Document Title: A Black Box Study of the Accuracy and Reproducibility of Tire Evidence Examiners' Conclusions

Author(s): Nicole Richetelli, Jan LeMay, Kensley M. Dunagan, William J. Chapman

Document Number: 310751

Date Received: July 2025

Award Number: 2020-DQ-BX-0026

A Black Box Study of the Accuracy and Reproducibility of Tire Evidence Examiners' Conclusions

Final Technical Research Report
Submitted on: 21 December 2023
Funded by: US Department of Justice, Office of Justice Programs, National Institute of Justice
Award Number: 2020-DQ-BX-0026

Authors:

Nicole Richetelli (Noblis)
Jan LeMay (Denver Police Department Crime Laboratory)
Kensley M. Dunagan (Noblis)
William J. Chapman (Noblis)

Principal Investigator:

William Chapman
Project Manager, Noblis
2002 Edmund Halley Drive
Reston, VA 20191
william.chapman@noblis.org
(703) 610-2983

Administrative Contact:

Kristi Lovelace
Contracts Manager, Noblis
kristi.lovelace@noblis.org
(703) 610-1562

Award Recipient:

Noblis, Inc.
2002 Edmund Halley Drive
Reston, VA 20191

Project Period:

January 1, 2020 to December 31, 2022; extended to December 31, 2023 (via NCE)

Award Amount:

$720,610

Abstract

This report provides an overview of the first formal black box study of tire impression evidence decisions, which was funded by the NIJ (2020-DQ-BX-0026). Herein, we provide summaries of the study goals, approach, test yield, key results and findings, limitations, and implications. We also describe the artifacts resulting from this project. Additional details, including a thorough description of the study methods and additional results, are available in the forthcoming manuscript [1].

This project was supported by Award No. 2020-DQ-BX-0026, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication/program/exhibition are those of the author(s) and do not necessarily reflect those of the Department of Justice.

1 Introduction

Tire impression evidence can be a valuable tool during a crime scene investigation—it can link vehicles to scenes or secondary locations and reveal information about the series of events surrounding a crime. However, forensic pattern interpretation disciplines, which include tire impression evidence, have faced scrutiny over the last several decades due to the subjective nature of the analysis process. In these disciplines, evidence interpretations are conducted by human examiners and largely rely on their accumulated training, knowledge, and expertise. These interpretations generally result in source attribution decisions; for tire impression evidence, an examiner provides their opinion regarding whether a known tire could be the source of a recovered tire tread impression. In order to understand the scientific reliability of these disciplines, it is necessary to characterize the accuracy and reproducibility of the decisions reported by examiners, which has been underscored by two well-known reports: the 2009 National Academy of Sciences (NAS) report [2] and the 2016 President's Council of Advisors on Science and Technology (PCAST) report [3,4]. These reports call for targeted empirical research efforts to help strengthen the scientific basis of pattern disciplines and characterize the reliability of these methods and of examiner decisions. The PCAST report specifically states that "the only way to establish the scientific validity and degree of reliability of a subjective forensic feature-comparison method—that is, one involving significant human judgement—is to test it empirically by seeing how often examiners actually get the right answer. Such an empirical test of a subjective forensic feature-comparison method is referred to as a 'black-box test.'"[3]. The National Commission on Forensic Science reinforced these recommendations and reiterated the calls for such testing [5]. Despite these calls for inquiry, there had never been a formal, rigorous empirical evaluation of the reliability of tire impression evidence decisions until this study was conducted. The results of this study will provide key information about the practice of tire impression examination to laboratory managers, practitioners, and the legal system that can help to facilitate improvements in standardization, training, and practice.

2 Project Summary and Overview of Study Results

2.1 Goals, Objectives, and Research Questions

This study was conducted to systematically and empirically evaluate the accuracy and reproducibility of decisions reported by practicing tire impression examiners on samples selected to be broadly representative of casework. In addition, this study was designed to produce a substantial dataset of ground-truth attributed tire impression imagery; this dataset will be publicly released in an effort to facilitate future research and support improvements to training and standardization for the discipline.

The primary research questions were as follows:

- How accurate are decisions reported by forensic tire impression examiners?
- How reproducible are these decisions among examiners?

- What factors are associated with the performance or reporting tendencies of tire impression examiners?

2.2 Study Design Overview

This research was modeled after previous black box studies that examined the accuracy and reliability of forensic examiners in the latent print, bloodstain, footwear, and handwriting disciplines [6–9]. The full study protocol was approved by the Advarra Institutional Review Board (Pro00051041), following NIJ requirements for human subjects research.

2.2.1 Study Period

Study registration opened in August 2022 and remained available throughout the entire course of the study; the full study was accessible for a period of 30 weeks (September 2022-April 2023). Participants received regular study-relevant notifications and reminders (via email) throughout the course of the study.

Prior to commencement of the study, two pilot tests were conducted to assess software functionality, examiner experience, and assignment logistics. The pilot test materials were not reused in the formal study, nor were the results included in analysis.

2.2.2 Study Eligibility and Participation Requirements

Participation in this study was open to U.S. and Canadian tire impression examiners who (1) had conducted operational casework within the past five years, (2) use (or have used) a categorical conclusion scale (e.g., SWGTREAD 2013), and (3) are proficient in English. In addition, U.S. and Canadian tire impression examiner trainees were also welcome to participate. Participants were solicited via postings and announcements through relevant scientific working groups and professional organizations, including the International Association for Identification (IAI), the IAI Footwear Certification Board, the Organization of Scientific Area Committees for Forensic Science (OSAC), and the International Association of Chiefs of Police (IACP). All interested parties from the footwear black box study [8] were also contacted about the TireBB study, given the high degree of overlap and cross-training between the two disciplines.

Interested tire impression examiners and trainees (hereafter "FTETs") were required to register for the study on the study website, which required completion of an eligibility check, a data use agreement, and an online consent form. After completing the registration process, registrants were required to complete a background questionnaire regarding their tire-impression related training and experience. Each participant was then asked to complete 40 digital tire impression comparisons, although partial participation[*] was accepted in this study in an effort to increase the number of responses received. After completing their

[*] *We accepted all comparisons submitted in this study and did not require that an FTET complete a minimum number of comparisons to be included in analysis. In other words, to be considered a participant in this study, an FTET was required to complete at least one assigned tire impression comparison.*

assigned comparisons, participants were asked to complete a post-test questionnaire regarding their testing experience and the representativeness of the study samples.

