Forensic Chemistry 35 (2023) 100513

Contents lists available at ScienceDirect

Forensic Chemistry

journal homepage: www.sciencedirect.com/journal/forensic-chemistry

Interpretation and report writing in forensic comparisons of paint evidence: An interlaboratory exercise

Andria Mehltretter [a], Meghan Prusinowski [b], Hal Arkes [c], David Flohr [d], Cedric Neumann [e], Scott Ryland [f], Donna Sirk [g], Tatiana Trejos [b,*]

[a] Laboratory Division, Federal Bureau of Investigation, 2501 Investigation Parkway, Quantico, VA 22135, United States
[b] Department of Forensic and Investigative Science, West Virginia University, 1600 University Avenue, Morgantown, WV 26508, United States
[c] Department of Psychology, The Ohio State University, 240N Lazenby Hall, 1827 Neil Ave, Columbus, OH 43210, United States
[d] Retired forensic scientist, Wedowee, AL 36278, United States
[e] Battelle Memorial Institute, Columbus, OH 43201, United States
[f] Retired forensic scientist, Orlando, FL, United States
[g] National Institute of Standards and Technology (NIST), 100 Bureau Drive, Gaithersburg, MD 20899, United States

ARTICLE INFO

Keywords:
Evidence interpretation
Report writing
Paint examinations
Interlaboratory
Verbal scale
Organization of Scientific Area Committees (OSAC) for Forensic Science

ABSTRACT

This interlaboratory study evaluated a guide for interpreting and reporting trace evidence examinations. The online survey aimed to assess the examiners' interpretation of casework scenarios designed by a subject matter expert panel (SMEP), specifically for paint evidence. A pool of 30 scenarios was created, and 15 were assigned to each participant using multi-factor design to evaluate agreement among examiners on case sets with different conclusion ranges and difficulty levels. Exploratory data analysis and three generalized mixed-effects models were used to assess the data. From the 1267 responses received from 85 participants, approximately 93% of responses were consistent between participants and within the SMEP consensus and the next best category, while 73% agreed with the SMEP consensus that was considered the ground truth. Most disagreements were observed in worst-case scenarios created with intended higher difficulty and complex circumstances.

The statistical models showed a strong positive relationship between the reported and expected conclusions, indicating that participants' findings align with the SMEP consensus. On the other hand, the exercise's difficulty level and participant's experience did not have a significant impact on the reported conclusions. However, the credible intervals for the probabilities of the different reported conclusions indicate that more experienced participants achieve greater consensus for a given exercise. The consensus reached among practitioners represents an advance in the trace community's efforts to standardize reporting of results and opinions when following systematic criteria.

Introduction

Forensic practitioners have the critical responsibility to communicate findings to customers in a manner that is objective, balanced, and helpful (i.e., easy to understand by a non-scientific audience and avoids ambiguity). Over the years, multiple studies by forensic practitioners, policymakers, legal professionals, statisticians, specialists in human factors (psychologists), and researchers have investigated effective approaches to communicate information to the relevant audience [1–4].

In general, there is a growing concurrence that Bayesian reasoning can be advantageous to represent the use of forensic observations when making decisions [5–18]. In the Bayesian framework, the probative value of the evidence is evaluated based on two elements: the similarity between the characteristics of the trace and control materials and the specificity (or rarity) of the characteristics observed for the trace material [8]. Alternatively, other approaches have been proposed to provide a qualitative explanation of interpretations in a case report [19–22].

The weakness associated with expressing evidentiary conclusions without explanations about significance was recognized years ago for

* Corresponding author.
E-mail addresses: ahmehltretter@fbi.gov (A. Mehltretter), mnp0006@mix.wvu.edu (M. Prusinowski), arkes@psy.ohio-state.edu (H. Arkes), davidbflohr@bellsouth.net (D. Flohr), neumannc@battelle.org (C. Neumann), sryland@cfl.rr.com (S. Ryland), donna.sirk@nist.gov (D. Sirk), Tatiana.trejos@mail.wvu.edu (T. Trejos).

https://doi.org/10.1016/j.forc.2023.100513
Received 14 April 2023; Received in revised form 5 June 2023; Accepted 12 June 2023
Available online 16 June 2023

forensic paint examinations [19–22]. Ryland wrote in 1981, "However, because of the class evidence nature of most paint evidence, the expert witness must be careful in a statement of interpretations drawn from the analytical results. The difficult responsibility of properly interpreting the meaning of these conclusions should not be avoided, for it is the expert's true reason for being there. Conclusions such as *could have originated from* and *originated from this source or another source having all of the same characteristics* are frequently given to the court at face value, with no attempt to further educate them as to the limited probability of the "did not originate from" or "the other source" possibilities. Although the former statements of these interpretations are technically correct, they tend to underplay the significance of the evidence" [19]. For over forty years, the trace evidence community has tried to fill the void with intermittent discrimination studies providing the means to better interpret the significance [23]. The task is not an easy one, for the profile of manufactured materials is constantly changing, and one study conducted years ago may not be accurate in today's environment. Laboratories have been slow to embrace incorporating significance interpretations in their formal reports in favor of relying on testimony opportunities to discuss the impact.

In response to these issues, Christopher Bommarito introduced a new associative scale at the 2006 U.S. Midwestern Association for Forensic Scientists/Midwest Forensic Resource Centre (MAFS/MFRC) Advanced Trace Evidence Symposium. The scale consisted of five "association types," noted as Type I through Type V. The scale was proposed as a standardized method to convey the various types of possible associations between items of trace evidence. The associative scale sparked interest within U.S. agencies as it was disseminated between 2006 and 2009 at scientific meetings, including SWGMAT, the Association of Forensic Quality Assurance Managers (AFQAM), and the U.S. National Institute of Justice (NIJ) Trace Evidence Symposium. The proposed scale addressed one of the subsequent recommendations (i.e., the need for standardization of terminology and reporting) identified in the National Research Council (NRC) report on the scientific foundations of forensic science [20].

Renewed focus on report wording and content was brought to bear by the NRC, U.S. National Academies, report in 2009 [20,21]. They noted several shortcomings with these abbreviated types of statements. First, forensic scientists' traditional audiences are not trained to understand the subtle nuances in report wording. Discussions during meetings of the Scientific Working Group for Materials Analysis (SWGMAT) led forensic practitioners to realize that concepts such as *could have*, *cannot be eliminated*, *indistinguishable*, and *consistent* are interpreted differently by different practitioners (personal communication, Andria Mehltretter). As a result, differences in these concepts may go unnoticed by an attorney or layperson. Second, most of these statements only express the level of similarity between the trace and control materials without accounting for the rarity of the observed characteristics in the relevant population, and thus, do not convey the full probative value of the forensic observations.[1]

Other initiatives were launched to improve the objectivity, fairness, and clarity of forensic conclusions to address the recommendation from the NRC. For example, beginning June 2010, the Paints and Polymers group of the Federal Bureau of Investigation (FBI) Laboratory started issuing reports that included an interpretation scale for any comparisons made between evidentiary items of paints, tapes, and polymeric materials. The entire scale was added to the report for context, together with a justification as to why a particular conclusion was reached. This new report structure detailed the analytical techniques applied to the evidence, assigned a level to the conclusion, and provided a justification for the assignment. This version of the scale was presented at SWGMAT in 2010, renewing discussions on the topic. An interpretation subcommittee was put together to standardize this approach for the trace evidence discipline.

In 2014, the SWGMAT working group was superseded by the Trace Materials Subcommittee of the Organization of Scientific Area Committees (OSAC) for Forensic Science. OSAC was and remains sponsored by the U.S. National Institute of Standards and Technology (NIST) and aims at strengthening forensic science by facilitating the development and promoting the use of high-quality, technically sound standards. In 2015, the OSAC Trace Materials Subcommittee formed an Interpretation Task Group to develop consensus-based guidelines for the interpretation and reporting of trace evidence.

The use of a verbal scale to report trace evidence was proposed. The verbal scale and recommended practice for its use became a draft OSAC interpretation guide [22] that has evolved over time based on this study and feedback. Each iteration of the document has (a) provided guidance to forensic examiners to standardize the interpretation of comparative examinations of trace evidence and (b) described the items that should be included in the report to aid the reader in interpreting the reported results. The document has focused primarily on fibers, glass, hair, paint, and tape but can be applied to other trace materials. The document was developed by a multidisciplinary working group, including administrators, practitioners, researchers, lawyers, statisticians, and human factors specialists. The document addresses concepts such as *exclusionary differences*, *indistinguishable*, *discrimination*, *match* versus *fit* terminology, and *rarity*, to mention some. The understanding of these terms by the end-users is fundamental to minimizing the ambiguity in forensic conclusions and their interpretation (see reference [22] for terminology).

The categories of the proposed verbal scale were defined based on a review of the literature and data-driven approaches for decision-making. [22] These categories are based on a) the foundational validity of the scientific methods used for the comparison of the items, b) discrimination capabilities of the analytical protocol, and c) existing knowledge of the specificity of the analytical observations based on survey studies, reference collections, or databases. The scale can easily be supplemented by error rates and formal statistical methods, if the foundation is available, to provide a fully quantitative approach. In other words, the proposed scale promotes a means to use data to support forensic scientists' opinions.

The scale is supported by multiple examples for each type of material (glass, paint, tape, fibers, and hair). These examples provide an extensive list of factors that can affect category choice in the scale of possible conclusions. Examples of casework scenarios that correspond to the different categories of conclusion and the associated report writing terminology are included. An effort was made to include scenarios that represent realistic casework situations and data to facilitate the application of the interpretation ranges, minimize ambiguity, and offer a platform to standardize the decision mechanisms to arrive at conclusions. The document enforces consistency between the materials-specific sections and in the way the probative value of the evidence is supported and reported [22]. A copy of the version distributed at the time of the study is included in the supplemental section.

Some of the most important aspects of the standard interpretation guide are that it:

1. Incorporates important factors to be considered when forming a conclusion;
2. Offers material-specific examples;
3. Provides specific guidelines on how to use the different factors when assessing the significance of observations;
4. Provides a context by including the entire range of possible conclusions, enabling all end-users to properly assess the significance of the forensic findings.

[1] From a logical point of view, the "could" statement expresses similarity and specificity but requires accounting for non-forensic case information to be justified. This may be deemed inappropriate. Furthermore, to be useful, "could" statements need to be relativized by a sense of the magnitude of the number of other sources that "also could" have generated the trace material.

