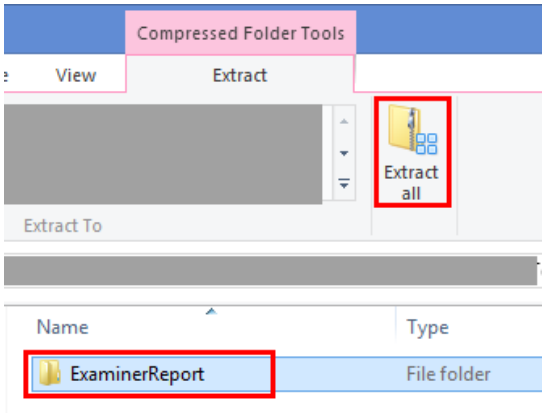

Fig. 2

   (d) You should now see a folder entitled "*ExaminerReport*."

   (e) Remove the CD from the drive and return it to the packet.

   (f) Please proceed to Section 4 on page 8 to launch the application.

3. XQuartz is an open-source graphical window environment, and is required prior to running the reporting application on a Mac machine. This package is provided to you on the enclosed CD. You will need administrator privileges to install the package onto your computer.

   (a) Using Finder, go to your CD drive. Double-click to open the "*XQuartz-2.7.9.dmg*" file.

   (b) Double-click on the "*XQuartz.pkg*" file to launch the installer.

   (c) Follow the instructions in the XQuartz installation dialog window to complete the installation (Fig. 3).

   (d) If your desktop or laptop is not compatible with XQuartz 2.7.9, please contact us for assistance.

   (e) Restart the computer to complete the XQuartz installation.

   (f) After installing XQuartz 2.7.9 onto your machine and restarting your computer, return to Finder and navigate to your CD drive.

   (g) Drag the "*MAC.zip*" file to your desktop.

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

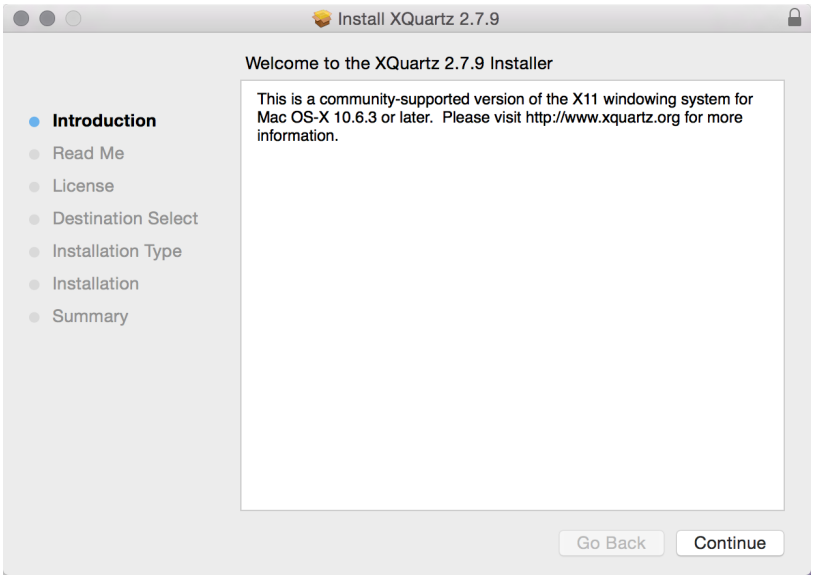

Fig. 3

(h) Double-click on the .zip file to extract the contents into a folder on the desktop. Note that the extraction process may take approximately 5 minutes to complete.

(i) Remove the CD and return it to the packet.

(j) Please proceed to Section 5 on page 9 to launch the application.

3.2 Downloading the application

1. The reporting interface can be downloaded from the following link: `https://tr.im/GUI_Application_Download`. Please note that the link is case-sensitive. Again, please note that it is recommended that you use the application on one machine. If you are using more than one computer, please contact us before submitting the data.

2. If you are a Windows user, please follow steps (2.(a)-2.(g)). If you are a Mac user, please skip to step (3.).

   (a) Download *"WINDOWS.zip"* onto the computer you will be using to submit your findings.

   (b) If you are signed in to Google Drive, right-click on the file that you wish to download and select **Download**. If you are not signed in to Google Drive or do not have a Google account, click on the **Download** icon in the top left corner as illustrated in Fig. 4.

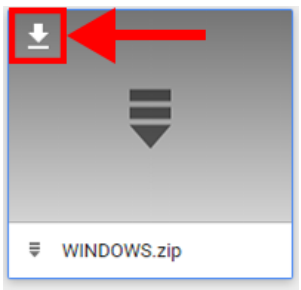

Fig. 4

   (c) A pop-up dialog box will appear to notify you that the file cannot be scanned for viruses. Click **Download anyway** (Fig. 5). You will be able to use your preferred antivirus software to scan the file later.

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

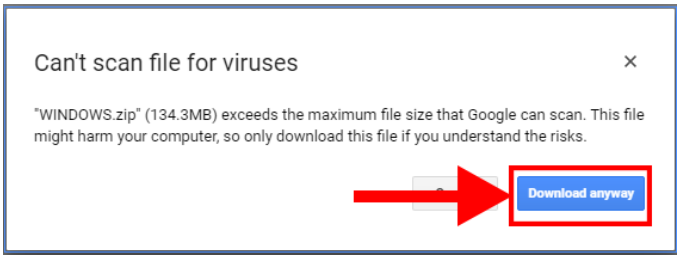

Fig. 5

(d) Copy the "*WINDOWS.zip*" file to your desktop.

(e) Extract the folder in the .zip file onto your desktop (Fig. 6). Note that this process may take approximately 5 minutes to complete.

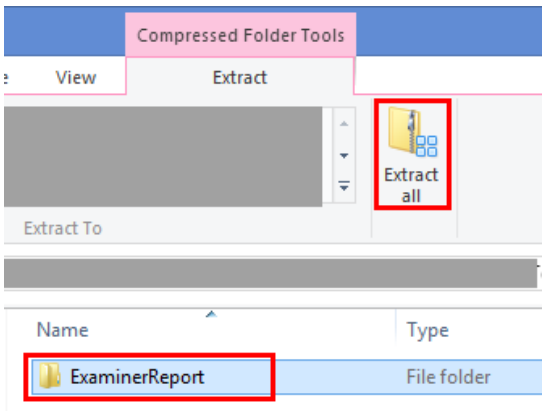

Fig. 6

(f) You should now see a folder entitled "*ExaminerReport.*"

(g) Please proceed to Section 4 on page 8 to launch the application.

3. XQuartz is an open-source graphical window environment, and is required prior to running the reporting application on a Mac machine. You will need administrator privileges to install the package onto your computer.

(a) Download both *"XQuartz-2.7.9.dmg"* and *"MAC.zip."*

(b) If you are signed in to Google Drive, right-click on the file that you wish to download and select **Download**. If you are not signed in to Google Drive or do not have a Google account, click on the **Download** icon in the top left corner as illustrated in Fig. 7.

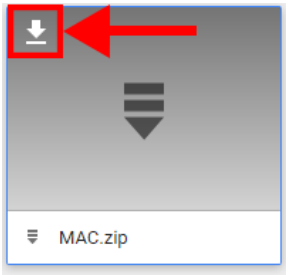

Fig. 7

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

(c) When downloading the *"MAC.zip"* file, a pop-up dialog box will appear to notify you that the file cannot be scanned for viruses. Click **Download anyway** (Fig. 8). You will be able to use your preferred antivirus software to scan the file later.

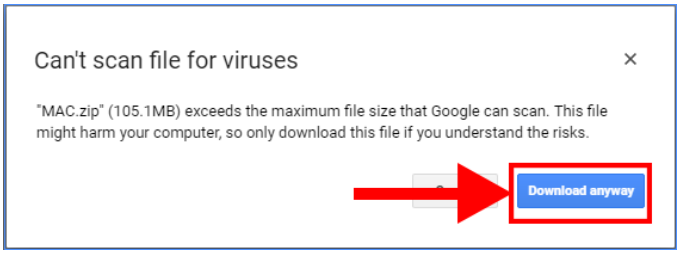

Fig. 8

(d) Open Finder and go to your *Downloads* folder.

(e) Double-click to open the "*XQuartz-2.7.9.dmg*" file.

(f) Double-click on the "*XQuartz.pkg*" file to launch the installer.

(g) Follow the instructions in the XQuartz installation dialog window to complete the installation (Fig. 9).

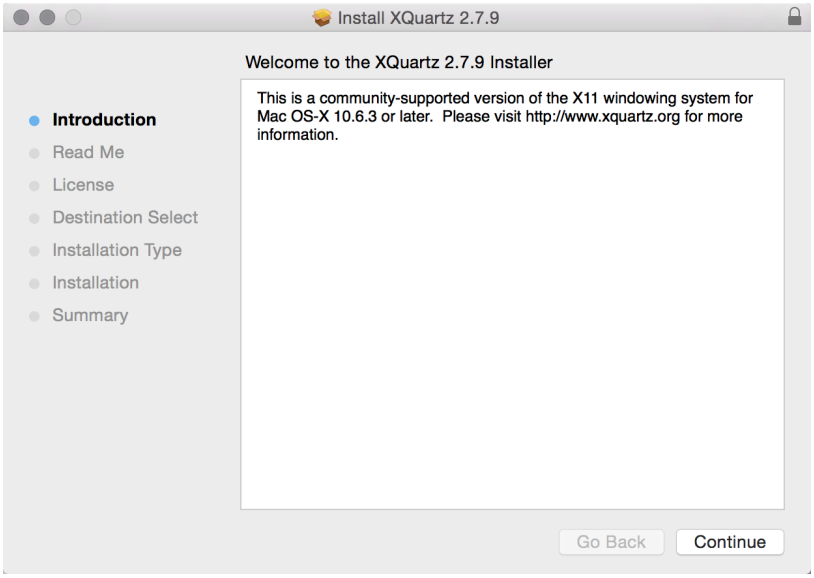

Fig. 9

(h) If your desktop or laptop is not compatible with XQuartz 2.7.9, please contact us for assistance.