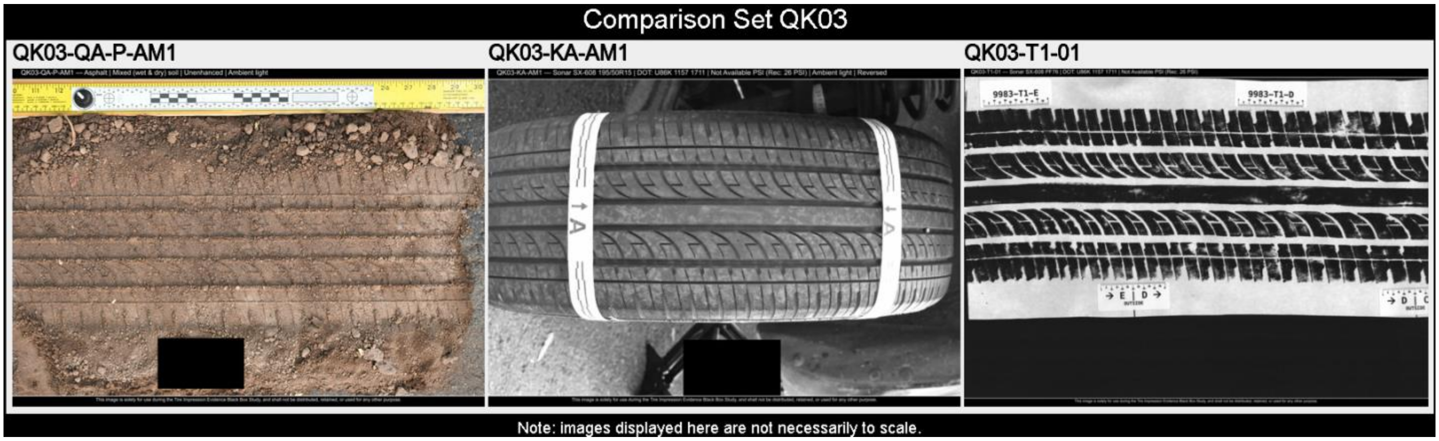
Participants were assured that all results would be anonymous, in accordance with human subjects research requirements, which was achieved using a blind coding system. Participants' study results were saved using random ParticipantIDs and were not linked to any personally identifiable information (PII). ParticipantIDs were replaced by Anonymous IDs (AnonIDs) prior to data analysis, precluding the analysis team's ability to cross associate participants, personally identifying information, questionnaire responses, or test results. Destruction of existing cross-references occurred prior to public presentation of results (e.g., indices associating ParticipantIDs with email addresses, or associating ParticipantIDs with AnonIDs). Therefore, participant identities could not be associated with the results at any point during analysis, or subsequently, such as for discovery.

2.2.3 Collection of Tire Examination Responses from Participants

Each participant was assigned 40 digital tire impression comparison sets (hereafter QKsets), of which 17 were *mated* (in which the K is the source of the Q), and 23 were *nonmated* (in which a tire other than the K is the source of the Q). Each assigned QKset included digital images of 1 questioned impression (Q) and 1 known tire (K) for comparison. Participants were not informed of the mated to nonmated QKset proportions, nor of the mating ground truth of each QKset; the order of packets as well as the order of mating was randomized between participants. Participants were instructed to conduct their examinations independently, using the same diligence employed in their operational casework. No time limitation was imposed for the completion of each assigned QKset comparison. As a quality assurance measure, participants were only permitted to access one QKset at a time to avoid the possibility of administrative errors or misunderstandings. Participants were required to complete the QKsets in the assigned order, and they could not view the images for any subsequent QKset prior to submitting their responses for the current QKset.

This study was entirely digital — all study materials were provided to participants as digital images only (no physical materials or tire items were distributed) — and was conducted using a custom web browser-based software interface. The study website was accessible using an ordinary web browser and did not require the download or installation of any additional software or plugins. The website facilitated the entire study — it presented questions, displayed proofsheet images, permitted download of full-resolution images, recorded participant responses, and transmitted responses back to study administrators.

QKsets were provided to participants as ZIP file downloads from the study website. The QKset file contents included a series of high-quality images[*] of questioned tire impressions (including original impressions as well as lifts and dental stone casts), known tire tread segment photographs, and replicate test impression scans. Images included rulers for calibration by participants. Figure 1 shows a proofsheet for QK03, which displays a representative questioned impression (3-dimensional, mixed wet and dry soil on asphalt), known tire tread section, and known tire test impression section image; proofsheets were not intended for comparison but were provided to participants on the study website for quality assurance purposes. K images (tread and test impression images) represented the full circumference of the tire. All images depicting tire tread sections, casts, and lifts were reversed (i.e., the images were flipped horizontally) so that they oriented with the impression on the ground; such images were labelled "Reversed". Participants were also provided with the following metadata for each QKset: questioned impression substrate/matrix, questioned impression processing/collection method, known tire make/model/size, known tire DOT number, and known tire PSI at time of collection (as well as the recommended PSI for the vehicle).


Comparison Set QK03
QK03-QA-P-AM1
QK03-QA-P-AM1 — Asphalt | Mixed (wet & dry) soil | Unenhanced | Ambient light
QK03-KA-AM1
QK03-KA-AM1 — Sonar SX-608 195/50R15 | DOT: U86K 1157 1711 | Not Available PSI (Rec: 26 PSI) | Ambient light | Reversed
QK03-T1-01
9983-T1-E
9983-T1-D
→ E | D →
→ D | C
Note: images displayed here are not necessarily to scale.

[redacted]

[redacted]

Figure 1. Proofsheet displayed on the study webpage for QK03, showing the questioned impression (left; photographed with ambient lighting), a single tread section (middle; photographed with ambient lighting), and a section of one of the test impressions (right; scanned with flatbed scanner).

For each of the assigned QKsets, the participants assessed impression quality; evaluated suitability; assessed the correspondence of tread design, size, mold variation, and wear; assessed the presence and correspondence of RACs; selected a decision category; rated difficulty and typicality; and detailed any limitations encountered.

[*] *All original images collected for this study (photographs and scanned) were converted from lossless TIFF to highest quality JPG using ImageMagick to reduce file sizes for both storage (on the web site) and download (by participants), while maintaining image quality. The decision to provide highest quality JPG images was based in part on lessons learned from the FootBB study* [8]*, which provided both lossless TIFF and high-quality JPG images; in practice, it was extremely difficult to detect any difference in the TIFF vs JPG images.*

The study required participants to report their source opinions using the following scale, which is a modified version of the SWGTREAD (2013) conclusion scale [10] that was also used in the FBI-Noblis footwear black box study [8]:

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

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

2.2.4 Tire Impression Sample Collection and Preparation

A total of 70 distinct tires were used to create the 80 QKsets (40 mated, 40 nonmated) used in this study, comprising 24 different tire makes/models/sizes, originating from 21 different vehicles. The amount of mileage for each tire was unavailable; the number of miles driven between the deposition of Q and the collection of K was not always available (however, this information would not generally be available during operational casework). All tires used in this study came from vehicles readily available to the study team.