In this study, we evaluated the extent of consensus in the interpretation of forensic observations and associated conclusions dealing with forensic paint examinations. Our study featured an interlaboratory experiment that relied on mock case scenarios that closely represented paint evidence cases. Participants were asked to use the standard interpretation guide created by the Interpretation Task Group of the OSAC Trace Materials Subcommittee to form and report their conclusions and justify their choices. Specifically, the study aimed to assist in answering the following questions:

1. Does the proposed approach assist in the standardization of interpretation and report writing for the forensic science community and its customers?
2. Is the document sufficiently clear and intelligible for easy application by practitioners?
3. Are there areas of the document that need to be improved or further developed?
4. Are the practitioners able to reach a consensus in their conclusions when interpreting the mock cases?

This study evaluated the data from over 1200 responses provided by 85 forensic practitioners over the two-month survey period to assess the robustness of the scale and application of the OSAC interpretation guide.

Material and methods

Study overview

The exercise consisted of an electronic survey of paint case-like scenarios provided to participants to interpret observations and form conclusions following the provided interpretation guide. A Google Docs form was set up as a questionnaire divided into sections with data tracking based on the participant's chosen conclusion.

The experiment was designed with input from a statistician, a psychologist, practitioners, and researchers. It considered circumstances representative of actual cases, as well as appropriate sample sizes and selection of the number of variables for statistical interpretation of the data. Also, human factors were considered during the description of the scenarios and elaboration of the survey questions to minimize bias in decision-making processes. A pool of 30 case-like scenarios was used in this study to simulate casework information and data a paint examiner would typically encounter. The scenarios included two levels of difficulty (easy, difficult) and five of the categories from the scale of conclusions proposed in the guide [Association with Highly Discriminating Characteristics (AHD), Association with Discriminating Characteristics (ADC), Association with Limitations (AL), Inconclusive (IN), and Exclusion/Elimination (EX)].

The scenarios were coded for assignment to the participants. Each participant received 15 scenarios in a multi-factor design set to accommodate a minimum of 60 participants. The assignment of the scenarios to each participant was random, conditioned on the following constraints:

1. Each participant would receive at least one scenario of each conclusion category.
2. The scenarios were organized so that no individual would receive all "difficult" or all "easy" scenarios.
3. The overall experiment would be mostly balanced in terms of the number of times each scenario, conclusion category, and level of difficulty was used.

With those conditions, a script coded in R (version 4.0.1) was used to randomly select and organize 60 individual sets of 15 scenarios. Separate Google Docs Form surveys were prepared for each participant based on the scenario selection. Additional participants beyond 60 received the same sets in order (i.e., participant 61 received the same scenarios as participant 1).

Each participant was assigned a unique identifier. Each survey was carved into sections, including three main segments:

1. **Introduction of the survey and instructions for completing the task:** The instructions requested the participant to download, review, and understand the proposed interpretation guideline before completing the exercise. Participants were to interpret the case-like scenarios based on the content of the provided guidance document and its specific examples related to paint evidence. Although they could use their own knowledge, experience, and intuition, they were asked to follow the interpretation guide to generate their responses.
2. **Respondent demographics questions:** The participants were asked to answer 12 demographic questions. These questions captured the respondents' background, expertise, and work environment. The demographic questions are provided in the supplemental section.
3. **Scenarios and Response Options:** Fifteen scenarios were provided to each respondent, with response options linked to various follow-up questions depending on what was selected for the conclusion category. An example of the scenarios and the specifics on the response and selection options are provided in the supplemental section.

Participants had one week (7 days) to complete and submit their responses; however, they could edit their responses as many times as necessary during that time frame. It was expected that all the work and responses would be made independently and without input from others.

Scenario development

Four subject matter experts were involved in developing the exercise scenarios to provide a variety of perspectives. The first task for the group was to draft the scenarios to cover a range of paint types, conclusion types/strengths, and difficulty levels. A scenario was deemed difficult if a similar example was not explicit in the guide, or when the factors included were borderline between two conclusion categories. Some considerations during drafting scenarios included writing style, whether to specify the analytical techniques, whether to provide data and in what format. Each author initially drafted approximately 20 scenarios, including an easy and difficult one for each conclusion type. These scenarios were compiled and circulated amongst the four members of the Subject Matter Expert Panel (SMEP).

Each author independently evaluated the scenarios drafted by their colleagues to assess the conclusion type and difficulty level. The group then met on numerous occasions to evaluate the scenarios regarding how each person arrived at a conclusion type, how difficult each of the scenarios was, and what edits were needed to better align content and writing style. As is common practice in studies considering consensus-based decisions, the scenarios in which the most consensus occurred between the SMEP members were selected for further deliberation and final editing [24,25]. For each scenario, a conclusion was considered the consensus answer or ground truth if all four SMEP members agreed to the same conclusion. The best scenarios were chosen for each conclusion type and difficulty level, ensuring that all combinations were covered. Several of the scenarios were slightly edited (e.g., color or chemistry of paint layers) to create new scenarios to fill any gaps. The final 30 scenarios are provided in the supplemental section.

Note that "physical fit" was removed as a possible conclusion type since it is extremely difficult to show or describe a physical fit without unintentionally biasing the expected answer; however, participants were not aware of this and could select that conclusion if they thought it was the appropriate response. Furthermore, references to specific techniques were removed unless they were necessary for evaluation. This was done to focus the study on the interpretation scale and not the choice of analytical techniques. Likewise, raw data was not included to keep data analysis from being a variable.

Interlaboratory study participants

The OSAC Trace Materials Subcommittee distributed advertisements through various venues to recruit participants to complete the exercise. In the Spring of 2020, the OSAC quarterly Newsletter, the OSAC monthly Standards Bulletin, the OSAC website, and OSAC "In Brief" bulletin distributed an invitation. The study was also announced to the American Society of Trace Evidence Examiners (ASTEE) and during the Center for Forensic Science Research and Education CFSRE 2020 Online Trace Evidence Symposium.

The intended participants for this study included trace evidence practitioners with experience in conducting paint examinations. This included participants who were actively working cases at the time of recruitment, in training, working in academia or industry, or former paint examiners (e.g., retired, shifted to management position).

The participation was voluntary, anonymous, and all work was to be done independently. A total of 117 participants registered for the study, with 102 completing training and receiving the survey. Of these 102 participants, 91 responses were received (8 responses were not received, and 3 participants withdrew). Six participants were further removed from consideration as they either did not report experience with paint examination or because they provided an incomplete scenario response, resulting in a total of 85 participants considered for the survey evaluation.

Exercise administration

Before participants received access to their exercise and response form, they were required to attend a 1-hour live session or recorded training session. Participants had to declare they had completed the training before receiving their individualized exercise links via a Training Completed notification form received by OSAC. In general, the following was covered during the training session: a) the goals of the exercise, b) a review of the guidance document, c) instructions on how to proceed, including reviewing the guidance document and relevant references in the trace review article, d) a demonstration of the Google Docs Form, and e) an in-depth review of the scenarios structure and the various selection offerings.

After completion of the virtual training, the participant received their survey and exercise based on their unique Participant Identification Number. The participants also received reference material from the training, slides, and the interpretation guide. Participants were free to evaluate each scenario and move forward and backward through the scenarios and change answers as needed until their allotted week (7 days) was finished.

Data analysis

Demographic information was evaluated from the survey, and the conclusions and comments from each exercise were collected and combined to assess the response rates and evaluate what factors contributed most to the accuracy of the responses (concerning the expected conclusions determined by the SMEP).

Exploratory and descriptive data analyses were used to help visualize trends in the data and percent agreement among participants and the consensus expected responses. The agreement with the consensus was evaluated in three ways (e.g., whether the participant reported the same interpretation category as the SMEP, a category that lies within the SMEP consensus and the closest category, or within the SMEP ranked best intended category).

Formal statistical analysis was conducted by using generalized linear mixed effect models. Raw and transformed conclusions reported by the participants were used as dependent variables. The expected conclusion and level of difficulty of each exercise, and years of experience of the participants were used as independent variables. The ordinal nature of the variable representing the reported and expected conclusions and the years of experience was preserved during the analysis. The analysis accounted for random variations between participants. Samples from the posterior distributions of the models' parameters were obtained by Hamiltonian Monte Carlo sampling [26] as implemented in Stan [27].

Results and discussions

Experimental design considerations and participant demographics

The demographic questions of the survey requested information from the participants, including their number of years of experience, the materials they analyzed most, whether they used a verbal scale for reporting conclusions, and whether they participated in or gave feedback on the development of the interpretation guide. These factors were evaluated to determine what influence they may have on reporting the results.

The responses from the 85 trace examiners were coded to transform the response into numerical inputs. The two-difficulty level, the five categories of conclusions, and the six levels of experience of participants were coded into numerical metrics used for further statistical analysis. The experiment was designed to be mostly balanced between the three factors for a minimum of 60 participants. An imbalance resulted from the additional number of participants and those who did not complete the 15 scenarios, but it is unlikely to impact the results given the overall number of responses ($n = 1267$) and the chosen method of convergence of the statistical analyses. An example of the random assignment of cases to participants is shown in the supplemental section.

Of the 85 respondents, 85% had more than six years of experience in trace evidence (Fig. 1A), and about half reported more than 15 years of experience. These demographics may reflect current trends in the U.S. practice in trace evidence, where there is a large population of senior scientists with vast expertise and a relatively low number of new hires in the specific trace evidence disciplines related to this study. Another factor considered was whether the participants were active practitioners, were in training, or had worked cases in recent years. Most participants reported that they were actively working cases, as seen in Fig. 1B. Approximately, 79.5% of the participants were from the United States, 16% were non-US, and 4.5% were unknown, although this question was not directly addressed in the survey and therefore not analyzed as a variable in this study.

Interestingly, the correct classification rates, when evaluated in comparison to the SMEP expected interpretation category, did not seem to be influenced by years of experience or current use of a verbal scale. On average, 73% of the participants selected the anticipated interpretation category, with no evident trend as the years increased or whether or not participants used a verbal scale in their laboratory (Fig. 2). A detailed description of how the performance rates were estimated is discussed in the following section.

Additional questions were used to evaluate the participants' familiarity with the document or other applicable scales related to the accuracy of the responses. Most respondents were not involved in developing the guidance document (Fig. 3A). However, about half of the respondents used some form of verbal scale for their conclusions and reports (Fig. 3B). This indicates an interest in the community for a reporting language that expresses the significance of the findings.