(i) Restart the computer to complete the XQuartz installation.

(j) Once your computer has restarted, using Finder, go to your *Downloads* folder and copy the "*MAC.zip*" file to the desktop.

(k) Extract the folder in the .zip file onto your desktop. Note that this process may take approximately 5 minutes to complete.

(l) You should now see a folder entitled "*ExaminerReport*."

(m) Please proceed to Section 5 on page 9 to launch the application.

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

4 Launching application in Windows

Note: If you are a Mac user, please proceed to Section 5 on page 9.

1. Open the folder "*ExaminerReport*." Select the **application** "*ExaminerReport*" as shown in Fig. 10.

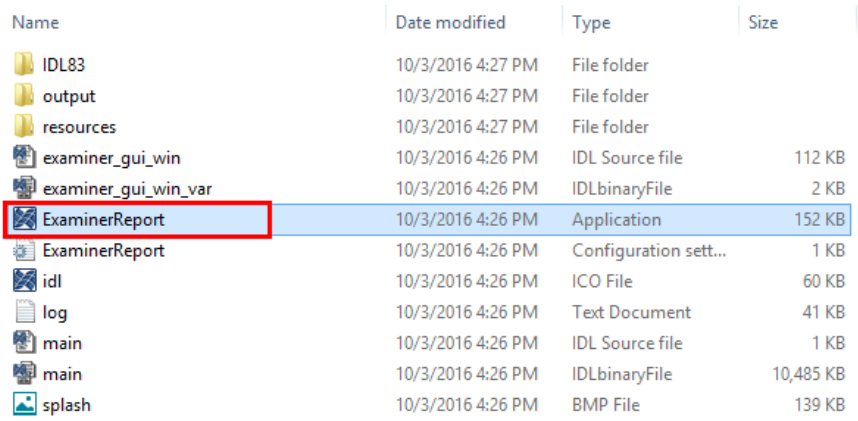

Fig. 10

Note that if you downloaded the reporting interface from the provided link, you may receive a security warning that the publisher cannot be verified. Please click **Run** to proceed.

2. This will launch a splash window as illustrated in Fig. 11. Click on the button **ExaminerReport** to proceed.

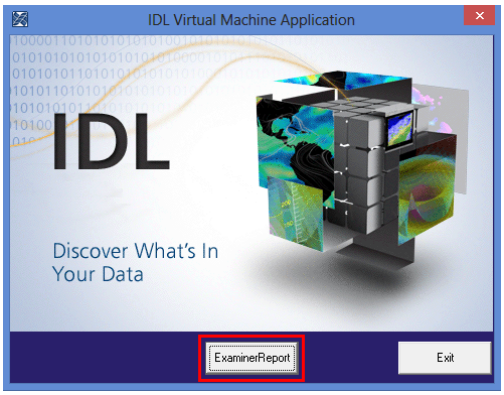

Fig. 11

3. The IDL Virtual Machine application will launch. Please click the **Click To Continue** button (Fig. 12).

Fig. 12

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

4. The "*Examiner Case Report*" window will launch, as illustrated in Fig. 13. This is the default view of the application each time it is launched (although the exact size of the window will vary as a function of your screen size).

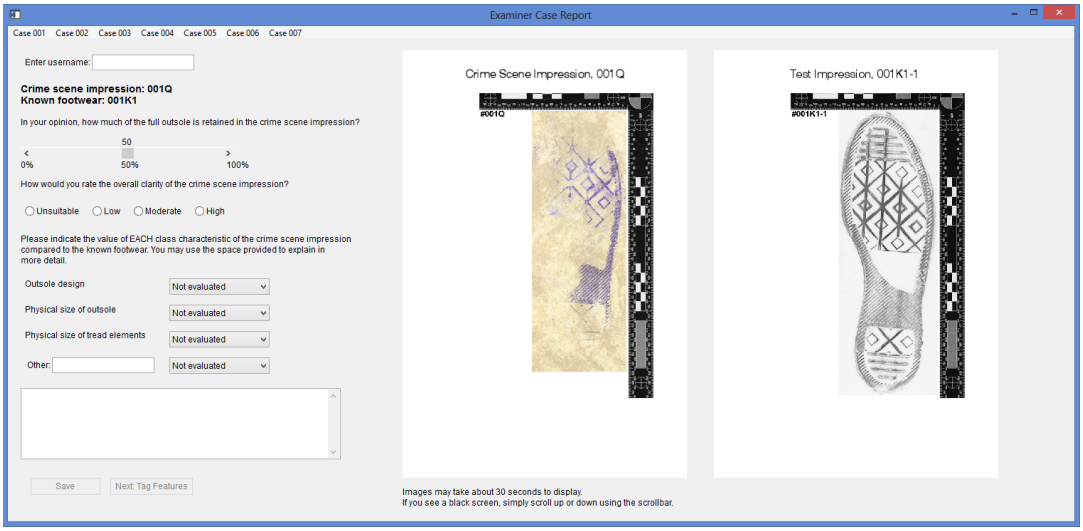

Fig. 13

5. Please proceed to Section 6 on page 11 to report your case findings.

5 Launching application in Mac

1. Open the folder "*ExaminerReport*." **Ctrl + click** on the "*ExaminerReport*" **application** (.app) file (Fig. 14). This will open a menu. Click **Open**.

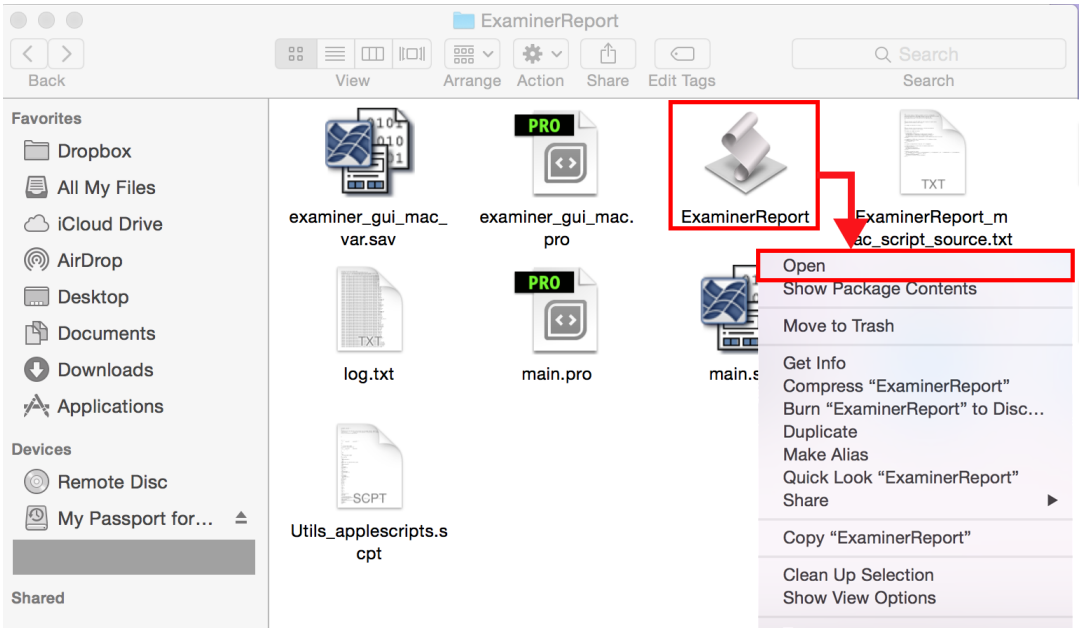

Fig. 14

2. This will prompt a dialog window, as illustrated in Fig. 15. Click **Open**.

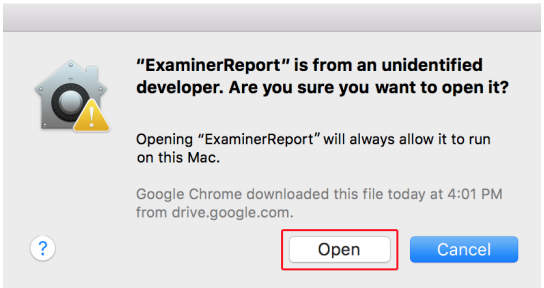

Fig. 15

3. This will launch the IDL Virtual Machine application, as illustrated in Fig. 16. Click **OK** to proceed.

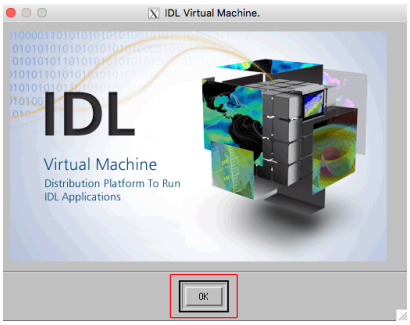

Fig. 16

4. The "*Examiner Case Report*" window will launch, as illustrated in Fig. 17. This is the default view of the application each time it is launched (although the exact size of the window will vary as a function of your screen size).