All tire impression samples included in this study were collected under controlled conditions with quality assurance measures designed to ensure the ground-truth (GT) source attribution (mating) of the QKsets. The study team sought to create test samples that were representative of those encountered in casework, spanned the spectrum of quality, incorporated a variety of tire designs, and provided the opportunity for participants to utilize the full range of conclusions.

Sample collection for this study was conducted from June through November of 2021 under controlled conditions, which involved a methodical process designed to maintain sample quality and ground truth source attribution. The sample collection process entailed the following steps: preparation, collection, and imaging of replicate questioned impressions from each tire; preparation and collection of a set of known tread images from each tire being used as a known in QKsets; and preparation, collection, and imaging of replicate test impressions from each tire being used as a known in QKsets.

In general, the preparation and collection of questioned and known impressions followed the best practice recommendations outlined in [11–13] and using the guidelines detailed in [14]. All images included in this study were captured either photographically or digitally scanned, at an image resolution of 300 ppi or greater. The photographic images were captured following the ASB best practices for photographic documentation of tire evidence [15]. Additional details regarding sample collection for this study are provided in the forthcoming manuscript [1].

Questioned Impressions

All questioned impressions (Qs) were produced under controlled conditions[*], by team members trained by subject matter experts; Qs included in the study were limited to those collected at low speeds to ensure the safety of the study team. A total of 130 questioned impressions were created, from which 80 were selected for inclusion in QKsets.

Questioned impressions were produced using various combinations of:

- Substrates: asphalt, cardboard, and wood
- Matrices: soil, sand, and residue (oil/grease)
- Conditions: wet/mixed/dry for soil and sand; unfinished/tempered/painted for wood
- Processing methods: none, black powder enhancement
- Collection methods: photography, gelatin lifters, dental stone casting

Both two-dimensional (2D) and three-dimensional (3D) impressions were included in this study and these Qs spanned a range of sizes (ranging from approximately 12-30 inches in length). Impressions with varying degrees of distortion, obfuscation, and

[*] *All collection was done in an outdoor parking lot equipped with a large tent for processing and temporary storage as well as an onsite, indoor laboratory for long-term storage and any additional required processing or photography. All collections occurred on dry days (no precipitation) during daylight hours to reduce the chances of interference or issues with the collection process.*

background interference were also generated to vary the quality of the resulting impressions. For examples of questioned impressions of varying quality, please see the QKsets provided in the public dataset [16].

Known Tire Tread Images and Test Impressions

For each known tire, a set of tire tread images and replicate test impressions were collected, using the best practices outlined in [12,15]. Knowns (Ks) from a given tire were collected from approximately two hours to 138 days (median 10 days) after the initial deposition of the Qs. All known tires included in the study were front tires in order to prevent the need to remove the tire from the vehicle for collection of tread images.

Prior to collecting test impressions, a set of known tire tread photographs was collected. To do so, each known tire was broken into segments by marking the location of the tread wear indicators on the sidewall of the tire. Tire chocks were placed behind the rear tires, and the vehicle was shifted into Neutral; the front of the vehicle was then lifted with a jack to raise the tire of interest off the ground and permit free rotation. Each tread segment was then photographed, using both ambient and oblique illumination, to capture the full circumference of the tire tread.

To ensure that the condition of the known tires was exactly the same in the known tread images and the test impressions, a set of two replicate test impressions was collected immediately after capturing the tread photographs. The test impressions were created using printer’s ink, following the best practices outlined by ASB [12]. All test impressions were collected with the tires mounted to the original vehicle. After drying for a minimum of 24 hours, test impressions were scanned in overlapping sections using a flatbed scanner (captured in grayscale as 600 ppi TIFF images).

2.3 Test Yield

2.3.1 Participants

A total of 39 FTETs registered for the study and completed the background questionnaire. Of those, 17 completed at least one QKset comparison and were included in analyses; 22 examiners did not complete any comparisons. We refer to the 17 FTETs who completed at least one tire comparison as “participants,” and we refer to the 22 FTETs who did not complete any comparisons as “registrants” for clarity. For brevity, we only discuss the participants in this report; for additional details about the registrants, see [1]. Of the 17 FTETs who participated in this study, three completed all 40 assigned comparisons, with each participant submitting responses to 14 assignments on average, ranging from 1-40 (median = 9).

Participants were required to complete a background questionnaire prior to starting their assigned comparisons. At the time of registration, 15 participants were qualified practicing tire examiners, 1 was conducting supervised casework, and 1 was in training for tire impression evidence. Twelve participants currently conduct tire examination casework at a US agency (4 local, 7 state, 1 federal) and five work at a Canadian agency (3 local, 2 provincial). The participants’ tire-related experience ranged from 1-15 years: seven had 1-5 years, six had 6-10 years, and three had 11-15 years. Ten participants received formal tire impression examination training (four had 12+ months training, four had 6-12 months training, six had less than 6 months training), one received informal, on-the-job training, and four only completed courses and/or workshops. The majority of participants (15/17) use the SWGTREAD 2013 conclusion scale in their operational casework, one uses the 2020 DOJ ULTR, and one uses another categorical conclusion scale. Tire impression evidence comparisons were not a primary responsibility of the 17 participants—all participants were also qualified in footwear and at least one additional discipline, most often fingerprints and/or crime scene investigation. In addition, all the participants only conduct tire examinations a few times a year, which suggests that tire evidence may not be commonly submitted to the laboratory for examination.

Given that the tire discipline is quite small and tire impression evidence examinations are not a common laboratory request, we expected that the number of participants for this study would be smaller than similar such studies in other disciplines (e.g., latent fingerprints, footwear); however, participation was lower than initially anticipated. Although there is no means of knowing the number of practicing tire examiners in the US and Canada, we can use publicly available proficiency test data to develop a rough estimate. From 2020 (when TireBB was funded) through 2022 (when TireBB registration opened), 57-78 tire examiners completed the Collaborative Testing Services (CTS) Tire Track Imprint proficiency test each year [17–19] (note that the CTS test is available to examiners from a variety of countries, but only US and Canadian examiners were eligible for our study). Of the 15 practicing tire examiners who participated in this study, 9 had completed a proficiency test in the last year; from this we could make a rough estimate that the study participants represent 12-16% of the population of FTEs who take CTS proficiency tests, which we expect is an underestimate given that international participation is accepted for the CTS test. Note that we have no way of estimating the population of FTEs who do not take proficiency tests: the roughly 1/3 of participants who had never taken a proficiency test may or may not reflect the overall population. Despite the small number of participants in this study, we do believe that we captured a non-trivial proportion of the practicing tire examiners in the US and Canada.