Assessment of performance measures

Table 1 summarizes the number of responses per interpretation category and difficulty level compared to the SMEP panel consensus. In the assignment of case combinations per examiner, an equal number of overall responses were anticipated per interpretation category and difficulty level among all participants, regardless of each volunteer's different proportions of case combinations of categories and difficulties. However, since more individuals were recruited than expected (60 targeted, 85 received) and not everyone responded to the intended

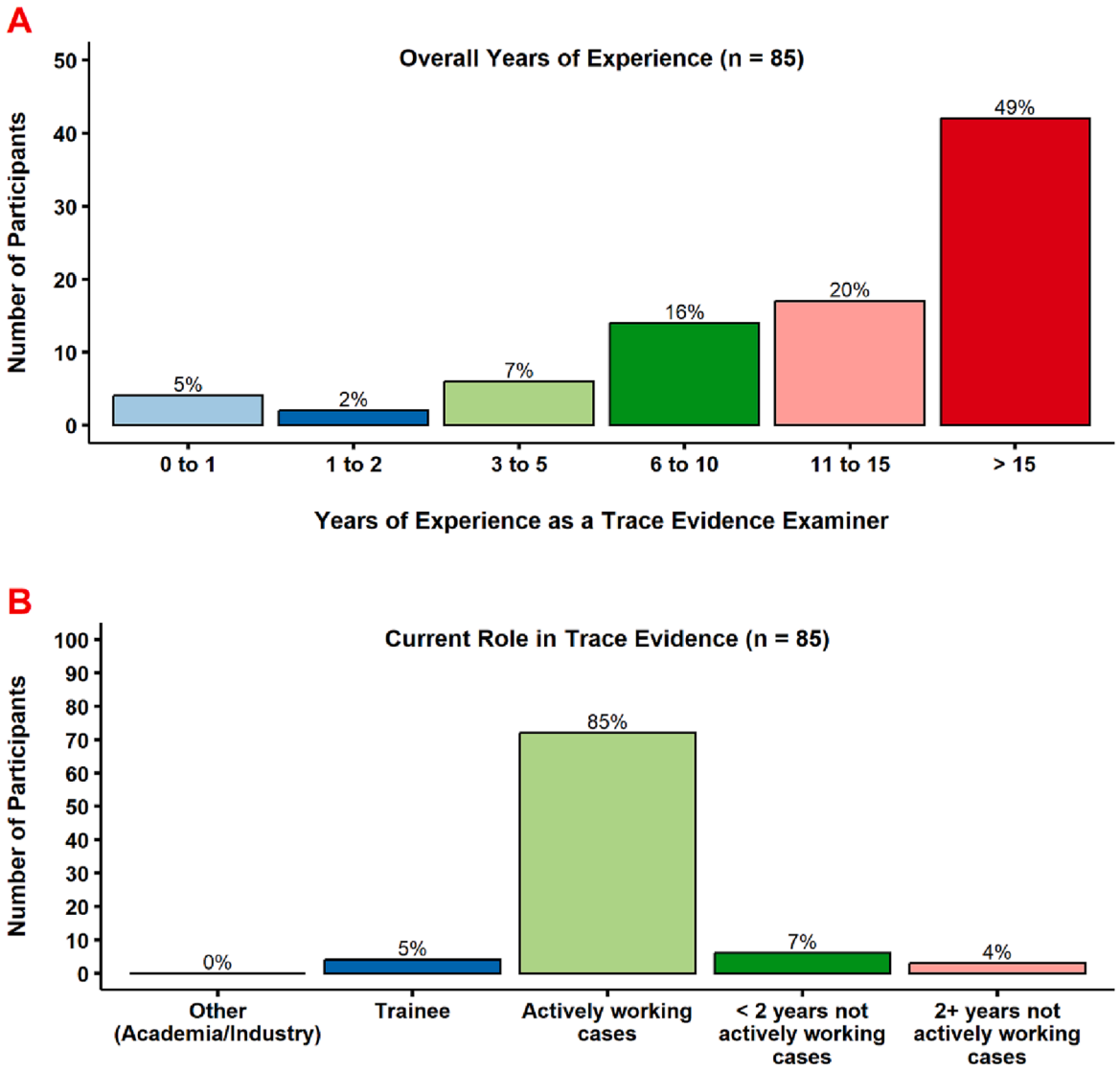

Fig. 1. Bar graphs showing the distribution of responses to demographic questions. A: Distribution of experience in trace evidence among respondents. B: Distribution of current occupation in trace evidence.

category across all their 15 scenarios, the number of responses per category varies slightly (Table 1). Therefore, the dataset of 1267 responses from 85 participants was used to evaluate the participant performance rates.

Fig. 4 displays the percent of responses reported by participants for each of the five expected SMEP-determined categories. For each interpretation category, it was evident that most of the participants agreed with the SMEP, with the next category above or below the expected classification generally being the second most common response. Also, Table 1 shows that for 28 out of the 30 scenarios (except scenarios 08–40 and 23–11), the participants reached a consensus among themselves that corresponded with the consensus reached by the SMEP. This indicated that independent of the "ground truth" established by the SMEP, most participants (93%) consistently reported the same conclusion. This reflected a high level of agreement among participants and can be used as an indicator of the effectiveness of the interpretation guide to assist with the standardization of criteria and thought processes to interpret data.

Performance rates were further evaluated using each of the following criteria to estimate the percent of correct or incorrect responses, with the correct answer defined in one of three ways:

a) The first approach captured whether the participant reports the same interpretation category as the SMEP.
b) The next approach expanded the correct answer to include one category higher and one lower from the expert panel's consensus.
c) For the third approach, responses were ranked by the SMEP from best to worst as determined for each specific circumstance per scenario. In other words, there were instances when the next best answer was not the neighbor on the scale. Then, performance measures were estimated as the closeness of the participant's responses to the best-intended category.

When using the first criterion of estimating the correct responses as the agreement of the examiner with the SMEP, 73% of the total responses were consistent with the expected consensus conclusion. Using the second criterion, 94% of the responses were within one interpretation category difference. When evaluating the distribution of performance rates by interpretation category, the categories of *Association with Highly Discriminating Characteristics (AHD)* and *Association with Limitations (AL)* had the highest percentage of correct responses. In contrast, the *Association with Discriminating characteristics (AD)* and the *Inconclusive (IN)* had the lowest percentage of correct responses (see Fig. 5A). When using the plus/minus one category, the number of correct

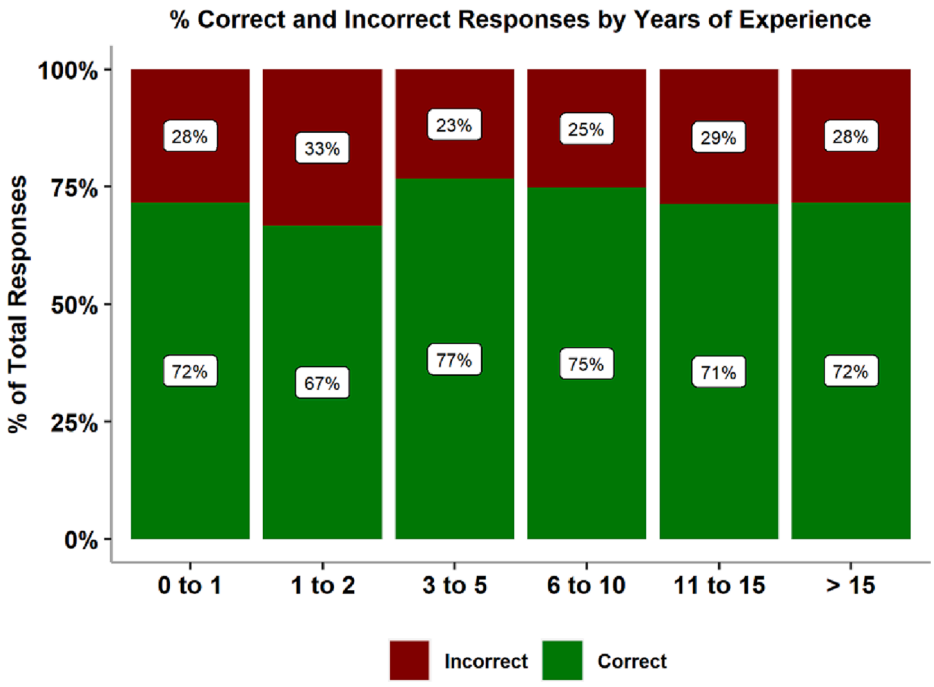

Fig. 2. Distribution of correct and incorrect responses (compared to the expected SMEP response category) by years of experience.

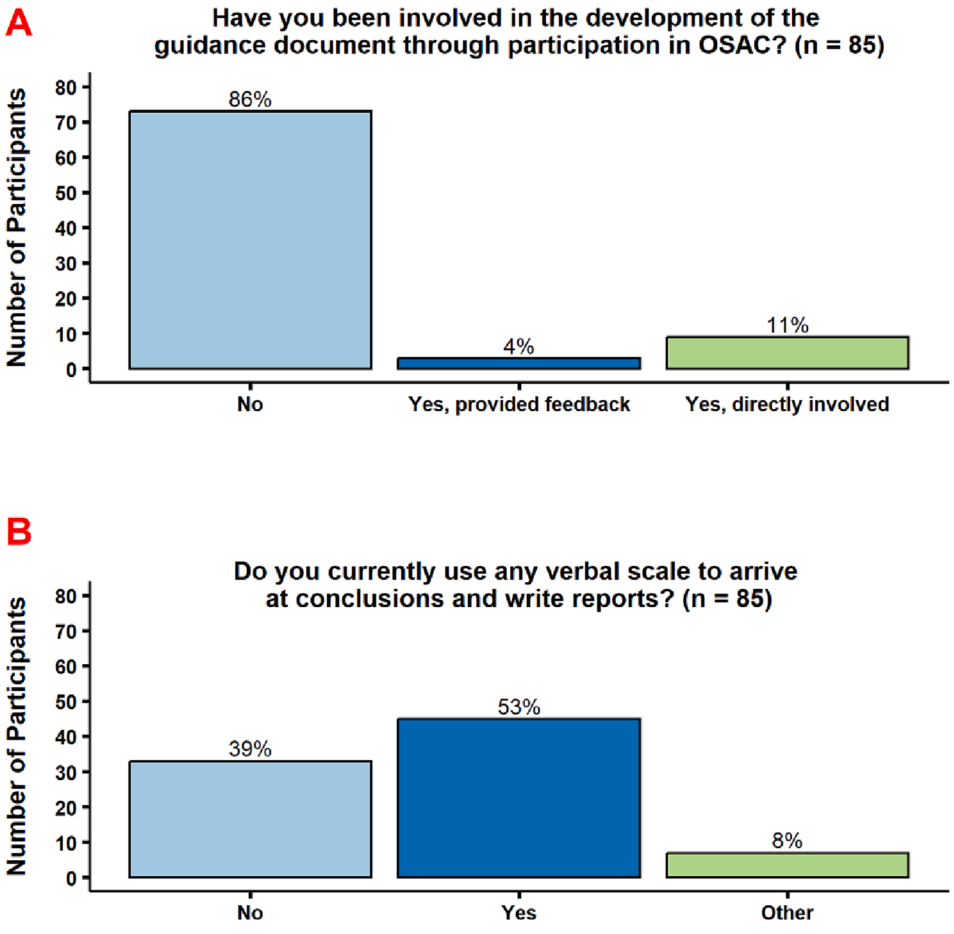

Fig. 3. Bar graphs showing the distribution of responses to demographic questions. A: Distribution of responses for involvement in developing the standard guide for interpretation and report writing. B: Distribution of the use of any type of verbal scale by participants. The "other" category included participants that were considering a verbal scale, but one was not yet implemented in their agency, or individuals who used one in a former job but not at the lab where they were employed at the time of the study.