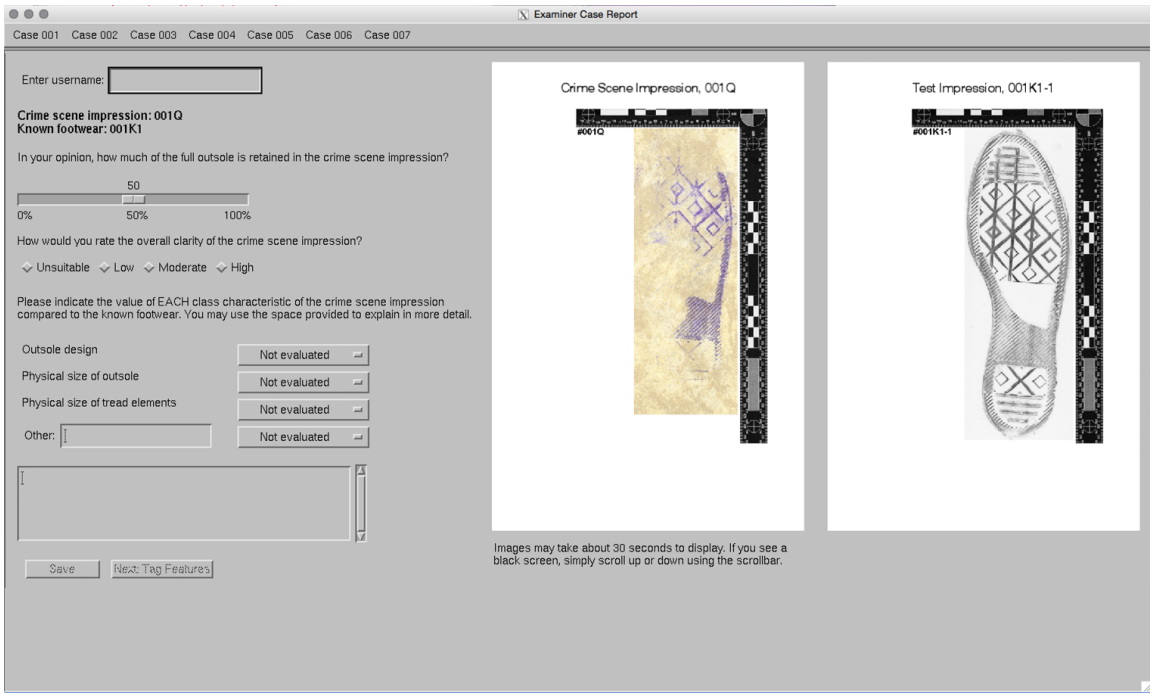

Fig. 17

5. Please proceed to Section 6 on page 11 to report your case findings.

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

6 Recording findings via reporting interface

6.1 Choosing a particular case

1. The username field must be filled prior to selecting a case. Enter your user ID in the field located at the top left of the window (Fig. 18). **This must be repeated for every case**. *Note:* Your user ID is printed on the face of the enclosed CD.

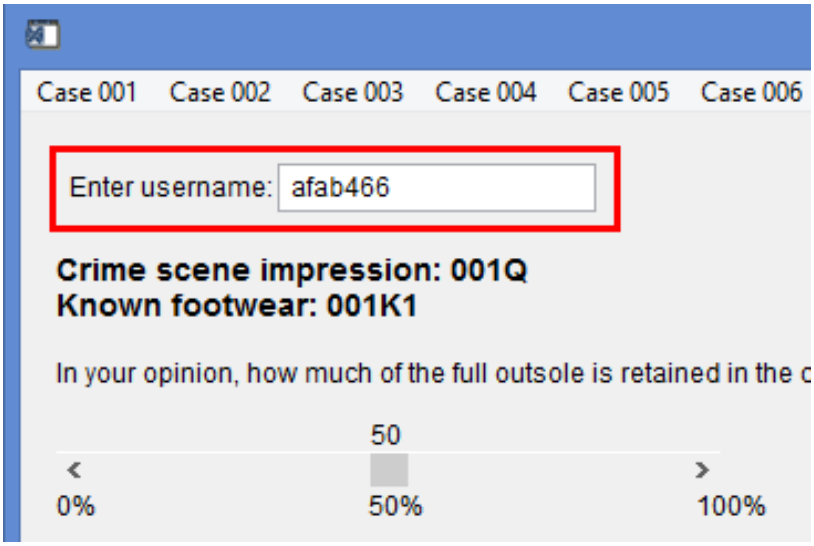

Fig. 18

2. If you wish to report for a particular case, use the menu bar to select the desired case (Fig. 19). The application is automatically set to load the test impression replica that corresponds with the *first known footwear*. Note that you will be able to report your conclusions regarding the second known *after* providing your results for the comparison between the crime scene impression and the first known footwear.

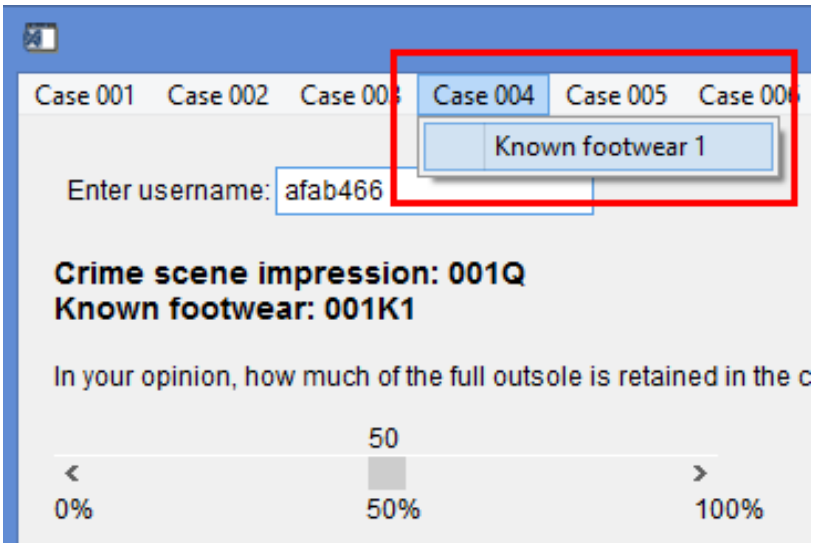

Fig. 19

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

3. The crime scene impression and first Handiprint replica for the specified case will load accordingly (Fig. 20). Please allow some time for this action to complete.

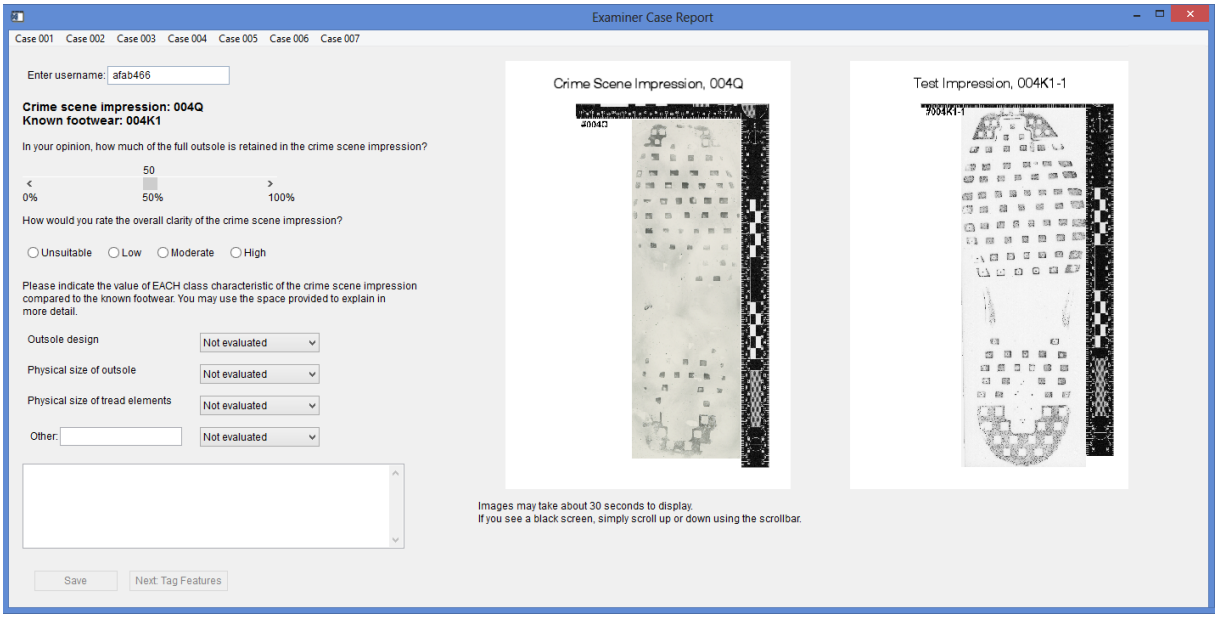

Fig. 20

6.2 Filling out the questionnaire

1. Answer questions about the crime scene impression overall. *Note:* The roller wheel of your mouse will not scroll through the reporting console; you must use the scrollbar at the bottom and/or right to move around the application window.