Appendix A1.1 Responses

A total of 80 distinct QKsets were created for this study: 40 mated and 40 nonmated. Of these, 77 received at least one response in the study (37 mated and 40 nonmated); each QKset received 3.1 responses on average (3.0 median), with the total number of responses per set ranging from 1-6. In total, we collected responses from 17 participants on 238 trials, across 77 distinct QKsets: 92 responses to mated QKsets and 146 responses to nonmated QKsets.

2.4 Summary of Study Results

Analyses reported in this section were based on a total of 238 responses from 17 participants on 77 distinct QKsets.

2.4.1 Decision Rates, Accuracy, and Errors

In general, accuracy can be described as the extent to which decisions are consistent with ground truth, or the avoidance of errors and incorrect responses. When discussing the accuracy of decisions in this study, we only use the term "error" to refer to definitive decisions that contradict ground truth (i.e., reporting *ID* for a nonmated QKset or reporting *Excl* for a mated QKset). We explicitly differentiate errors from probable decisions that are inconsistent with ground truth (e.g., reporting *HighAssn* for a nonmated QKset), which we describe as "incorrect." We make this distinction between errors and incorrect decisions because FTETs explicitly differentiate between definitive and probable decisions in reporting, based upon the amount of information available in the comparison and their interpretations of their observations; it would therefore be inappropriate to consolidate them together into a single category of error, which effectively removes the important and intentional nuance required to appropriately convey the intended weight of the FTET's decision.

Furthermore, the use of an eight-level scale requires that traditional methods of measuring accuracy and errors be adapted, particularly to consider probable decisions and class associations. Various approaches can be used to calculate decision rates for multi-level conclusion scales. As recommended by OSAC [20] and similar to previous studies [6,8,9], decision rates can be reported in four ways: 1) PRES, which includes all presentations; 2) COMP, which includes all comparisons, excluding *NotSuitable* trials, 3) CALLS, which includes only comparisons resulting in non-indeterminate responses, excluding *NotSuitable* and *Inc* trials and 4) DEF, which includes only comparisons resulting in definitive decisions (*ID* and *Excl*), excluding trials with indeterminate (*NotSuitable* or *Inc*), class association (*LimitedAssn* or *Assn*), or probable decisions (*NonAssn* or *HighAssn*).

Table 1 reports decision rates computed based upon PRES, COMP, CALLS, and DEF. In addition to the measured rates, we also provide 95% Clopper Pearson[*] confidence intervals (CIs) [21] for each metric. We must acknowledge that the limited sample size in this study (both overall study size and by participants/QKsets) does limit the precision of the rate estimates and underscores the importance for considering the confidence intervals when assessing performance as measured in this study. Note that the method used to compute decision rates can have a notable effect on the resulting rates, particularly when only considering trails resulting in definitive trials—for example, the proportion of nonmated QKsets resulting in *Excls* (true exclusion rate) varies from TER_{PRES}=15.8%, TER_{COMP}=16.1%, $\text{TER}_{\text{CALLS}}$=16.7%, TER_{DEF}=95.8%. Although we provide the decision rates computed in all four manners for completeness, in the remainder of this report we specifically summarize the results based upon all presentations (PRES). Because non-definitive determinations comprise such a large proportion of the decisions

[*] *The Clopper-Pearson CI is a commonly utilized binomial CI that generally produces conservative estimates of the interval* [21,26]*. Note that since rates are not evenly distributed by QKset or by participant, any approach for measuring confidence intervals is necessarily imperfect. Furthermore, the Clopper-Pearson estimate assumes independence among decisions; because our data includes commonalities of examiners and image pairs, the CIs presented here may be narrower than appropriate for the data.*

reported by FTETs, omitting these determinations from the computation of decision rates would not be a good representation of the decision space. In addition, the results based upon all presentations allow for direct comparison with the comparable rates from the footwear black box study [8], which may be of interest because the two disciplines are related and all participants in this study were cross-trained in both footwear and tire examination.

	Mated					*Nonmated*				
	#	*%PRES*	*%COMP*	*%CALLS*	*%DEF*	#	*%PRES*	*%COMP*	*%CALLS*	*%DEF*
Excl	3	3.3% [0.7%-9.2%]	3.3% [0.7%-9.2%]	3.3% [0.7%-9.2%]	12.0% [2.5%-31.2%]	23	15.8% [10.3%-22.7%]	16.1% [10.5%-23.1%]	16.7% [10.9%-24.0%]	95.8% [78.9%-99.9%]
NonAssn	1	1.1% [0.0%-5.9%]	1.1% [0.0%-5.9%]	1.1% [0.0%-5.9%]	---	14	9.6% [5.3%-15.6%]	9.8% [5.5%-15.9%]	10.1% [5.7%-16.4%]	---
NotSuitable	0	0.0% [0.0%-3.9%]	---	---	---	3	2.1% [0.4%-5.9%]	---	---	---
Inc	0	0.0% [0.0%-3.9%]	0.0% [0.0%-3.9%]	---	---	5	3.4% [1.1%-7.8%]	3.5% [1.1%-8.0%]	---	---
LimitedAssn	22	23.9% [15.6%-33.9%]	23.9% [15.6%-33.9%]	23.9% [15.6%-33.9%]	---	40	27.4% [20.3%-35.4%]	28.0% [20.8%-36.1%]	29.0% [21.6%-37.3%]	---
Assn	39	42.4% [32.1%-53.1%]	42.4% [32.1%-53.1%]	42.4% [32.1%-53.1%]	---	60	41.1% [33.0%-49.5%]	42.0% [33.8%-50.5%]	43.5% [35.1%-52.2%]	---
HighAssn	5	5.4% [1.8%-12.2%]	5.4% [1.8%-12.2%]	5.4% [1.8%-12.2%]	---	0	0.0% [0.0%-2.5%]	0.0% [0.0%-2.5%]	0.0% [0.0%-2.6%]	---
ID	22	23.9% [15.6%-33.9%]	23.9% [15.6%-33.9%]	23.9% [15.6%-33.9%]	88.0% [68.8%-97.5%]	1	0.7% [0.0%-3.8%]	0.7% [0.0%-3.8%]	0.7% [0.0%-4.0%]	4.2% [0.1%-21.1%]
Total PRES	92					146				
Total COMP	92					143				
Total CALLS	92					138				
Total DEF	25					24				

Table 1. Decision rates and 95% Clopper-Pearson confidence intervals, computed based upon PRES, COMP, CALLS, and DEF. Errors are highlighted in yellow and incorrect decisions are highlighted in blue; indeterminate decisions are shown in gray.