Table 1
Description of the cumulative responses to all scenarios per category compared to the SMEP-assigned conclusion category and difficulty level.

Scenario ID	AHD	AD	AL	IN	EX	Total	SMEP Conclusion	SMEP Difficulty
01–19	32	4				36	**AHD**	**Difficult**
02–61	36	7	3	2	1	49	**AHD**	**Difficult**
03–27	24	10		5	1	40	**AHD**	**Difficult**
04–02	39		1			40	**AHD**	**Easy**
05–54	43	4				47	**AHD**	**Easy**
06–59	33	6	6			45	**AHD**	**Easy**
07–18	1	23	17			41	**AD**	**Difficult**
08–40		11	32	1		44	**AD**	**Difficult**
09–58	18	20	2		1	41	**AD**	**Difficult**
10–05	4	42				46	**AD**	**Easy**
11–17	6	33			1	40	**AD**	**Easy**
12–37		27	14			41	**AD**	**Easy**
13–31		4	41			45	**AL**	**Difficult**
14–42	1	9	33	1		44	**AL**	**Difficult**
15–78	2	14	24			40	**AL**	**Difficult**
16–14		3	44			47	**AL**	**Easy**
17–16		1	33	8		42	**AL**	**Easy**
18–21			38	2		40	**AL**	**Easy**
19–32			6	35	1	42	**IN**	**Difficult**
20–38			9	31	2	42	**IN**	**Difficult**
21–43			4	35		39	**IN**	**Difficult**
22–64		5	16	20	1	42	**IN**	**Easy**
23–11			25	16	1	42	**IN**	**Easy**
24–25			16	22	5	43	**IN**	**Easy**
25–1					45	45	**EX**	**Easy**
26–8		2	1		39	42	**EX**	**Easy**
27–26			1		41	42	**EX**	**Easy**
28–60		12	4	9	20	45	**EX**	**Difficult**
29–68	11	2	2	4	20	39	**EX**	**Difficult**
30-60B		10	2	7	17	36	**EX**	**Difficult**
Expected Total *	**257**	**253**	**258**	**250**	**249**	**1267**		
Total Received	**250**	**249**	**374**	**198**	**196**	**1267**		

*Expected Total represents the number of responses that would have been received for each conclusion category had all participants reported the intended conclusion.

responses generally increased across all categories, with higher percentage of incorrect responses on the two interpretation categories at the extremes of the scale (Fig. 5B).

In general, the number of "incorrect" conclusions (as in conflict with the SMEP consensus) for all categories of conclusions increased with the level of the assigned difficulty of the case. The inconclusive category was the only exception where incorrect responses did not increase with difficulty (Fig. 6). Difficult cases were the ones where the combination of differences and similarities in the observed features or the relevant case information regarding the transfer of the items was more likely to prevent more assertive conclusion categories. Thus, a greater level of consensus was observed for this difficult-inconclusive category of conclusions than in easier cases.

It was noted that the most variation resulted from the scenarios rated as difficult by the SMEP. Two of the scenarios, 02–61 and 03–27, were intended to result in an Association with Highly Discriminating Characteristics (AHD) conclusion, while scenarios 28–60, 29–68, and 30-60B were expected to be Exclusion/Elimination (EX). Therefore, we discuss what can be learned from these complex "worst" case scenarios.

Scenario 02–61 was a scenario involving architectural paint of which the questioned (Q) paint had a sequence of four layers that corresponded to layers within the known (K) but was only a partial structure compared to the K; specifically, the Q did not contain multiple lower layers, nor the topmost layer represented in the K. Of the corresponding layers between the Q and K, each of the four presented a different color and chemistry than the other three, which is an example of an Association with Highly Discriminating Characteristics listed in Section 9.5.2.3 of the version of the interpretation guide provided: *Architectural paint system with two or more different layers*). We speculate that the participants who chose Association with Discriminating Characteristics used the missing layers in the Q as the reason to lower the strength of the chosen conclusion. When writing the scenario, the SMEP did not think that the missing layers were reasons to lower the strength of the conclusion or to exclude because 1) the Q was removed from a prybar, so it could reasonably be understood why underlying layers were present on the K and not the Q (they hadn't been removed during the prying), and 2) the top layer on the Q was heavily damaged so the top missing layer might also not have remained on or transferred when it was separated from the K source. The SMEP reasoned that one must consider the nature of the transfer when evaluating differences between the Q and the K, keeping in mind that the K is purposely selected to represent all layers present on the K source.

Scenario 03–27 involved bicycle paint in which the Q layers corresponded to the K, except that the K had white and yellow overspray particles on its surface, whereas the Q only had white overspray. Similar to Scenario 02–61, participants may have used that missing component as the reason to reduce the strength of the conclusion, whereas the SMEP did not. The rationale of the SMEP was that there is still some overspray present in the Q, which leads to an increase in the strength of the conclusion. In this situation, one must consider that the overspray deposits are random and that the questioned sample may be of limited size. Bicycle paints were not explicitly categorized in the interpretation document, so there was some complexity involved in reaching a conclusion on this scenario.

Regardless, for both 02–61 and 03–27, the SMEP response (AHD) and the participant response (AD) were both Associations of Evidence with Class Characteristics, and reports can further explain the rationale of why one association type was chosen over another. More familiarity with the interpretation guide by users of the document might improve future consensus rates in difficult scenarios, particularly if discussion sessions are encouraged within and between agencies prior to implementation in their quality management system.

Automotive paints were the focus for the Exclusion/Elimination (EX) Scenarios 28–60, 29–68, and 30-60B. For these, the Q had an additional layer present on it that was not present in the K. For scenarios 28–60 the chemistry of the Q suggested an additional factory layer, while for 30-60B, the chemistry of the Q Layer 1 suggested that it also could be a

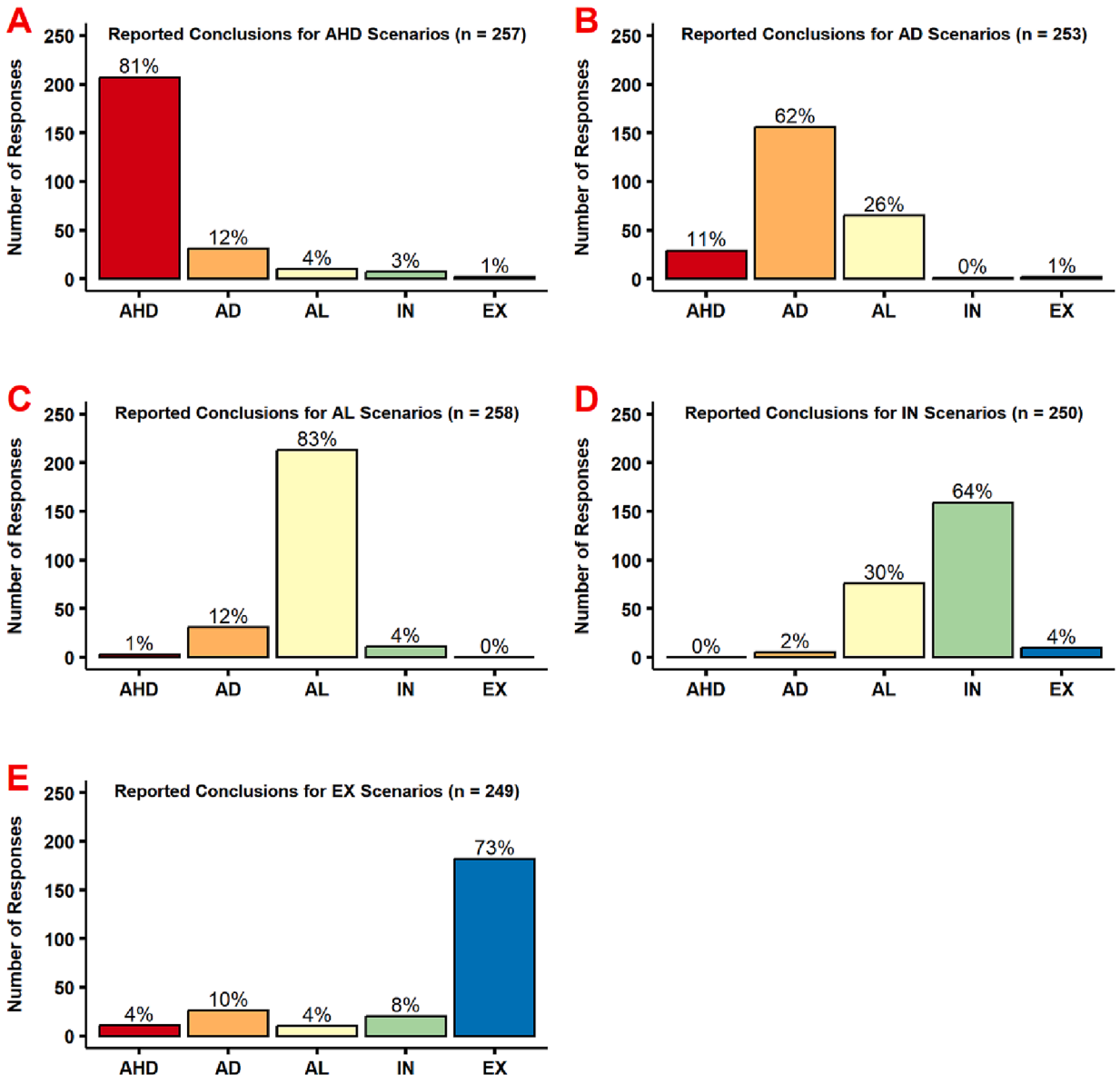

Fig. 4. Percent of responses reported by participants per intended panel consensus category. A: Association with Highly Discriminating Characteristics (AHD), B: Association with Discriminating Characteristics (AD), C: Association with Limitations (AL), D: Inconclusive (IN), and E: Exclusion/Elimination (EX).

refinish paint applied over the original finish. On the other hand, Scenario 29–68 had an additional internal refinish layer present within the layer structure. (Q, Layer 3). We speculate that the participants either (1) did not thoroughly read the paint section of the interpretation document to the end, (2) were reluctant to exclude based on the many other similarities between the samples and disregarded the direction in the interpretation document, or (3) did not have enough experience to identify if the chemical compositions were indicative of factory applied or refinish systems.