2. Click **Save** to save your answers (Fig. 21). *Note:* The **Save** button will be disabled once you have clicked on it, but will be re-enabled if you change your answers.

3. Click **Next: Tag Features** (Fig. 22) to proceed to the next step (Subsection 6.3), in which you will be prompted to answer questions regarding specific features you observed on the crime scene impression and/or known footwear (inclusive of the outsole and both test impression replicas).

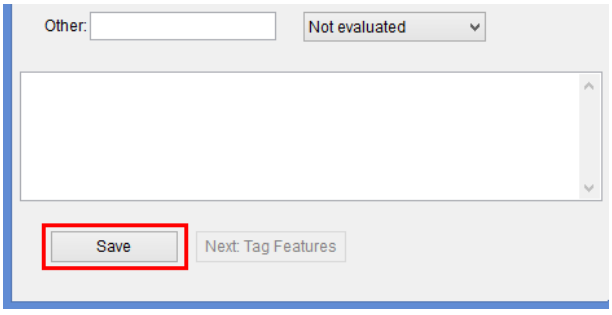

Fig. 21

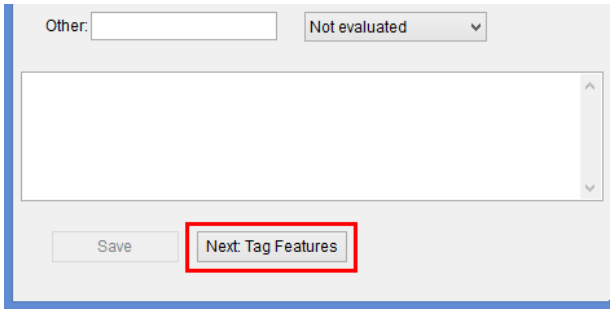

Fig. 22

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

6.3 Tagging features on crime scene and test impressions

1. Clicking on **Next: Tag Features** will take you to the screen shown in Fig. 23. The image in Panel A is the crime scene impression, and the image in Panel B is the first Handiprint test impression replica. Use the scrollbars in the markup windows to navigate around the images.

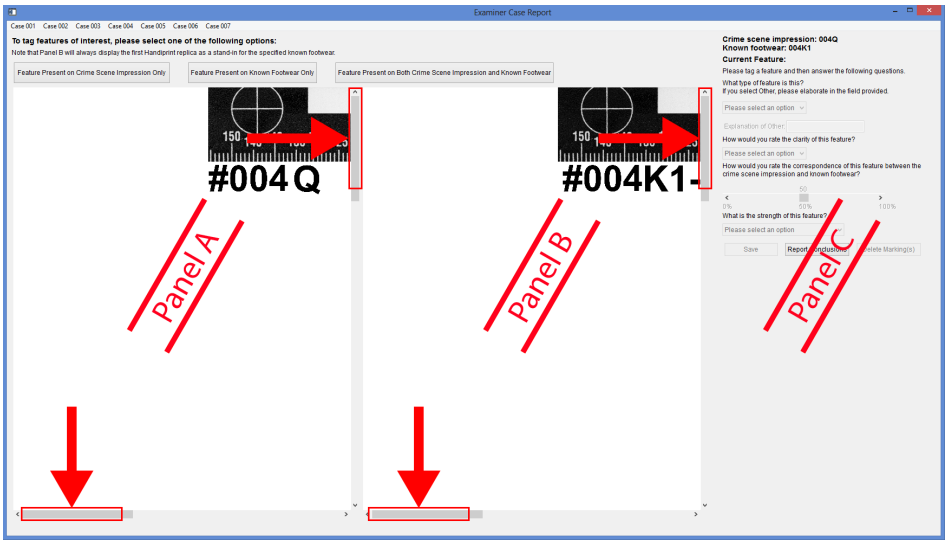

Fig. 23

2. To tag a feature of interest, please select (via a left mouse click) one of the options illustrated in Fig. 24. If there are no features to tag, please skip to page 16 of this document, step (6.).

Fig. 24

(a) If you select **Feature Present on Crime Scene Impression Only**, Panel A will activate; using a left mouse click, please tag a feature of interest on the crime scene impression.

For features present on known footwear: Panel B will always display the *first Handiprint* ***replica*** as a stand-in for the known footwear/outsole being compared. If a feature has been identified on the known footwear (or any of its associated test impressions), but does not reproduce in the *first Handiprint replica*, please proceed by marking the feature's ***general location*** in Panel B.

(b) If you select **Feature Present on Known Footwear Only**, Panel B will activate; using a left mouse click, please tag a feature of interest on the test impression replica.

(c) If you select **Feature Present on Both Crime Scene Impression and Known Footwear**, Panel A will activate *first*. Again, using a left mouse click, please tag a feature of interest on the crime scene impression. After tagging the crime scene impression, Panel B will activate. Using a left mouse click, please tag the corresponding feature on the test impression replica.

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

3. Each left mouse click will create an enumerated label on the corresponding image (Panel A, Panel B or both Panel A and Panel B, as illustrated in Fig. 25). After marking a feature of interest on the crime scene impression and/or Handiprint replica, proceed to Panel C to answer a series of questions regarding your markup(s).