Overall, class associations (*Assn* and *LimitedAssn*) were the most commonly reported decision group (68% of all 238 responses), followed by definitive decisions (21% *ID* or *Excl*) and then probable decisions (8% *HighAssn* or *NonAssn*). Just 3% of responses reported neutral determinations, all of which were reported on nonmated QKsets—it is important to consider the limited sample size in this study (both overall and by QKset/participant); therefore, this observation is likely an artifact due to sample size and not necessarily a true effect of mating on the tendency to report neutral determinations.

On mated QKsets, 29.3% of all decisions were consistent with ground truth (23.9% True ID Rate[*] (TIR$_{\text{PRES}}$); 5.4% correct *HighAssn* rate (CAR$_{\text{PRES}}$)) and 66.3% were class associations. Conversely, 3 trials resulted in an erroneous *Excl* (False Exclusion Rate (FER$_{\text{PRES}}$) = 3.3%) and 1 trial resulted in an incorrect *NonAssn* (Incorrect Non-Association Rate (IAR$_{\text{PRES}}$) = 1.1%); thus, 4 out of the 92 responses on mated QKsets contradicted ground truth. For nonmated QKsets, approximately one-quarter of all decisions on nonmated QKsets were consistent with ground truth (15.8% True Exclusion Rate (TER$_{\text{PRES}}$); 9.6% correct *NonAssn*

[*] *We use "True ID rate" (TIR), "True Exclusion rate" (TER), etc. because they are more explicit than "true positive" and "true negative," etc., especially when using a multi-level conclusion scale.*

rate (CNR_{PRES})) and 68.5% were class associations. Just one of the 146 responses on nonmated QKsets contradicted ground truth: one erroneous ID was reported (False ID Rate (FIR_{PRES}) = 0.7%) and no incorrect *HighAssns* were reported.

This study included both examiners (fully qualified to conduct independent casework) and trainees (currently conducting supervised casework or in training) in an effort to increase the number of participants: 15 participants were examiners and 2 were trainees. The distribution of decisions reported by examiners and trainees were generally comparable. Of the five erroneous and incorrect responses, trainees reported one (erroneous *Excl*). Similar to examiners, the majority of decisions reported by trainees were class associations. Given that performance was not notably different between participant type (examiner vs trainee) and decision rates, we pool their responses for all analyses moving forward.

Computation of traditional accuracy and error rates requires *a priori* information about ground truth (i.e., knowledge about whether the known tire did or did not make the impression); this information is rarely, if ever, available in operational casework. An alternative method of describing accuracy is using posterior probabilities, which instead estimate the chance that a reported decision is correct, by assessing the proportions of responses for a given decision category that are consistent with ground truth. Overall, 96% of *IDs* reported in this study were correct (i.e., on mated QKsets; positive predictive value, PPV), as were 89% of *Excls* (i.e., on nonmated QKsets; negative predictive value, NPV). Because these rates are posterior probabilities, they are notably affected by the proportions of mated vs nonmated data in the sample set (i.e., the prior probability of mated vs nonmated QKsets); when computationally rescaled to a 50% mate prevalence, 97% of *IDs* and 83% of *Excls* would be expected to be correct. A 50% mate prevalence is a useful point of comparison because it assumes no informative priors (i.e., there is an equal chance that the comparison is mated as there is that it is nonmated), but this assumption may or may not be appropriate for any particular operational scenario.

2.5 Reproducibility of Decisions

Reproducibility refers to inter-examiner consistency: the extent to which responses from different participants agree when given the same QKset. Therefore, to assess reproducibility a given QKset must have at least two responses from FTETs. In total, 67 of the QKsets in this study were compared by two or more examiners and received 228 individual decisions, thus yielding 642 inter-examiner decision pairings for evaluation. The estimates of reproducibility in this study are limited by the number of responses per QKset: each QKset received just 3.1 responses on average (median 3). The number of responses per QKset in this study is notably lower than our previous studies; for example, each QKset in the 2019-2020 FBI-Noblis footwear black box study [8] received an average of 22.4 responses, each QKset in the 2019-2020 FBI-Noblis handwriting black box study received an average of 36.5 responses [9], and each image pair in the original 2009-2011 FBI-Noblis latent print black box study [6] received an average of 23 responses. For this reason, we do not present analyses of examiner consensus on each sample as we did in previous studies [6–9].

Figure 2 describes reproducibility of decisions as a function of the absolute difference in responses. To report this, we assigned each decision category a numerical score (*Excl*=1, *NonAssn*=2, *NotSuitable*/*Inc*=3, *LimitedAssn*=4, *Assn*=5, *HighAssn*=6, *ID*=7) and computed the absolute value of the difference between the categories reported for each decision pairing. For example, if two participants both reported *ID* on a given QKset, the absolute difference in their decisions would be 0 (i.e., 7-7); if, however, one participant reported *ID* and one reported *Assn*, then the absolute difference would be 2 (i.e., 7-5).

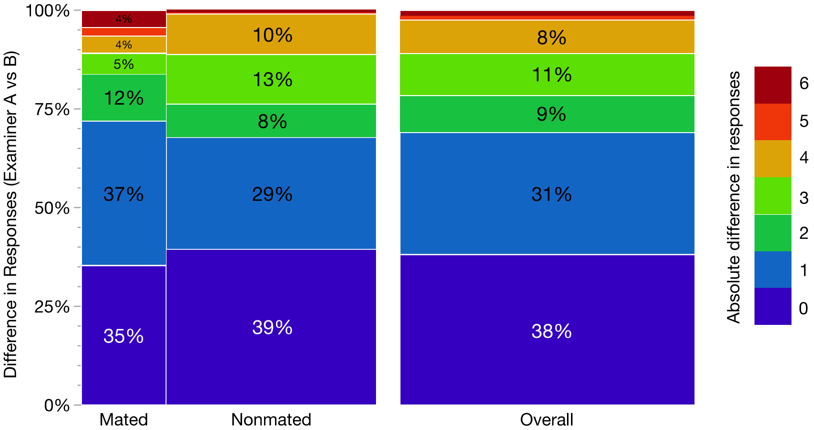


Figure 2. Reproducibility of decisions, summarized by absolute difference in decisions: 0 indicates that the same decision was reported by two participants, and 6 indicates that diametrically opposed decisions (*ID* and *Excl*) were reported by two participants. Note this analysis does not distinguish *NotSuitable* and *Inc*. (n = 642 inter-examiner decision pairs from 228 individual decisions on 67 QKsets with at least two responses)

Overall, 39% of decisions were reproduced exactly (35% mated; 39% nonmated), and 69% (72% mated; 68% nonmated) were reproduced within one decision category. On mated QKsets, *IDs* were the only decisions that were reproduced by a majority of participants (55% of *IDs* were reproduced exactly); on nonmated QKsets, none of the decisions were reproduced by a majority of participants, but *Excls* were reproduced by a plurality (38% of *Excls* were reproduced exactly). When considered as a group, class associations (*Assn* and *LimitedAssn*) were often reproduced — 76% (78% mated; 75% nonmated) of class association responses were reproduced (as either *Assn* or *LimitedAssn*) by a different examiner. Some of the observed variability in decisions, particularly shifts of responses between neighboring categories, can likely be attributed to the size of the conclusion scale used for tire impression examinations, which includes 7 levels. This larger scale allows examiners to use the scale as a continuum, which grants finer granularity in reporting their decisions by accounting for the various types of characteristics that they evaluate and the perceived weight of observed correspondences and/or dissimilarities. However, using such a scale also means that we can expect additional variability due to the number of decision categories available [22–24]; in fact, two examiners may observe the same features and interpret them in a similar manner, but assign slightly different weights to their significance, which could result in the selection of neighboring decision categories despite the fact that they are using the same features to reach this decision.