The final text of the paint section read as follows, which specifically addressed this type of scenario:

> *"9.7.6 Vehicles can have different paint systems on different panels of the same vehicle. Therefore, it is possible that one known vehicle part differs from the questioned sample and is eliminated as a possible source, but that enough similarities exist to warrant the request of additional samples from other parts of the vehicle to determine whether the entire vehicle should be eliminated."*
>
> *"9.7.6.1 Example report wording. The unknown paint did not come from that area of the vehicle where the known sample was taken (Exclusion/Elimination). Vehicles can have different paint systems on different panels of the same vehicle. Further comparisons can be performed if additional known samples are submitted."*

As is expected with this type of interlaboratory exercise, the feedback received from the participants led to improvements to the interpretation guide. For example, the discrepancy observed between SMEP and the participants in scenarios 28–60, 29–68, and 30-60B triggered some modifications to the guide. The OSAC Trace Materials Polymer Task Group re-evaluated the guide and recommended that this type of scenario should result in an "Exclusion with Limitations (EL)" conclusion. This conclusion category was not previously thought to be applicable to paint, and therefore was not used in this study. The newest version of the interpretation guide, which went through subcommittee ballot and adjudication of public comments, includes this conclusion category. At the time of this writing, the definition is as follows: "Exclusion with Limitations – The item exhibits differences from the comparison sample that support that it did not originate from the source, as represented by the comparison sample; however, limiting factors prevented an Exclusion (Elimination) from being reached. This provides more support for

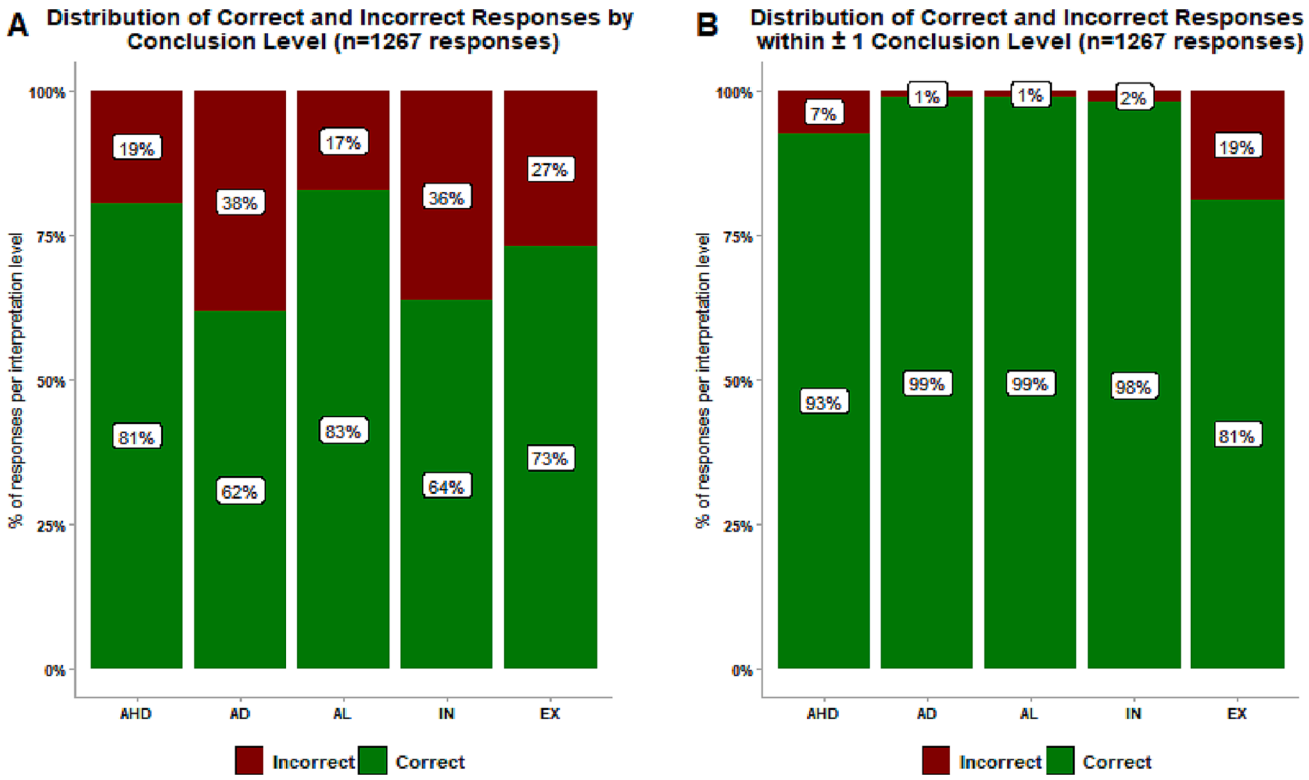

Fig. 5. Represents the percentage of correct and incorrect responses for each interpretation category. A: correct responses as per agreement with the SMEP consensus, B: correct responses as per agreement within one interpretation category of the SMEP intended response.

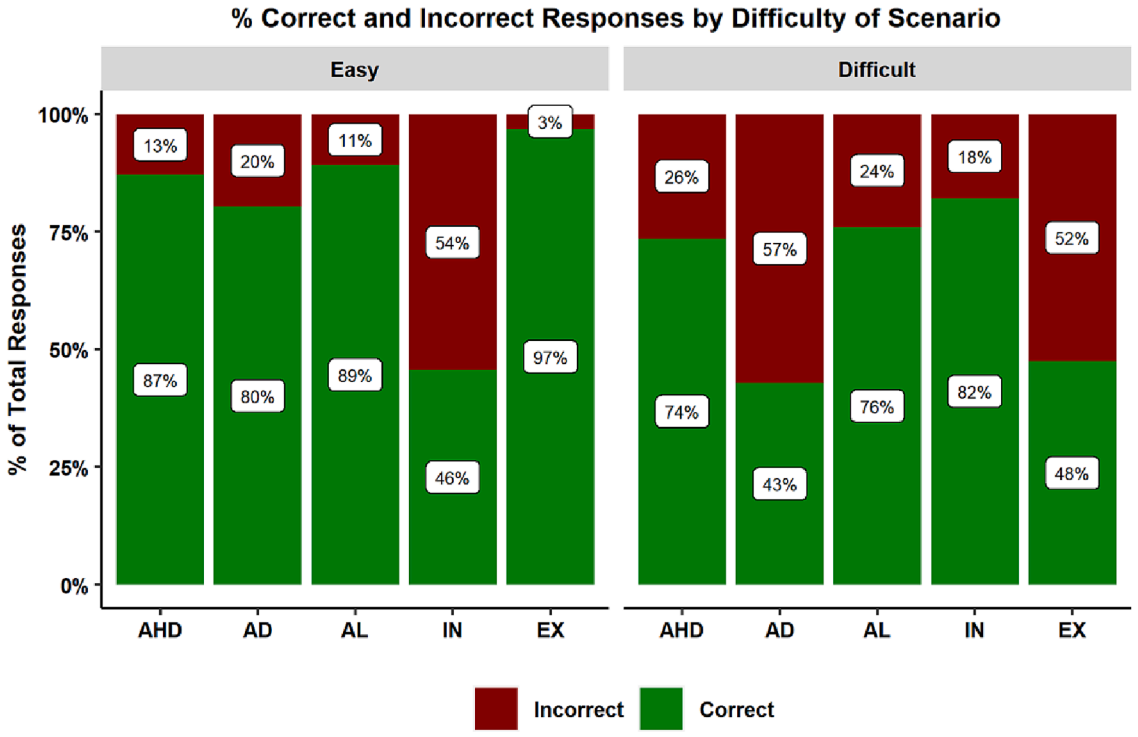

Fig. 6. Percent of correct (in agreement with SMEP consensus) and incorrect (in conflict with SMEP consensus) responses per level of difficulty.

the proposition that the items originated from different sources as opposed to the same source."

As applied to paints, an EL example was added to the guide that reads "Vehicles can have different paint systems on different panels of the same vehicle. Therefore, it is possible that one known vehicle part differs from the questioned sample (e.g., one sample has an anti-chip layer). No analytical differences were observed in the corresponding layers. The vehicle is no longer available to collect an additional known sample.".

It was noteworthy that the proposed interpretation criteria were not being implemented in most crime laboratories. Thus, most participants (85%) reported not being familiar with the guidance document. It is expected that increased understanding and familiarization will enhance consistency in using the scale under various case circumstances.

The administrators of this study obtained indications from some participants regarding why they chose a given category using the survey responses[2] and a live discussion session with some test takers at the NIST-OSAC subcommittee. When analyzing the basis for answers that did not correspond to the consensus of the SMEP, it was noticed that some resulted from not following the guide criteria. Others were dependent upon the test taker's judgment, the scenario complexity, or resulted from a lack of clear instructions from the OSAC interpretation guide. During these discussions, it appeared that the adjacent category to the consensus answer from the SMEP (e.g., AL vs. AD, or IN vs. EX) was not always the second-closest correct conclusion. As a result, the SMEP ranked each scenario response from the best possible answer (e.g., panel consensus) to the worst possible answer. This was conducted first independently, and then the panel met to evaluate consensus in the weighted responses. A weighted "correctness" matrix on the interpretation categories was created for all scenarios. Ranks ranged from 0 to 4, with 0 corresponding to the consensus answer and 4 to the furthest from the consensus interpretation. The sign of the rankings was also documented, with a + sign indicating the conclusion was a higher category than the consensus response and a - sign indicating the conclusion was lower than the consensus (Table 2).

Table 2
Table describing the weighted rankings of each conclusion category as determined by the SMEP. Each possible category was assigned a value between 0 and 4, with 0 corresponding to the consensus response and four being the furthest from the consensus interpretation. A + sign indicates the conclusion is a higher category than the intended response, and a - sign indicates the conclusion is lower than the intended response.