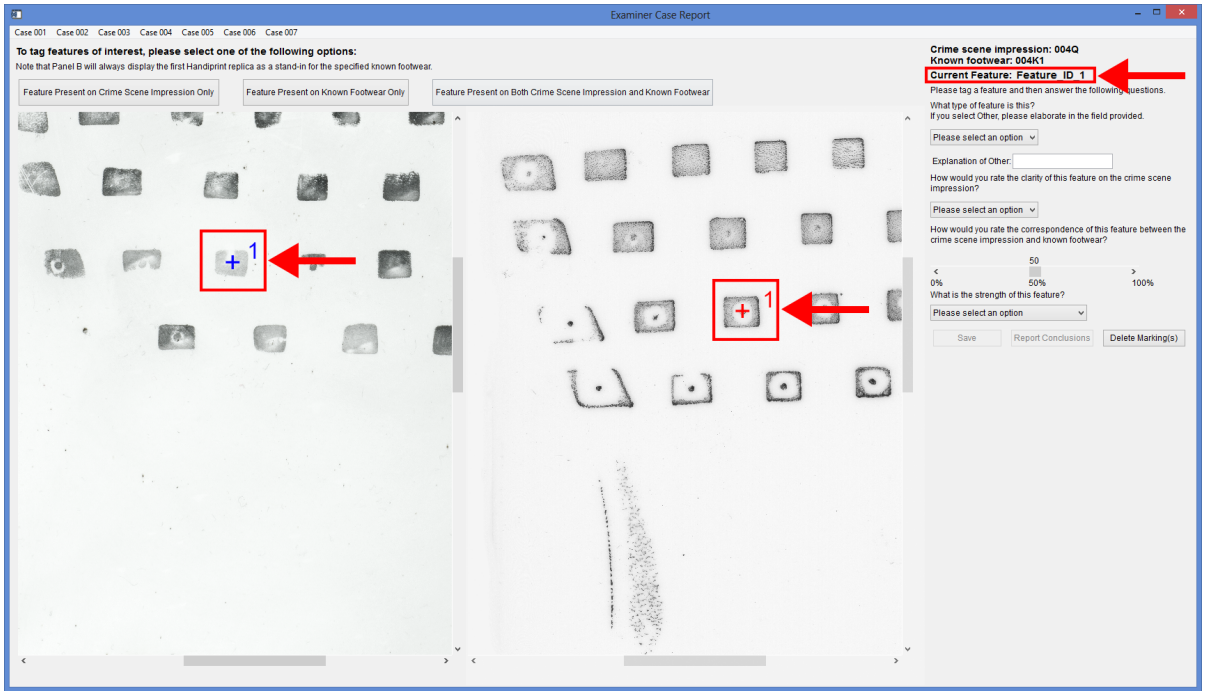

Fig. 25

4. Click **Save** to save your responses (Fig. 26).

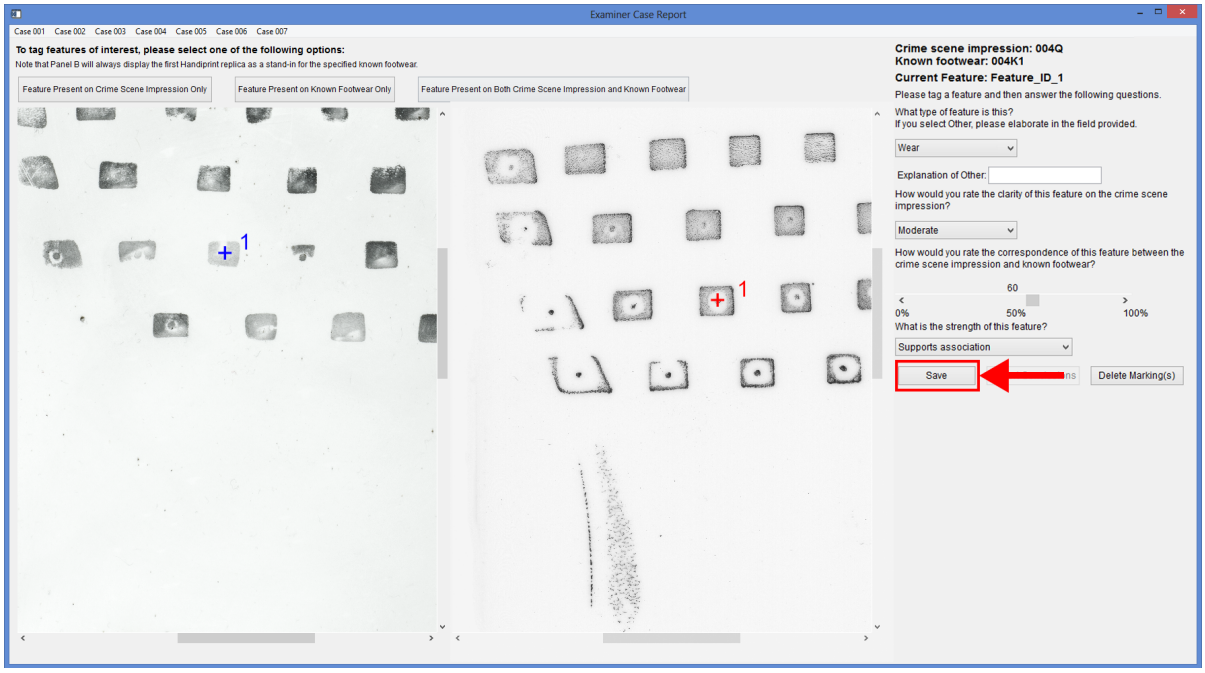

Fig. 26

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

6.3.1 Deleting marked feature(s)

(a) If at any point you wish to delete a marked feature, click **Delete Marking(s)** (Fig. 27).

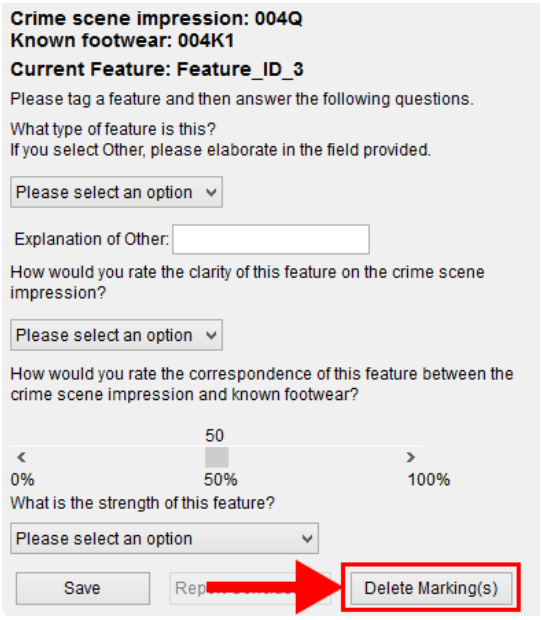

Fig. 27

(b) You will be able to select the feature(s) you wish to delete (Fig. 28). Click **Delete Selection(s)** or **Cancel** to continue.

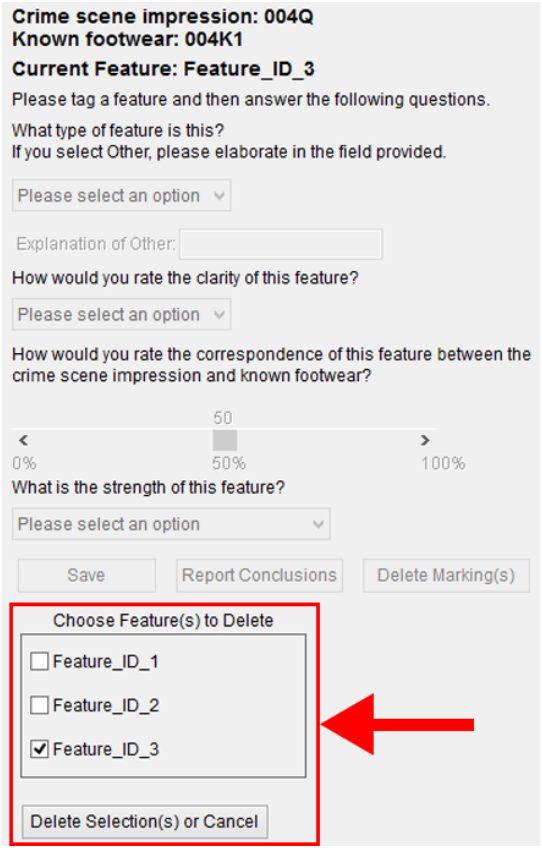

Fig. 28

(c) A dialog box will appear to confirm your action (Fig. 29). Clicking either **Yes** or **No** will allow you to return to the markup screen so you can continue marking additional features or proceed to report your conclusions.

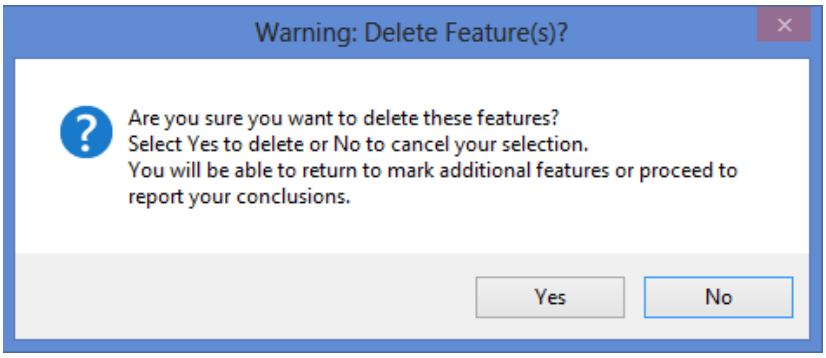

Fig. 29

5. Repeat steps (2.)-(4.) (of Subsection 6.3) to tag additional features, as necessary, and answer questions regarding each newly tagged feature.

6. Click **Report Conclusions** once you have tagged all the features of interest and answered all the questions (or if there were no features to tag) (Fig. 30). You will be taken to the last screen in the application to report your final conclusions (Subsection 6.4).

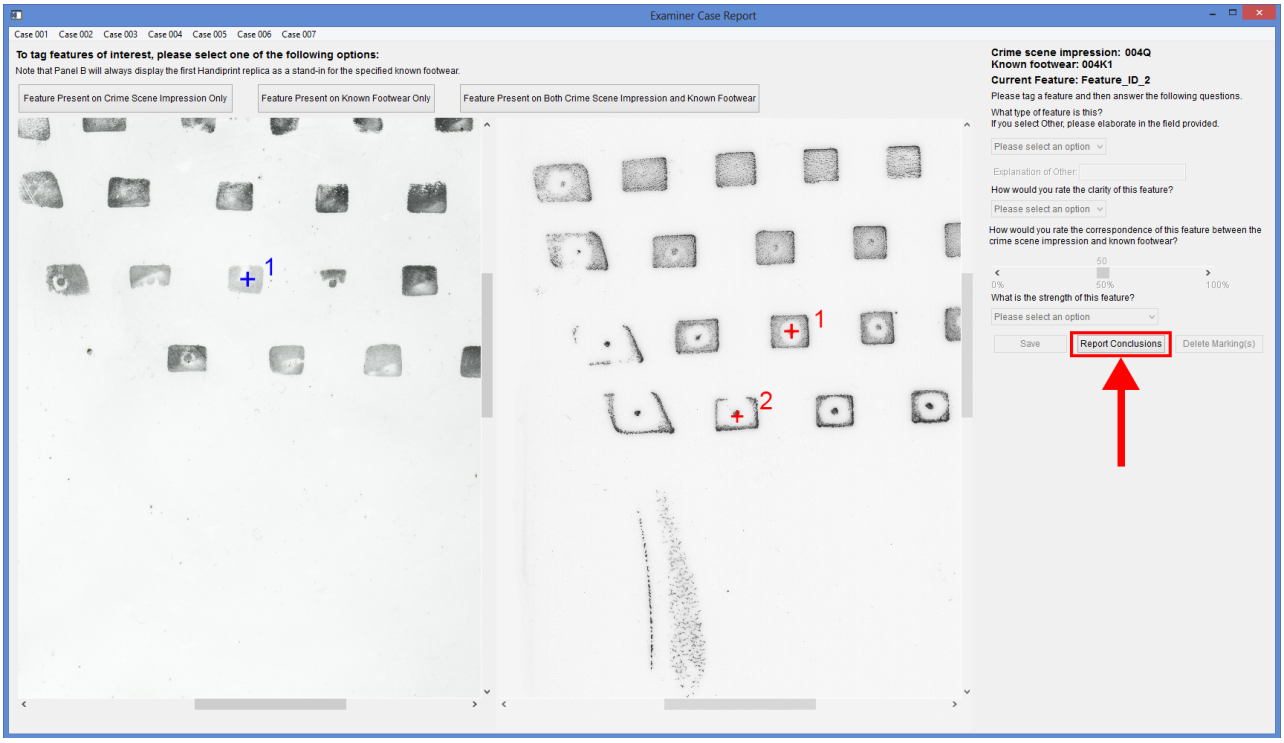

Fig. 30

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

6.4 Reporting final conclusions

1. The status message at the top of the final report screen will indicate whether an image with markup(s) has been saved, or there were no features tagged (Figs. 31 and 32, respectively).