Just under 2% of decision pairs were diametrically opposed (i.e., *ID* vs. *Excl*; 4% on mated QKsets and 0.4% on nonmated QKsets). These contradictory decisions represent instances in which an error could be flagged during verification, technical review, or independent reanalysis—if both *ID* and *Excl* are reported for a given comparison, one is necessarily incorrect and would be revealed by disagreement during these quality assurance (QA) checks. The result of greater interest is the chance that errors or incorrect decisions are reproduced, and would therefore potentially be missed during QA and possibly even reported out. With respect to mated QKsets, one of the false *Excls* (of three total) was reproduced as an incorrect *NonAssn*—the other 4 decisions on that QKset were all correct, including two *Assns* and two *IDs*. For nonmated QKsets, we only observed one false *ID* and no incorrect *HighAssns*, and thus there were no reproduced errors/incorrect decisions on nonmated QKsets; it is important to note that this result may be due to the small scale of the study and we cannot preclude that the proportion of incorrect *HighAssns* and false *IDs* could increase with additional data, which may also increase the chance of such decisions being reproduced.

2.6 Effects of Questioned Impression Quality

This study was designed to include questioned impressions that spanned a range of quality from bad to excellent. In order to evaluate the effect of quality on decisions, participants answered a series of questions (prior to suitability assessment) pertaining to the quality of the Q for each assigned comparison (see [1] for additional details). These questions were adapted from the questioned impression quality rubric developed in the FBI-Noblis footwear black box study [8,25]. In that rubric, 10 attributes pertaining to quality were evaluated, and each was ranked using an ordinal scale, ranging from 0 (poor) to 2 (good), thus yielding a final questioned impression quality (QQ) score out of a maximum possible score of 20 points for each Q. For this study, we adapted that quality rubric: we included 8 attributes[*] for 2D Qs and 7 attributes for 3D Qs (matrix was not applicable for 3D impressions). Because the maximum QQ score differed for 2D and 3D impressions (16 for 2D and 14 for 3D), we normalized the QQ scores for each impression, thus yielding QQ_{Norm} scores ranging from 0 to 1.

As shown in Figure 3, decision rates varied notably as a function of QQ_{Norm}—as QQ_{Norm} increased, so did the proportion of definitive decisions, whereas the proportion of class associations decreased. Note, however, that the proportion of both true and false definitive decisions increased with quality: all but one of the erroneous and incorrect decisions reported in this study occurred on the highest quality Qs (QQ_{Norm} = 0.8-1.0), and the remaining false *Excl* was reported on a Q in the second highest quality bin (QQ_{Norm} = 0.6-0.8). Similarly, the reproducibility of decisions was related to the questioned impression quality: the

[*] *We asked participants to rate each of the following quality attributes in this study: contrast, quantity, pattern, distortion, overlap, substrate, matrix (2D Qs only), and clarity.*

proportion of decisions that were reproduced exactly or within one decision category increased as the quality of Qs decreased, which is likely due to the increased proportion of class associations for the lower QQ_{Norm} bins.

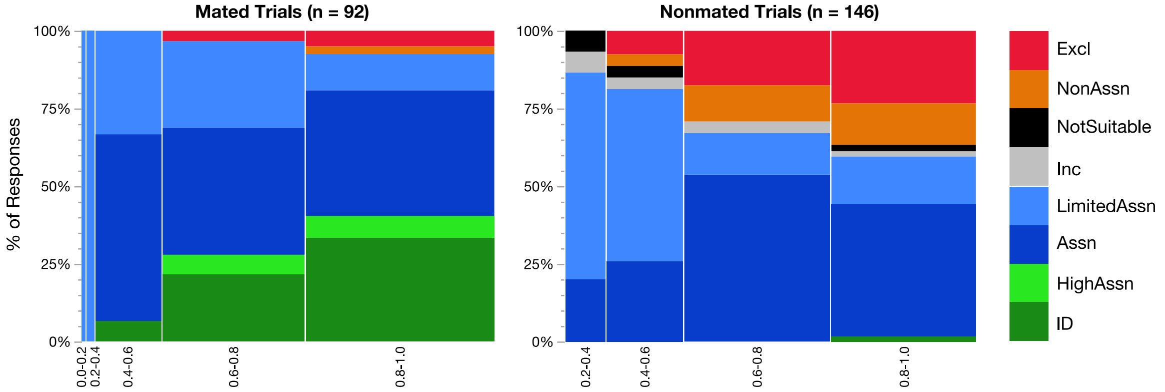


Figure 3. Decision rates based upon all presentations (PRES) by normalized quality score (QQNorm) for the 238 comparison responses.

2.7 Additional Results

The main focus of this study was assessing the accuracy and reproducibility of tire impression decisions reported by practicing FTETs, but various additional assessments were also performed, the results of which are briefly summarized here and reported in detail in [1].

Effect of Dimensionality: Two-dimensional impressions yielded more definitive conclusions than 3-dimensional impressions, and therefore they had higher rates of both true/correct and false/incorrect decisions. On the other hand, the majority of comparisons involving 3D impressions resulted in class associations and therefore these decisions exhibited higher reproducibility (within one conclusion category) than decisions reported on 2D impression comparisons.

Quality vs Dimensionality: We found that dimensionality and quality were moderately correlated (bias-corrected Cramer's V correlation coefficient = 0.2503). Based upon a chi-square analysis, quality was dependent upon dimensionality (p = 0.0008): specifically, 2D impressions were more likely than 3D impressions to yield the highest quality scores (QQ_{Norm} = [0.8-1.0]) and less likely to yield mid-range scores (QQ_{Norm} = [0.4-0.6]). Because these factors are correlated and both are associated with the decisions reported by FTETs, evaluations of differences in decision rates based upon Q factors should consider both quality and dimensionality to avoid potentially confounding effects.