Scenario ID	SMEP Intended Conclusion	Rankings of each possible category				
		AHD	AD	AL	IN	EX
01–19	**AHD**	0	−1	−4	−4	−4
02–61	**AHD**	0	−1	−3	−4	−4
03–27	**AHD**	0	−1	−3	−3	−4
04–02	**AHD**	0	−3	−4	−4	−4
05–54	**AHD**	0	−3	−4	−4	−4
06–59	**AHD**	0	−1	−3	−4	−4
07–18	**AD**	3	0	−1	−3	−4
08–40	**AD**	4	0	−1	−3	−4
09–58	**AD**	1	0	−2	−4	−4
10–05	**AD**	2	0	−4	−4	−4
11–17	**AD**	2	0	−4	−4	−4
12–37	**AD**	4	0	−1	−3	−4
13–31	**AL**	4	1	0	−3	−4
14–42	**AL**	4	2	0	−3	−4
15–78	**AL**	4	1	0	−3	−4
16–14	**AL**	4	2	0	−3	−4
17–16	**AL**	4	4	0	−2	−4
18–21	**AL**	4	4	0	−2	−4
19–32	**IN**	4	4	2	0	−3
20–38	**IN**	4	4	2	0	−3
21–43	**IN**	4	4	3	0	−3
22–64	**IN**	4	3	1	0	−3
23–11	**IN**	4	4	1	0	−2
24–25	**IN**	4	4	1	0	−2
25–1	**EX**	4	4	4	4	0
26–8	**EX**	4	4	4	4	0
27–26	**EX**	4	4	4	4	0
28–60	**EX**	4	3	2	1	0
29–68	**EX**	3	4	2	1	0
30-60B	**EX**	3	4	2	1	0

This approach considered the flexibility afforded by the interpretation scale. Although most scenarios clearly corresponded to one of the conclusions, there were circumstances in which some scenarios could not easily fit into a single category. In these situations, minor considerations could lead the expert opinion to lean towards one category or the other. Either choice could be considered correct when well justified and documented.

On the other hand, there were circumstances where some conclusions were not possible or appropriate at all for some scenarios. For instance, scenario 01–19 involves a vehicle hit-and-run fatality. Two large questioned (Q) fragments recovered from the scene were compared to paint from the hood of the suspect pickup truck (K). All fragments were indistinguishable after comprehensive analytical comparisons of replicates of the corresponding paint layers in the Q and K samples. The K and Q fragments had the following color/layer structure and composition:

1. Very weathered blue alkyd topcoat
2. Gray alkyd primer
3. Light gray alkyd-epoxy primer
4. Bright red polyester spot putty
5. Colorless to pale pink polyester body filler

In this example, repainted/refurbished automobile paint with five layers and weathered features on the top layer were compared. The OSAC interpretation guide indicated that, based on the rarity of the features and characteristics of K/Q paint fragments, the comparison of the K and Q fragments were associated with highly discriminating characteristics (AHD), which was the highest category of the scale of conclusions. In this example, AHD was assigned a 0 as it corresponded to the consensus of the SMEP.

The only other acceptable conclusion for scenario 01–19 was an Association with Discriminating Characteristics (AD) (corresponding to −1 since it is one category lower than the highest category), which would be relatively reasonable since the examples in the OSAC interpretation guide did not address entire layer structures of an aftermarket refinish. The OSAC interpretation guide clearly excluded any other interpretation category in the circumstances of this scenario, and therefore the remaining categories were all ranked with a "4".

On the other hand, the problematic case 28–60 represented a hit-and-run where three multi-layered paint fragments (Q) were recovered from the debris collected from the victim's garments and compared to the known (K) samples collected from an entire area of the right front fender of a vehicle of interest. They exhibited the following color/layer structure and composition:

Q fragments' layer structure	K fragments' layer structure
1. Colorless acrylic-urethane clearcoat	Missing / not present
2. Colorless acrylic-melamine-styrene clearcoat	1. Colorless acrylic-melamine-styrene clearcoat
3. Green -blue acrylic-melamine basecoat containing two types of decorative flake	2. Green -blue acrylic-melamine basecoat containing two types of decorative flake
4. Medium red-brown epoxy-polyester-melamine primer	3. Medium red-brown epoxy-polyester-melamine primer
5. Dark gray epoxy-urethane primer	4. Dark gray epoxy-urethane primer

Comprehensive analytical comparisons of replicates of the corresponding paint layers in the Q and K fragments revealed that Q layer 2 was indistinguishable from K layer 1, Q layer 3 was indistinguishable from K

[2] Although in many instances the justifications from the participants were not provided or were too brief for the authors to understand their reasoning.

layer 2, Q layer 4 was indistinguishable from K layer 3, Q layer 5 was indistinguishable from K layer 4.

Given the differences in the layer structure and the absence of the first top layer on the K fragments, the consensus from the SMEP and most of the respondents was that the interpretation led to the exclusion of the K, submitted as the source.

The SMEP considered the closest possible interpretation in this scenario to be inconclusive (+1) because the questioned sample had an extra layer, not in the known, so they were different. Still, the examiner may not have made a decision of association or exclusion because of the other marked similarities. So, they simply could not say, one way or the other.

The next option of "Association with Limitations (AL)" was given a + 2. A possible explanation was that the OEM layers match, but there was a missing layer in the K; so, it was still an association but with the limitation that there was an extra layer in the Q. Although unlikely, it is possible the examiner considered the possibility that the first layer may have cleaved off the K.

An "Association with Discriminating characteristics (AD)" was given a + 3, assuming the examiner considered the matching OEM automotive layers and ignored the extra layer in Q. However, one cannot simply ignore an extra layer in the questioned sample without an explanation. This was not supported in the interpretation document. Finally, the worst possible conclusion was considered (+4), which none of the participants reported for this scenario.

Fig. 7 shows the results of responses by the consensus interpretation category in a graded color scale. For example, dark green represents agreement with the ranked SMEP and moves to light green, yellow, orange, and red the more it distances from the defined ground truth. Using this alternative approach, on average, 73% ± 10% of the responses agreed with the SMEP, and 89% ± 5 % of the responses fall within the consensus agreement and the next best possible category (rankings 0 and 1 combined).

These combined findings show that a high level of agreement was achieved among practitioners regarding the significance of results in comparative examinations when using the OSAC interpretation guide. Moreover, those instances of disagreement were narrowed down to two leading causes:

1) The understanding of the "*source*" by the participants may have been unclear (e.g., whether it was the whole vehicle, or the sample submitted as the known), resulting in a lack of consensus. This led to improvements to the OSAC interpretation guide to remove any ambiguity.
2) Possibly participants did not understand or misinterpreted the interpretation guide. The impact of this type of issue can be minimized as more training, discourse among practitioners, and familiarization with the interpretation approach is offered to the end-users.

Statistical analysis

Formal statistical analysis was conducted to investigate the effect of experience and scenario complexity on the divergence of the participants' answers from the consensus of the SMEP. The statistical analysis was performed using three generalized mixed effects models. All models were of the following general form:

$$Pr(Y_{i,j} \leq y) = f(X'_{i,j}\beta)$$

where:

Y_{ij} was an ordinal variable representing the category of the conclusion of exercise i by participant j (or the amount of divergence from the consensus conclusions);

y *was* an index corresponding to the categories of the scale of possible conclusions (or the different levels of divergence from the consensus conclusion). Note that the indices are ordered;

$f()$ was the softmax function,

X_{ij} was a vector of independent variables for exercise i and participant j. In this analysis, X_{ij} includes the ID and years of experience of participant j, and the consensus conclusion and level of difficulty of exercise i as determined by the SMEP.

β *was* a vector of parameters.

The first model studied the raw reported conclusions as a function of the independent variables. Thus, we had. $Y_{i,j} \in \{EX, IN, AL, AD, ADH\} \forall i,j$

Table 3 indicates that the SMEP consensus conclusion was a strong positive predictor of the participants' reported conclusion (positive mean, small relative standard deviation and entirely positive 89% credible interval). In other words, there was a strong positive relationship between the reported and SMEP consensus conclusions. On the

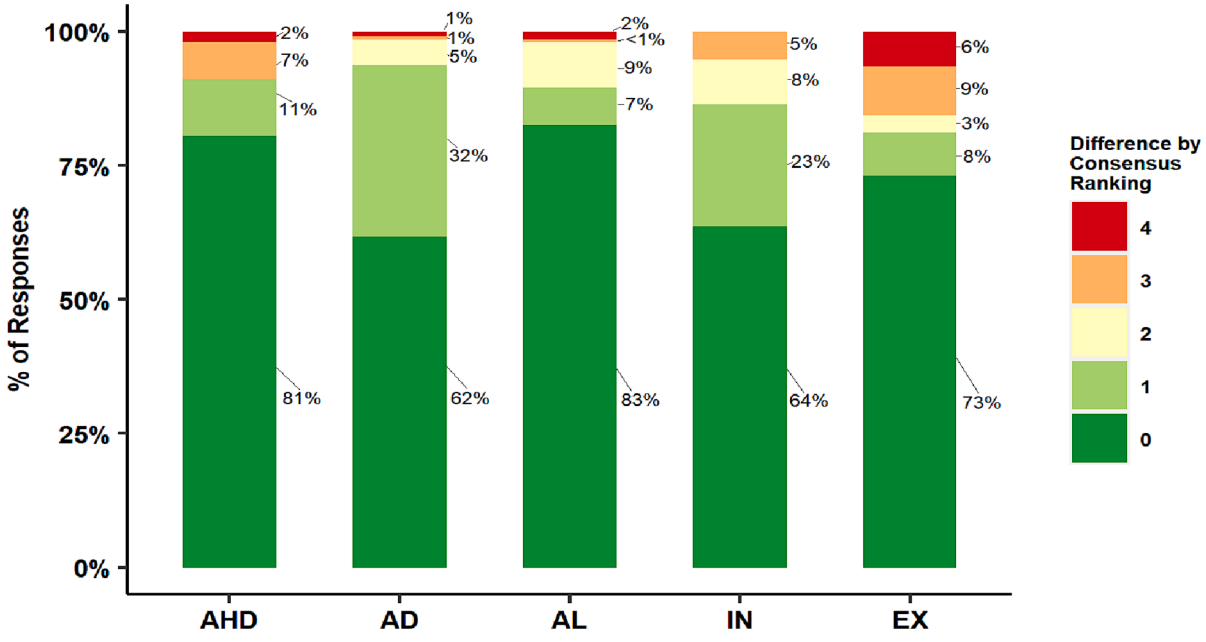

Fig. 7. Distribution of weighted responses by interpretation category. Ranking of correct weight coded by color from 0 (best) to 4 (worst).

Table 3
Mean, standard deviation, and 5.5% and 94.5% quantiles (89% credible interval) of the posterior distributions of the regression parameters for the SMEP consensus conclusion and level of difficulty of the exercises, and participants' years of experience.