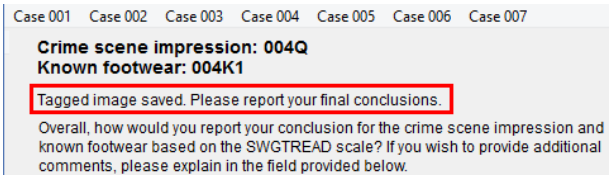

Fig. 31

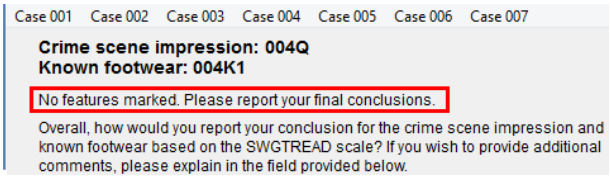

Fig. 32

2. Report your final conclusions regarding the **specified known footwear** (Fig. 33) according to the SWGTREAD scale, as well as the scale used in your laboratory/agency if there are differences in reporting verbiage. A copy of the SWGTREAD scale and how to interpret conclusions is enclosed for your convenience.

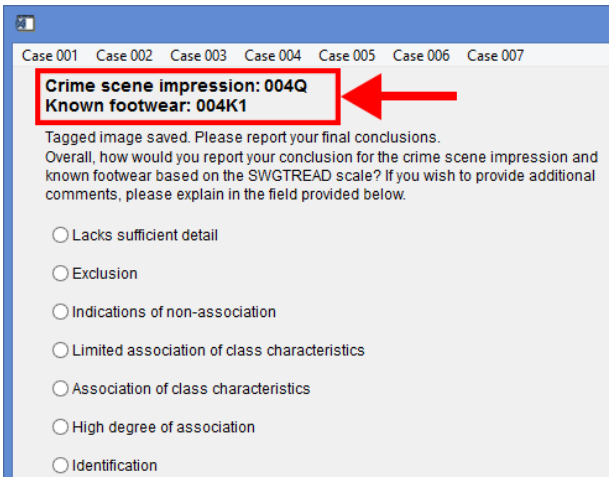

Fig. 33

3. Indicate any limitations encountered during the analysis and elaborate in the provided space, if desired.

4. If you are reporting for cases with one known (Cases 002 and 005), please follow steps (4.(a)-4.(d)). For cases with two knowns (Cases 001, 003, 004, 006 and 007), please proceed to step (5.).

   (a) Please indicate the difficulty of the case to facilitate future research studies.

   (b) When finished, please click **Save Conclusion** (Fig. 34).

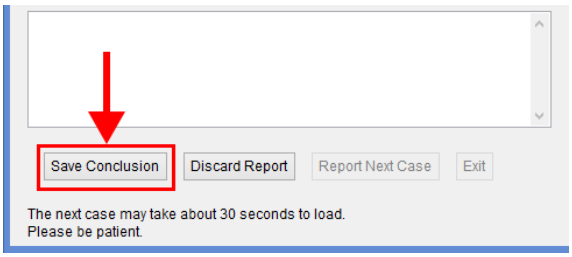

Fig. 34

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

(c) If you would like to continue reporting for a different case, click **Report Next Case** (Fig. 35). Refreshing the application may take up to 30 seconds; please be patient. *Note:* The next case does not load automatically. You must select the case according to Subsection 6.1, page 11, step (2.), Fig. 19.

(d) If you would like stop for now, click **Exit** to close the application (Fig. 36); ***please avoid closing the application using the X button in the top right corner (Windows) or top left corner (Mac).*** *If you do so accidentally, a dialog box may pop up to notify you of an error. Click* ***OK****.*

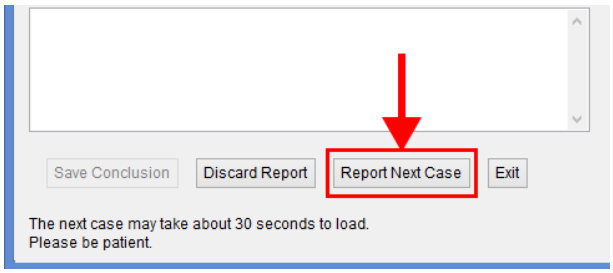

Fig. 35

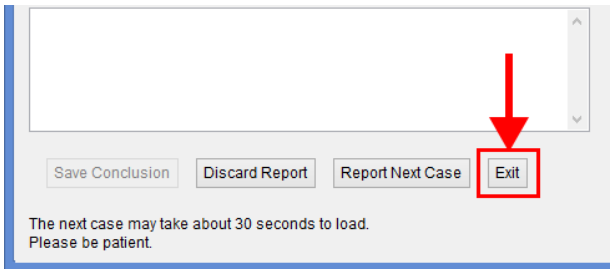

Fig. 36

5. For cases involving two knowns, click on the **Save Conclusion** (Fig. 37) button to save your conclusion for the crime scene impression versus the *first known footwear*.

Fig. 37

6. Click on the **Examine Second Known in this Case** button (Fig. 38) to report your findings regarding the *second known footwear* in the current case (Fig. 39). You must complete reporting for both knowns in a given case in order to exit or move on to another case.

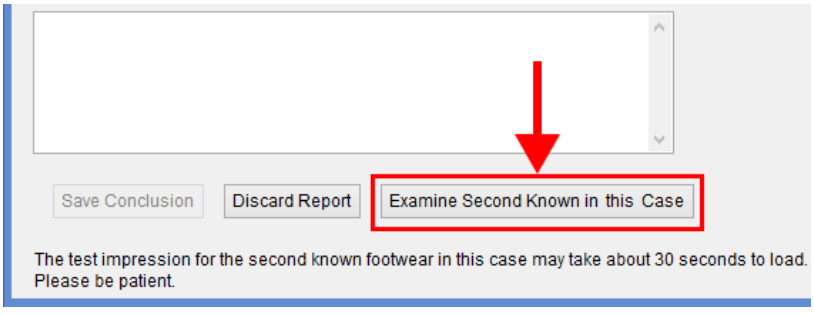

Fig. 38

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

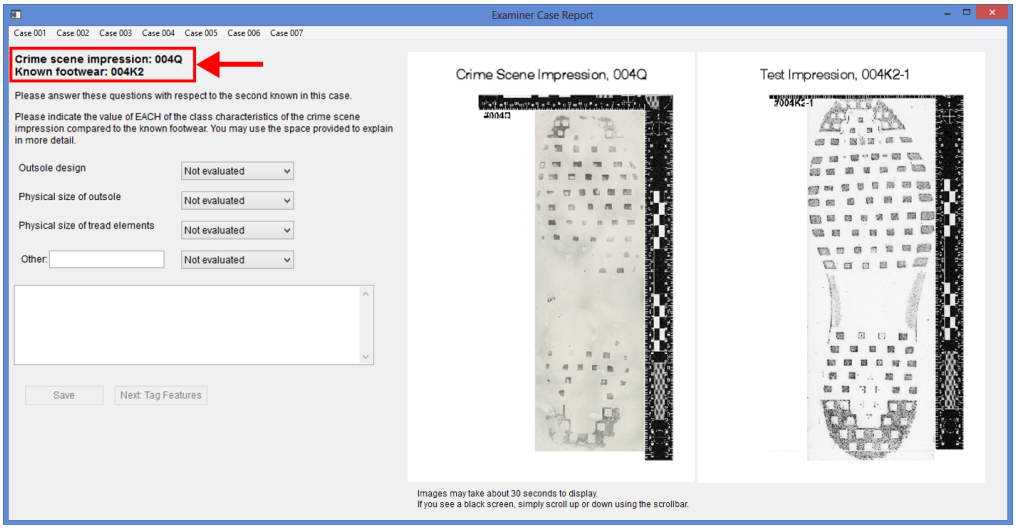

Fig. 39

7. Repeat the reporting process for the crime scene impression versus the *second known footwear* according to Subsections 6.2 to 6.4 step (3.).

   (a) Please indicate the difficulty of the case (inclusive of both first and second knowns) to facilitate future research studies.

   (b) When finished, click **Save Conclusion** (Fig. 40).

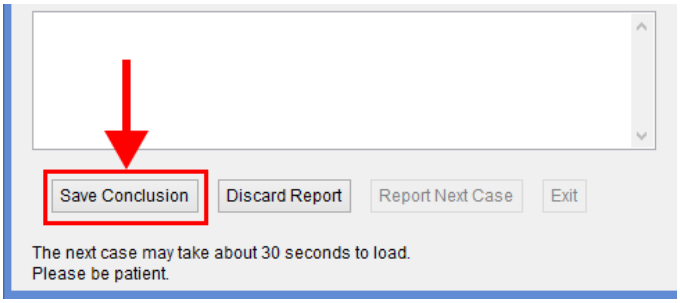

Fig. 40

   (c) If you would like to continue reporting for a different case, click **Report Next Case** (Fig. 41). Refreshing the application may take up to 30 seconds; please be patient. *Note:* The next case does not load automatically. You must select the case according to Subsection 6.1, page 11, step (2.), Fig. 19.