Typicality of Qs: Most participants indicated that the questioned impressions were typical of those encountered during casework (27% of trials indicated "encountered often" and 45% indicated "encountered infrequently"). Three-dimensional impressions were viewed as somewhat more typical of operational casework than 2D impressions. However, decision rates were generally comparable irrespective of the typicality rating; we did not detect an association between typicality ratings and reporting tendencies or performance.

Effect of Difficulty: On mated QKsets, the proportion of ID decisions decreased as difficulty increased from Very Easy to Moderate. On nonmated QKsets, the proportion of class associations decreased as difficulty increased. Although decision rates were associated with both dimensionality and quality, neither of these factors were notably associated with the resulting difficulty ratings.

Limitations Encountered: Overall, 84% of trials deemed suitable for comparison (n=235) listed at least one limitation. The quality/clarity of the Q and background/substrate noise or interference with the questioned impression were the most frequently cited limitations (noted in 55-57% of trials). On definitive decisions, participants were much less likely to indicate that there were limitation(s). Due to study logistics, it was not possible to provide physical items to participants for comparison; however, this was not generally considered a limitation by participants: only 10% of trials listed the lack of known tires and/or lack of original evidence items was a limitation.

3 Study Outcomes

The performance of forensic examiners practicing tire evidence analysis is of interest to both the legal and forensic science communities. In this report, we summarize the results of the only formal black box study of forensic tire impression evidence conducted to date. This study constitutes exploratory research characterizing the accuracy and reproducibility of decisions reported by forensic tire examiners and trainees. Although the scale of this study was relatively small, which limits the precision of the measured rates, the findings remain valuable by providing empirical understanding of a discipline for which no such data has previously been available. In particular, the results of this study help to characterize reporting tendencies on tire impression comparisons, offer insight into decision rates, and evaluate the potential impact of sample characteristics (e.g., quality, dimensionality, substrate/matrix).

3.1 Summary of Study Findings

Research Question #1: How accurate are decisions reported by forensic tire impression examiners?

Just over two-thirds of all decisions reported in this study were class associations (68% *Assn* or *LimitedAssn*), followed by definitive decisions (21% *ID* or *Excl*), probable decisions (8% *HighAssn* or *NonAssn*), and neutral determinations (3% *NotSuitable* or *Inc*). In operational contexts, class associations may be useful for investigation purposes if a definitive decision cannot be reached in a comparison (e.g., for scene-linking, identifying the make and model of a tire, determining the type and size of tire and therefore inferring the type of vehicle that may have left an impression at a crime scene). All nonmated QKsets in this study included questioned impressions that were deposited by tires that were the same make, model, and size as the known tires;

therefore, additional focused research on classification accuracy of tire impressions may be worthwhile and help to further characterize performance in the relevant decision space.

When examiners did report definitive decisions, they were generally correct: 96% of *IDs* and 89% of *Excls* were consistent with ground truth regarding the source of the questioned impression. We observed four errors (3 false *Excls*; 1 false *ID*) and one incorrect decision (1 incorrect *NonAssn*) in this study; this is consistent with both the footwear black box study [8] and the latent print footwear black box study [6], in which false *Excls* were more common than false *IDs*.

Research Question #2: How reproducible are these decisions among examiners?

Regarding the reproducibility of conclusions, 38% of decisions agreed exactly, and an additional 31% were reproduced within one decision category (e.g., ID vs *HighAssn*). Although participants did not reach the exact same conclusion the majority of the time, they rarely reported contradictory opinions: just under 2% of the responses resulted in diametrically opposed decisions (*ID* vs *Excl*). With respect to repeated errors, which would represent instances in which technical review or verification might not flag an error before reporting, none of the three false *Excls* were repeated exactly, but one was reproduced as an incorrect *NonAssn*; the single false *ID* was not reproduced as a *HighAssn*, but this does not necessarily mean that such an error would not be reproduced (exactly or within one decision category) in a larger study.

Research Question #3: What factors are associated with the performance or reporting tendencies of tire impression examiners?

Decision rates, accuracy, and errors were notably associated with questioned impression quality and dimensionality, which were found to be correlated factors. Higher quality impressions and/or two-dimensional impressions were more likely to result in definitive decisions, increasing the rates of both true (TID and TEX) and false (FID and FEX) definitive decisions. Due to the higher proportion of class associations, trials that included lower quality and/or three-dimensional questioned impressions exhibited higher rates of reproducibility—the majority of decisions reported on these trials resulted in *Assn* or *LimitedAssn* (neighboring decisions) and thus were often reproduced within one decision category. These results show that decision rates vary based upon questioned impression attributes, and therefore, the overall rates are not necessarily applicable to all comparisons.

3.2 Limitations and Implications

The relatively low participation in this study underscores several issues with respect to both decision analysis studies as a whole as well as specific to tire examination. First, these studies rely on volunteer participation, and due to IRB restrictions, we cannot mandate participation. Therefore, participation in these studies often competes with casework for examiners' limited time—backlogs, number of cross-trained disciplines, casework demands, other professional responsibilities, and lingering effects on laboratory operations from the COVID pandemic may all contribute to higher demand on examiners' time. Second, specific to the tire examination discipline, the community is smaller relative to other pattern disciplines and because these requests are seen less often in laboratories, participating in tire-related research may be viewed as lower priority. Third, tire comparisons are a much more time-intensive examination process than many other pattern disciplines; the sheer size of tire impressions relative to footwear or fingerprints, for example, generally increases the amount of time it takes to complete a detailed comparison. Lastly, the level of completion for the 17 participants was lower than anticipated: asking for 40 comparisons was clearly an unreasonable expectation for volunteers and any future study should consider reducing the requested level of effort.

The rates measured in this study are intended to serve as overall assessments to inform decision making and guide future research. They should not be taken to be precise measures of operational decision rates. The participants (especially given such a small number) should not be assumed to be representative of the overall population of practicing forensic tire examiners. The samples included in this study should not be assumed to be representative of all operational casework. In addition, the study procedures may constitute a deviation from what is done in standard practice in operational casework—for example, participants did not have access to the physical samples for comparison, the examinations were entirely digital, the reporting was conducted using an online interface, and the conclusion scale used may differ from that used by participants for casework. The results do not account for operational quality assurance measures such as verification or technical review. Given the relatively small study size, all effects should be interpreted with caution. Nonetheless, these results do provide insight into an area that has never been explored in a published study before, and we expect that the results will be of value to practitioners, laboratory managers and decision makers, and the legal community.

In addition to providing the first empirical assessment of forensic tire examiner decisions, this study also produced a rich dataset of tire impression imagery with ground truth attribution, including questioned impressions, known tire tread photographs, and known tire test impressions. This dataset, including all high-quality images and associated metadata, will be made publicly available so that it can be further leveraged by a variety of stakeholders, such as to facilitate future research studies and increase tire impression training examples.