	Mean	SD	5.5%	94.5%
$\beta_{conclusion}$	7.805	0.268	7.379	8.235
$\beta_{difficulty}$	0.097	0.107	−0.073	0.267
$\beta_{experience}$	−0.143	0.243	−0.542	0.229

other hand, the level of difficulty of the exercises seemed to have little impact on the reported conclusions (the credible interval is mostly positive, but 0 is a likely value for the parameter). Similarly, the level of experience of the participants did not seem to have any predictive value (the credible interval is mostly negative but 0 is a likely value).

The counterfactual plots in Fig. 8 show the effect of the three independent variables of interest. While the probability of the reported conclusion was strongly influenced by the expected SMEP conclusion, the exercise's level of difficulty and participant's experience did not have a major impact. Interestingly, the credible intervals for the boundaries between the probabilities of the different reported conclusions were wider when participants had less experience than when they were experienced. This seems to indicate that more experienced participants achieved greater consensus on the conclusion for a given exercise.

The second model studied the divergence between the reported and consensus conclusions as a function of the same independent variables as in the previous model. In this model, $Y_{ij} \in \{0, 1, 2, 3, 4\} \forall i, j$, where 0 indicated no difference between reported and expected conclusions, 1 indicated that the reported and consensus conclusions are in two successive categories in the set $\{EX, IN, AL, AD, ADH\}$, 2 indicated that the reported and consensus conclusions are one category apart in the set of possible conclusions, and so on.

Table 4 indicates that there was a negative relationship between the

Table 4
Mean, standard deviation, and 5.5% and 94.5% quantiles of the posterior distributions of the regression parameters for the consensus conclusion and level of difficulty of the exercises, and years of experience of the participants.

	Mean	SD	5.5%	94.5%
$\beta_{conclusion}$	−0.725	0.194	−1.039	−0.419
$\beta_{difficulty}$	0.753	0.125	0.553	0.954
$\beta_{experience}$	−0.354	0.259	−0.782	0.045

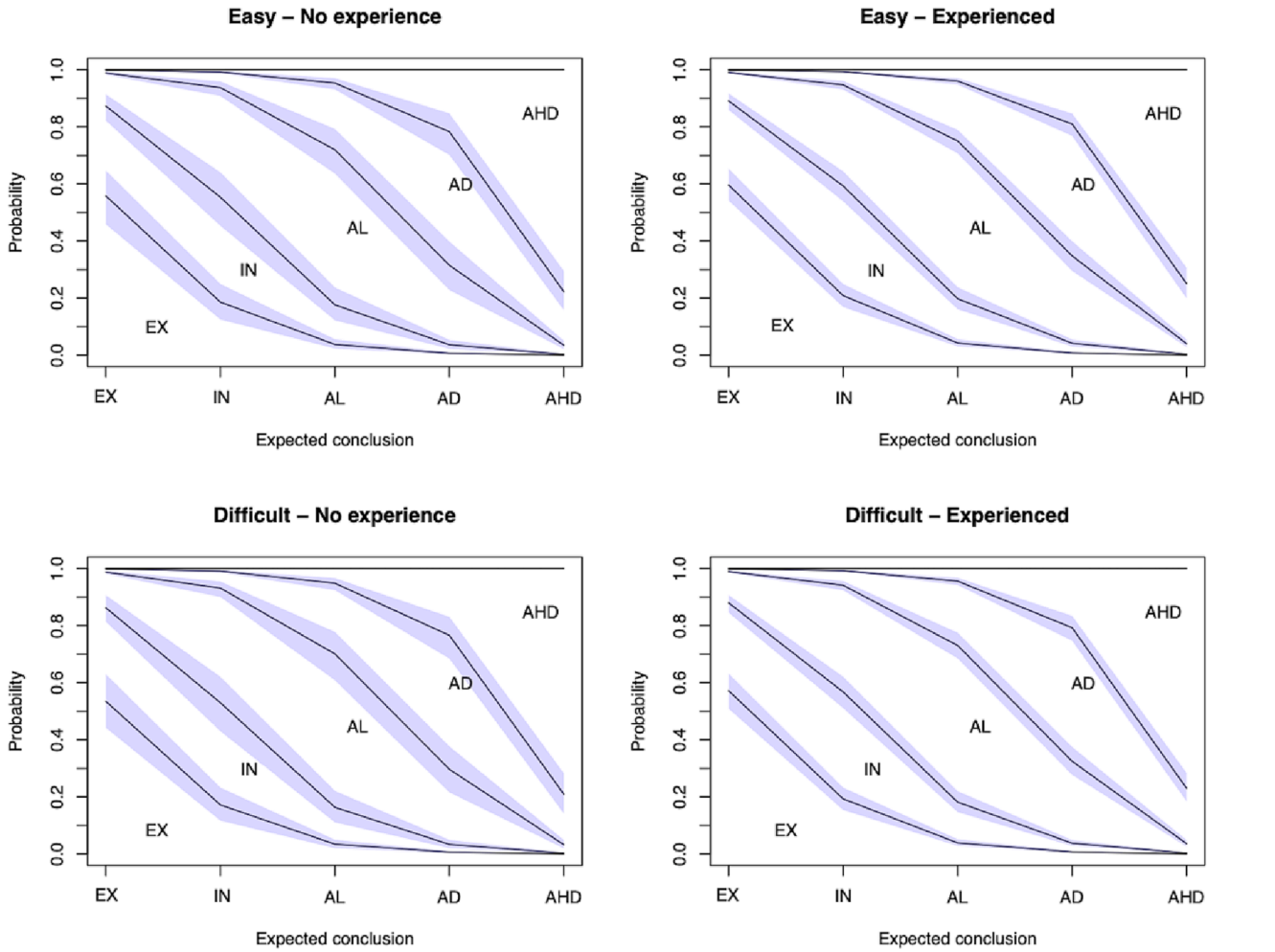

Fig. 8. Counterfactual plots showing the model's predictions for the probability of reporting the different levels of the conclusion scale as a function of the consensus conclusions, level of difficulty of the scenarios (easy vs. difficult), and the participants' experience (no experience vs. most experienced). The solid lines represent the predicted value of $Pr(Y_{ij} \leq y)$ in the different conditions, the shaded areas represent the 89% credible interval of $Pr(Y_{ij} \leq y)$. The intervals between the solid lines represent $Pr(Y_{ij} = y)$. For example, the probability that a reported conclusion would be an exclusion given that the consensus conclusion is an exclusion for an easy scenario and an inexperienced participant (upper left plot) was between 0.47 and 0.65 (with an expected value of approximately 0.58). By extension, the probability that a reported conclusion would be inconclusive in the same scenario would be between 0.82 and 0.65 = 0.17 and 0.91—0.47 = 0.44.

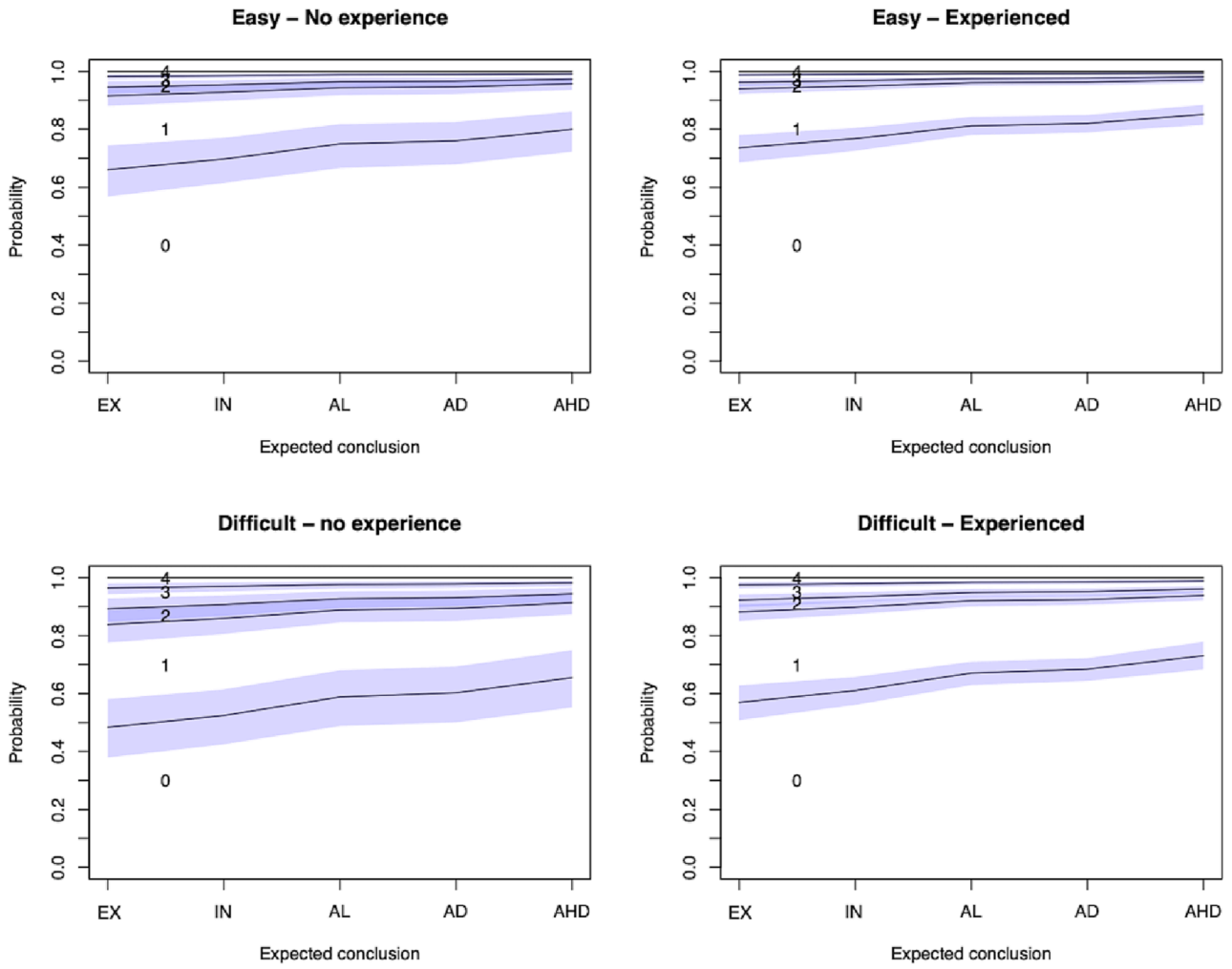

Fig. 9. Counterfactual plots showing the model's predictions for the probability of the reported and consensus conclusions diverging by 0,1,2,3,4 levels, as a function of the consensus conclusions, level of difficulty of the scenarios (easy vs. difficult) and the experience of the participants (no experience vs. most experienced). The solid lines represent the predicted value of $Pr(Y_{i,j} \leq y)$ in the different conditions, the shaded areas represent the 89% credible interval of $Pr(Y_{i,j} \leq y)$. The intervals between the solid lines represent $Pr(Y_{i,j} = y)$.

expected conclusion and the divergence between expected and reported conclusions. The counterfactual plots in Fig. 9 show that the probability of a divergence between consensus and reported conclusions decreased as the level of the consensus conclusion increased. Thus, there was a greater chance of consensus between the participants and the SMEP for the exercises that were meant to result in strong associations, and there was a greater chance of discrepancy for the exercises that were meant to result in exclusion. This confirmed the result presented in Section 3.2.