   (d) If you would like stop for now, click **Exit** to close the application (Fig. 42); ***please avoid closing the application using the X button in the top right corner (Windows) or top left corner (Mac).*** *If you do so accidentally, a dialog box may pop up to notify you of an error. Click* ***OK****.*

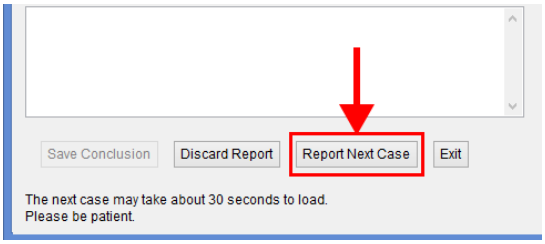

Fig. 41

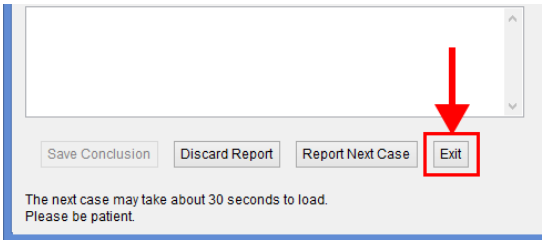

Fig. 42

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

6.5 Discarding a report

1. If you would like to discard the current report, you may only do so at the **Report Conclusions** page by clicking on **Discard Report** (Fig. 43).

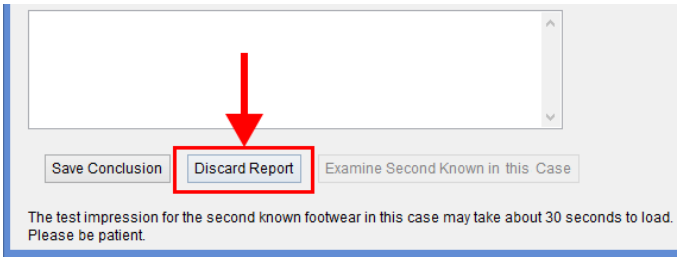

Fig. 43

2. A dialog box will appear to confirm your action (Fig. 44). Click **Yes** to confirm or **No** to return to the **Report Conclusions** page. ***Attention:*** Selecting **Yes** will delete files related to the current case, which may include images and text responses. This means that if you opt to discard after completing the report for the *second known footwear*, it will also discard your report for the *first known footwear* and you will have to restart your report for the *entire case* at a later time.

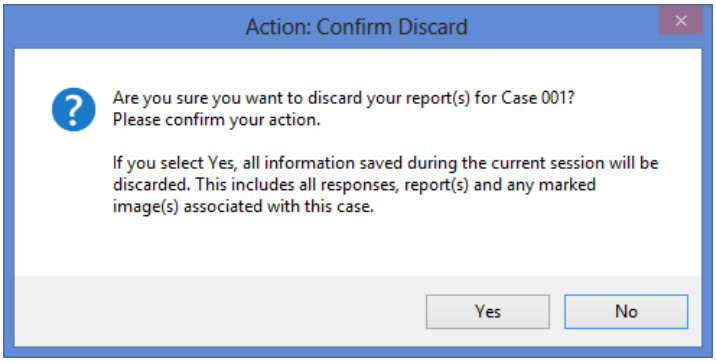

Fig. 44

3. If you choose to discard, a confirmation dialog box will appear (Fig. 45), asking if you would like to refresh the application.

   (a) Click **Yes** to continue using the application and report a new case.

   (b) Click **No** to quit the application and resume your reporting at a later date.

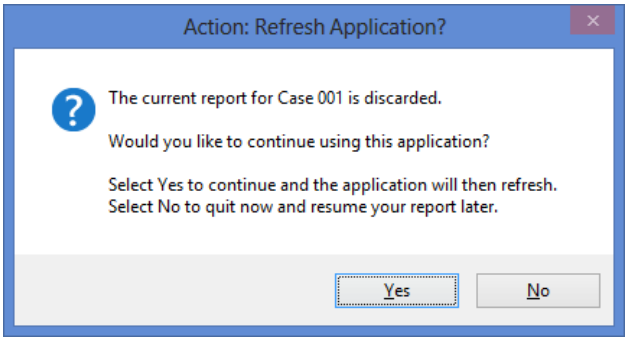

Fig. 45

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

4. If you would like to re-report on a case, but after the discard option has passed, simply report on the case a second time. In other words, if you reported a case twice (or more), your most recent responses will be used by the research group.

6.6 Help with application

1. If the images seem to have disappeared, simply adjust your scroll position using the scrollbars (Fig. 46), and the images will refresh.

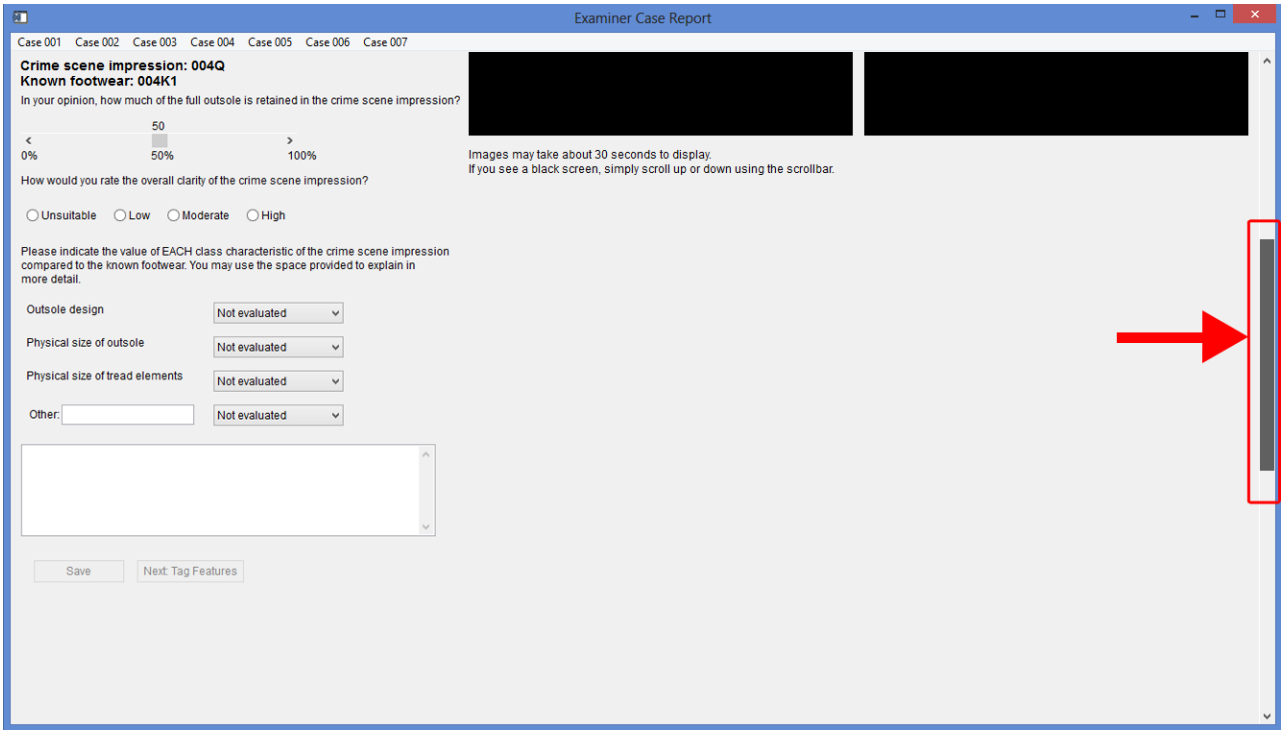

Fig. 46

2. If you neglect to fill in the username field, a warning dialog box will pop up (Fig. 47). Click **OK** to close the box. Enter your username as in Subsection 6.1, page 11, step (1.), Fig. 18 and proceed.

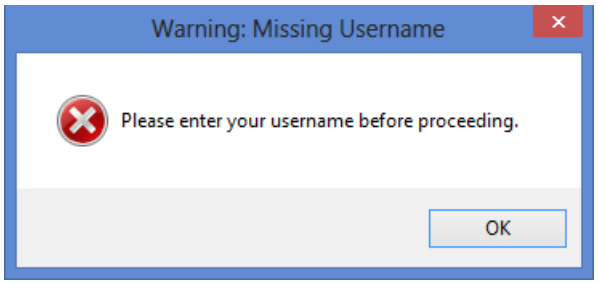

Fig. 47

3. If you receive a warning that some image files may be corrupted, you must replace your resources subfolder. Navigate to your *"ExaminerReport"* folder and locate the *"resources"* subfolder. Right-click on the subfolder and select **Delete** to discard. A new copy of the *"resources"* subfolder is available for download from the following link: `tr.im/Resources_Download`. Download the subfolder and place into the *"ExaminerReport"* folder. You should be able to proceed as normal using the reporting interface.

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

7 Submitting data online

1. After you have completed reporting for *all cases*, please submit your conclusions.
2. Locate the desktop folder "*ExaminerReport*" and open it.
3. Inside this folder, select the subfolder labeled "*output*."
   (a) For Windows users, right click on the folder. Select **Send to** and **Compressed (zipped) folder** (Fig. 48). This may take approximately 2-5 minutes.

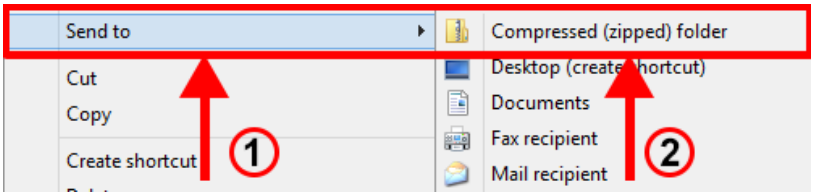

Fig. 48

   (b) For Mac users, right click on the folder. Select **Compress "output"** (Fig. 49). This may take approximately 2-5 minutes.