4 Project Artifacts

4.1 Participants and Collaborating Organizations

A total of 39 US and Canadian tire impression examiners and trainees engaged with this research, by registering for an account on the study website. The primary results are a function of 17 FTETs who participated in the study by completing at least one assigned tire impression comparison.

In addition to the study participants, the study team was comprised of the following six personnel, who were jointly responsible for completing the work:

- William J. Chapman (Noblis; principal investigator)
- Connie L. Parks (Noblis; former principal investigator)
- Jan LeMay (Denver Police Department Crime Laboratory; subject matter expert)
- Nicole Richetelli (Noblis; key personnel)
- Kensley Dunagan (Noblis; analyst)
- R. Austin Hicklin (Noblis; Director of Forensic Science Group)

4.2 List of Products and Dissemination Activities

This study resulted in one peer-reviewed publication, one public presentation of results, and two datasets:

- ***Publication***: Richetelli N, LeMay J, Dunagan KM, Parks CL, Hicklin RA and Chapman WJ (October 2023). Accuracy and Reproducibility of Forensic Tire Examination Decisions (Submitted; *Forensic Science International*).
- ***Presentation***: Richetelli N and LeMay J (August 2023). Black Box Evaluation of Tire Impression Evidence Conclusions. *Oral Presentation: International Association for Identification Annual Educational Conference: National Harbor, MD*.
- ***Datasets***: see Section 4.3 below.

4.3 Datasets Generated

This study generated two datasets that will be made publicly available:

- ZIP download of the 77 QKsets that received comparison responses in this study, including 532 full-size high-quality images of the questioned impressions and known tires (tread photographs and test impression scans).
    - *Noblis, Tire Impression Black Box Study Dataset (Oct 2023). DOI: 10.17605/OSF.IO/FRTX4*
- De-identified response data from this study, including responses to the background questionnaire (for all 39 registrants and participants) and comparisons (238 comparisons).
    - *Supplementary Data 1 from: Richetelli N, LeMay J, Dunagan KM, Parks CL, Hicklin RA and Chapman WJ (October 2023). Accuracy and Reproducibility of Forensic Tire Examination Decisions (Submitted; Forensic Science International).*

References

[1] N. Richetelli, J. LeMay, K.M. Dunagan, C.L. Parks, R.A. Hicklin, W.J. Chapman, Accuracy and Reproducibility of Forensic Tire Examination Decisions, Forensic Sci Int. (Submitted) (2023).

[2] National Research Council, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington, D.C., 2009. https://www.ojp.gov/pdffiles1/nij/grants/228091.pdf (accessed March 22, 2023).

[3] President's Council of Advisors on Science and Technology (PCAST), Report to the President. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, (2016). https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf (accessed March 22, 2023).

[4] President's Council of Advisors on Science and Technology, An Addendum to the PCAST Report on Forensic Science in Criminal Courts, Executive Office of the President, Washington, D.C., 2017. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_addendum_finalv2.pdf (accessed March 21, 2023).

[5] National Commission on Forensic Science, Views of the Commission Facilitating Research on Laboratory Performance, 2016.

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

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

[8] R.A. Hicklin, B.C. McVicker, C. Parks, J. LeMay, N. Richetelli, M. Smith, J. Buscaglia, R. Schwartz Perlman, E.M. Peters, B.A. Eckenrode, Accuracy, Reproducibility, and Repeatability of Forensic Footwear Examiner Decisions, Forensic Sci Int. 339 (2022). https://doi.org/10.1016/j.forsciint.2022.111418.

[9] R.A. Hicklin, L. Eisenhart, N. Richetelli, M.D. Miller, P. BelCastro, T. Burkes, C. Parks, M. Smith, J. Buscaglia, P.E. M., R. Schwartz Perlman, J. Abonamah, E.B. A., Accuracy and Reliability of Forensic Handwriting Comparisons, Proceedings of the National Academy of Sciences. 119 (2022) e2119944119. https://doi.org/10.1073/pnas.2119944119.

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

[11] AAFS Standards Board (ASB), Best Practice Recommendation for the Detection and Collection of Footwear and Tire Impression Evidence, ANSI/ASB Best Practice Recommendation 052, First Edition. (2022). https://www.aafs.org/sites/default/files/media/documents/052_BPR_e1.pdf (accessed May 16, 2023).

[12] AAFS Standards Board (ASB), Best Practices for the Preparation of Test Impressions from Footwear and Tires, ANSI/ASB Best Practice Recommendation 021, First Edition. (2019). https://www.aafs.org/sites/default/files/media/documents/021_BPR_e1.pdf (accessed May 16, 2023).

[13] AAFS Standards Board (ASB), Best Practice Recommendation for Lifting of Footwear and Tire Impressions, ANSI/ASB Best Practice Recommendation 049, First Edition. (2020). https://www.aafs.org/sites/default/files/media/documents/049_BPR_e1.pdf (accessed May 23, 2023).

[14] W.J. Bodziak, Tire Tread and Tire Track Evidence: Recovery and Forensic Examination, CRC Press, Boca Raton, 2008.

[15] AAFS Standards Board (ASB), Best Practice Recommendation for Photographic Documentation of Footwear and Tire Impression Evidence, ANSI/ASB Best Practice Recommendation 050, First Edition. (2021). https://www.aafs.org/sites/default/files/media/documents/050_BPR_e1 wErrata1_0.pdf (accessed May 16, 2023).

[16] Noblis, Tire Impression Black Box Study Dataset, (2023). https://doi.org/10.17605/OSF.IO/FRTX4.

[17] Collaborative Testing Services Forensic Testing Program, Tire Track Imprint Evidence Test No. 20-5351/5 Summary Report, 2020. https://cts-forensics.com/reports/20-5351.5_Web.pdf (accessed May 31, 2023).

[18] Collaborative Testing Services Forensic Testing Program, Tire Track Imprint Evidence Test No. 21-5351/5 Summary Report, (2021). https://cts-forensics.com/reports/21-5351.5_Web.pdf (accessed May 31, 2023).

[19] Collaborative Testing Services Forensic Testing Program, Tire Track Imprint Evidence Test No. 22-5351/5 Summary Report, (2022). https://cts-forensics.com/reports/22-5351.5_Web.pdf (accessed May 31, 2023).

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

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

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

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

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

[25] B. McVicker, C.L. Parks, J. LeMay, B. Eckenrode, R.A. Hicklin, A Method for Characterizing Questioned Footwear Impression Quality, J Forensic Identif. 71 (2021) 205–217.

[26] A. Agresti, B.A. Coull, Approximate Is Better than “Exact” for Interval Estimation of Binomial Proportions, Am Stat. 52 (1998) 119–126.