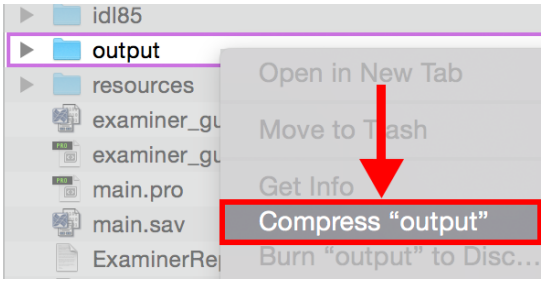

Fig. 49

   (c) If you receive a message indicating that you are denied permission to compress the folder, copy the "*output*" folder to a different location on your desktop or laptop and attempt to zip again.
4. Rename the archive (.zip) file as your user ID (*e.g., afab466.zip*).
5. Using your preferred web browser, go to: `https://tr.im/ex_main`. Please note that the link is case-sensitive.
6. The link will take you to a page where you can upload the compressed file anonymously (Fig. 50). In the top right corner, click on the button **Choose Files**, and locate the .zip file. You may also drag the .zip file to the highlighted box.

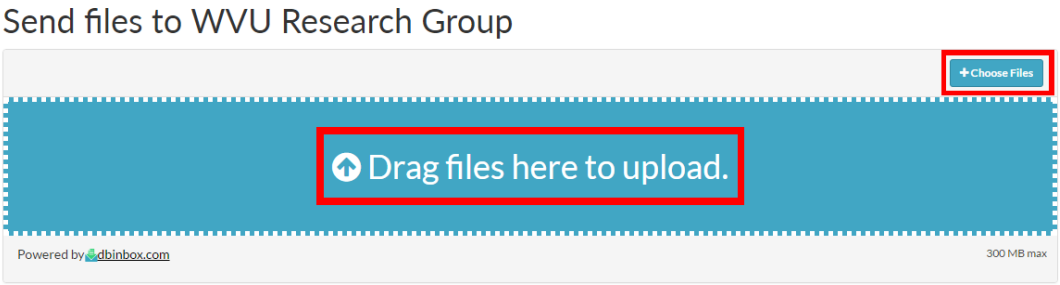

Fig. 50

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

7. Once successfully uploaded, the row indicated with your filename will be highlighted in green (Fig. 51). *Please do not close the window until after the application has completed transferring your files.*

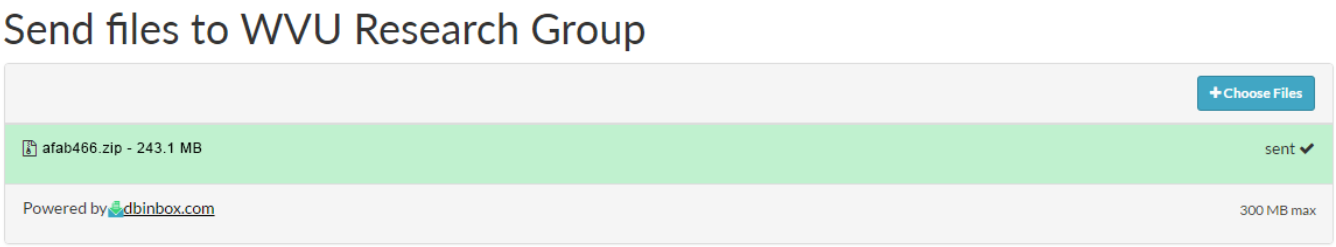

Fig. 51

8. Finally, please do not delete the "*ExaminerReport*" folder until after you have received an email confirming that your .zip file has been received by the WVU research group. Please allow 24-48 hours for this notification. Once you have been informed of the receipt of your data, please feel free to delete all electronic files related to this study.

Note: If you are using a government-owned computer to run the reporting application, and your agency's security restrictions limit (a) access to file sharing websites (such as the results submission portal at `https://tr.im/ex_main`), and (b) the use of USB flash drives to transfer files from your agency's computer to a personal computer or device, please contact the Speir Research Group regarding other submission options.

A.3 DRSA Validation

Although the theoretical underpinnings of DRSA exist within the literature, its coded implementation invariably includes liberties. Initially, it was hoped that the footwear dataset could be analyzed using existing and open-source code. Unfortunately, the complexity of this dataset (as well as the combination of variables) could not be analyzed in this manner. Thus, an `R-Shiny` version was contracted. The resulting implementation is the intellectual property of Dr. Endre Palatinus. However, validation efforts were conducted by the WVU research group. Each data validation is discussed below, including a general summary.

1. Pawlak Z. Rough sets and intelligent data analysis. Information Sciences 2002; 147:1-12.
    - DRSA Primer example
    - Rules slightly vary from those suggested in the paper, but they are all correct
    - Contains only dominance variables
    - Paper does not mention unions, GUI does (therefore coverage calculation differs)
2. Greco S, Matarazzo B, Slowinski R, Stefanowski J. An Algorithm for Induction of Decision Rules Consistent with the Dominance Principle. 2000 Oct; 304-313.
    - Example on pg. 310
    - GUI generated an extra decision rule that is not mentioned in the paper, but the rule is correct
    - Contains only dominance variables
    - Paper does not provide support, certainty, coverage, and strength metrics for direct comparison
3. Slowinski R, Greco S, Matarazzo B. Rough Sets in Decision Making. In: Meyers R, editors. Encyclopedia of Complexity and Systems Science. New York, NY: Springer, 2009; 7753-7786.
    - Student example on pg. 7763
    - Almost all rules differ from those suggested in the paper, but they are all correct
    - Paper does not mention unions, GUI does
    - Contains only dominance variables
    - Paper does not provide support, certainty, coverage, and strength metrics for direct comparison
4. Greco S, Matarazzo B, Slowinski R. Rough sets theory for multicriteria decision analysis. European Journal of Operational Research. 2001; 129:1-47.
    - Student example on pg. 23

- Results comparable
- Contains only dominance variables
- Paper does not provide support, certainty, coverage, and strength metrics for direct comparison

5. Greco S, Matarazzo B, Slowinski R. The use of rough sets and fuzzy sets in MCDM. In: Gal T et al., editors. Multicriteria Decision Making. New York, NY: Springer Science + Business Media, 1999: 14-3 – 14-52.

   - Warehouse example on pg. 14-12
   - Decision rules from GUI do not mention QUAL variable (A2), paper does
   - Rules from GUI are correct
   - Contains only dominance variables
   - Paper does not provide support, certainty, coverage, and strength metrics for direct comparison

6. Greco S, Matarazzo B, Slowinski R. A New Rough Set Approach to Multicriteria and Multiattribute Classification. In: Polkowski L, Skowron A, editors. Rough Sets and Current Trends in Computing. Heidelberg, Germany: Springer-Verlag, 1998: 60-67.

   - Warehouse example on pg. 64
   - Results are comparable
   - Contains dominance and indiscernible variables
   - Paper does not provide support, certainty, coverage, and strength metrics for direct comparison

7. Pawlak Z. In Pursuit of Patterns in Data Reasoning from Data- The Rough Set Way. In: Alpigini JJ et al., editors. Rough Sets and Current Trends in Computing. Heidelberg, Germany: Springer-Verlag, 2002: 1-9.

   - Weather example on pg. 3
   - Paper does not state decision rules, just uses the initial facts to make the decision table
   - GUI rules are correct
   - Contains only dominance variables
   - Paper provides support, certainty, coverage, and strength metrics. Match with GUI for Rule 1, slightly differ for Rule 2 because it is a combination of rules from the paper

8. Slowinski R. Dominance-based Rough Set Approach to Multiple Criteria Decision Aiding. Poznan University of Technology, 2012.

   - Student example on pg. 20

- Decision rules from GUI do not mention MATH variable, paper does
- GUI rules that differ from paper are still correct
- Contains only dominance variables
- Paper does not provide support, certainty, coverage, and strength metrics for direct comparison

9. Slowinski R, Vanderpooten D. Similarity relations as a basis for rough approximations. ICS Research Report 53/95, Warsaw Univ. Technology, 1995.

    - Example on pg. 15
    - Paper gives more rules than the GUI
    - One of the GUI rules has two similarity ranges for the same variable on the LHS (left hand side)
    - All GUI rules are correct
    - Uses a combination of similarity (uses α and β values) and indiscernible variables
    - Paper does not provide support, certainty, coverage, and strength metrics for direct comparison

Summary

- The `R-Shiny` GUI works well for the examples discussed above. There were several instances where the rules generated from the GUI differed from those listed in the paper, but the GUI rules were still correct. Papers often did not use unions, but the GUI does and they are correct.
- Most papers do not compute support, certainty, coverage, and strength metrics for comparison, but those that did were consistent with the GUI when the rules were written the same way.
- A variety of variable types were tested, but most data sets used exclusively dominance variables. One included a combination of dominance and indiscernible variables, which was successful. One included a combination of similarity and indiscernible variables, which was also successful but differed from the paper. Unfortunately, an existing dataset with all types of variable (indiscernible, similarity and dominance) was not available for comparison.
- Datasets of different sizes were also examined, ranging from 6 to 980 objects; this variation did not seem to impact the results, and all examples ran reasonably quickly within the GUI.