While the level of difficulty of the exercises impacted the probability of a divergence between expected and reported conclusions, the participants' level of experience did not seem to impact the divergence directly (See Table 5). However, Fig. 9 shows that the level of experience of the participants seemed to have the same effect as in the previous model: there appeared to be less consensus between less experienced participants than between more experienced ones as observed by the wider shaded credible intervals in less experienced. Finally, the last model studied the SMEP-defined divergence (see Table 2) between the reported and consensus conclusions as a function of the same independent variables as in the previous model. In this model, $Y_i \in \{0, 1, 2, 3, 4\} \forall i$, where the possible values were the absolute values of the divergence weights reported in Table 2.

We noted that the values for the parameters of the second and third model are very similar. Thus, we expected the effect of the different independent variables on the weighted divergence to also be similar, which was confirmed by the counterfactual plots in Fig. 10. When comparing Figs. 9 and 10, it seemed that the weighted divergence magnified the potential lack of consensus between less experienced examiners when compared to experienced ones.

Table 5
Mean, standard deviation, and 5.5% and 94.5% quantiles of the posterior distributions of the regression parameters for the expected conclusion and level of difficulty of the exercises, and years of experience of the participants.

	Mean	SD	5.5%	94.5%
$\beta_{conclusion}$	−0.682	0.190	−0.990	−0.380
$\beta_{difficulty}$	0.728	0.126	0.526	0.929
$\beta_{experience}$	−0.291	0.250	−0.699	0.096

Conclusions

A significant collaboration from forensic paint examiners made it possible to collect nearly 1300 responses from assessing 30 case scenarios and analytical information. The survey was designed to comply with various ideal characteristics for an interlaboratory study:

1. It included a coordination body responsible for the organization and distribution of surveys.

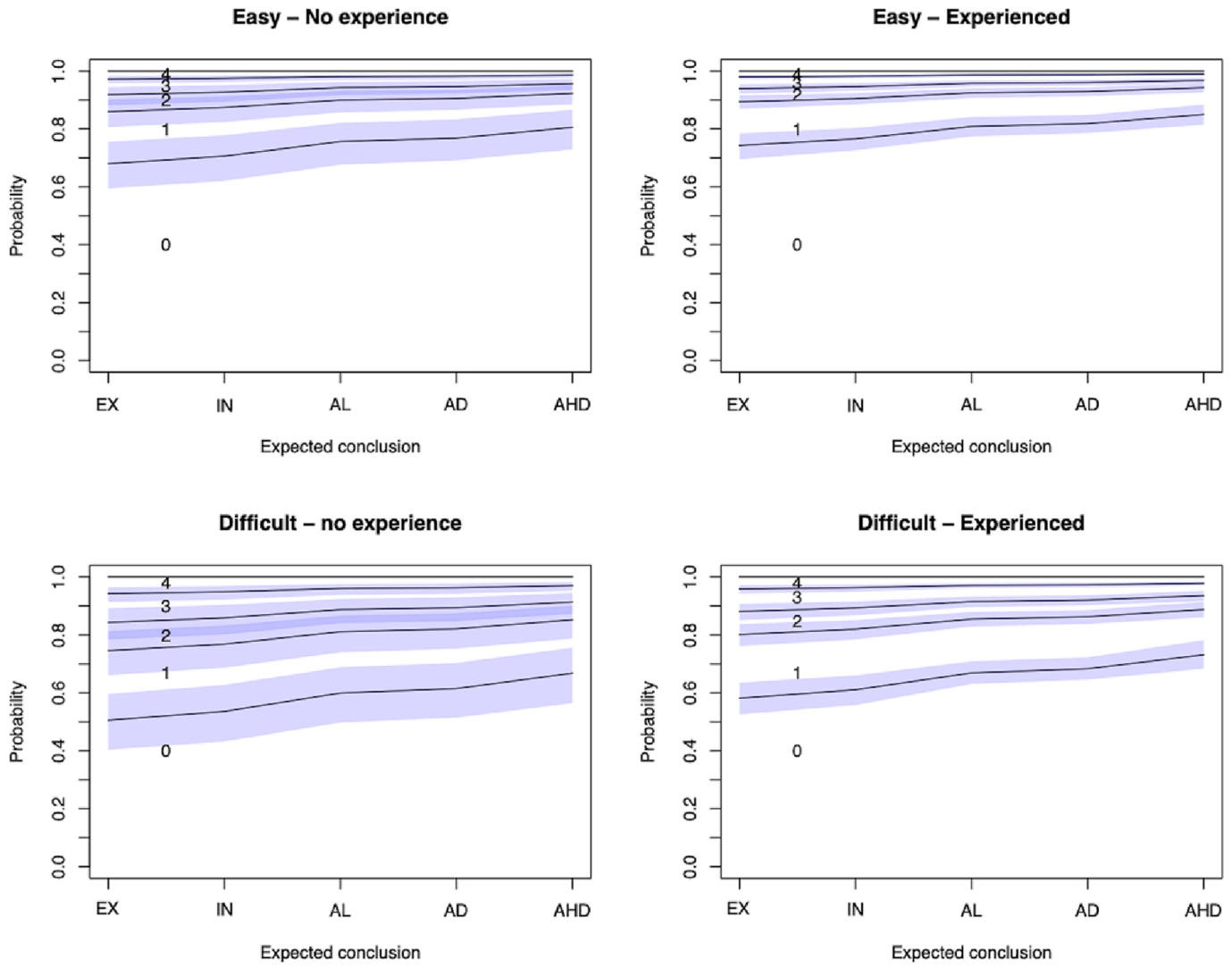

Fig. 10. Counterfactual plots showing the model's predictions for the probability of the reported and consensus conclusions diverging by 0,1,2,3,4 as defined by the SMEP, as a function of the consensus conclusions, level of difficulty of the scenarios (easy vs. difficult) and the experience of the participants (no experience vs. most experienced). The solid lines represent the predicted value of $Pr(Y_{i,j} \leq y)$ in the different conditions, the shaded areas represent the 89% credible interval of $Pr(Y_{i,j} \leq y)$. The intervals between the solid lines represent $Pr(Y_{i,j} = y)$.

2. It was comprised of samples that represented casework as closely as possible, including various difficulty levels, worst-case scenarios, and diverse case types.
3. The study had a known ground truth established as consensus determined by the SMEP, which was kept blind to participants.
4. The performance of the method was evaluated by forensic practitioners with experience in paint examinations.
5. The exercise was set up to measure uncertainty by assessing performance rates and showed repeatable and traceable results that can lead to consensus standards.

This study employed various exploratory data analysis and generalized mixed-effects models to assess the effects of conclusions categories, difficulty level, and practitioner's experience in reaching an agreement between participants and with the SMEP. Regardless of the model utilized, the interlaboratory study results demonstrated that a high level of agreement was achieved among practitioners regarding the significance of evidence in comparative examinations when using the criteria in the proposed interpretation guide. The generalized mixed models suggested that greater consensus on the conclusion for a given scenario was achieved among more experienced participants.

One of the most valuable aspects of an interlaboratory exercise such as this is that the process leads to improved decision-making and communication channels, community engagement, and cultural change. The discussions surrounding these scenarios, both from expert panels and participants, created lively debates in the community regarding case scenarios practitioners have encountered. Participation in these exercises can expand an individual practitioner's perspective and experience and improve the forensic community's collective experience.

Finally, by developing the exercise and evaluating the results, we have established consensus criteria that lead to transparency and consistency in the discipline. The impact of this research on the trace community is critical in evolving our practice, as it permits us to establish a scientific basis to support an examiner's opinion and provide means to replace vague reporting results with consensus language that communicates the significance of the evidence. These collaborative studies will eventually lead to forensic science customers receiving better service and more explicit information for decision-making.

Declaration of Competing Interest

The authors declare that they have no known competing financial

interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments and disclaimers

This is publication number 21.42 of the FBI Laboratory Division. Names of commercial manufacturers are provided for identification purposes only, and inclusion does not imply endorsement of the manufacturer, or its products or services by the FBI. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the U.S. Government.

NIST Disclaimer: Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

The authors want to thank Korina Menking Hoggatt for initial support with the edits of scenarios, Christopher Bomarito and Jose Almirall for their insights, and the SWGMAT and OSAC Interpretation and Discipline Task Groups for their contributions in the interpretation of trace materials.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.forc.2023.100513.

References

[1] C. Neumann, D.H. Kaye, G. Jackson, V. Reyna, A. Ranadive, Presenting quantitative and qualitative information on forensic science evidence in the courtroom, Chance 29 (1) (2016) 37–43, https://doi.org/10.1080/09332480.2016.1156365.

[2] C Aitken, R Paul, J Graham. Communicating and Interpreting Statistical Evidence in the Administration of Criminal Justice – Practitioner Guide No. 1. *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings*. London: Royal Statistical Society, London. 2010. https://www.maths.ed.ac.uk/~cgga/Guide-1-WEB.pdf.

[3] I. Dror, J. Kukucka, Linear Sequential Unmasking-Expanded (LSU-E): A general approach for improving decision making as well as minimizing noise and bias, Forensic Sci. Int.: Synergy 3 (2021), 100161, https://doi.org/10.1016/j.fsisyn.2021.100161.

[4] J. Keijser, H. Elffers, Understanding of forensic expert reports by judges, defense lawyers and forensic professionals, Psychol. Crime Law 18 (2) (2012) 191–207, https://doi.org/10.1080/1068316100373674.

[5] I.J. Good, Weight of Evidence: A Brief Survey, Bayesian Stat. 2 (1985) 249–270. https://www.cs.tufts.edu/~nr/cs257/archive/jack-good/weight-of-evidence.pdf.

[6] R. Cook, I. Evett, G. Jackson, P. Jones, J. Lambert, A model for case assessment and interpretation, Sci. Justice 38 (1998) 151–156, https://doi.org/10.1016/S1355-0306(98)72099-4.

[7] G Vignaux, B Robertson. Bayesian Methods and Court Decision-Making. In: A Mohammad-Djafari, G Demoment. (eds) Maximum Entropy and Bayesian Methods. *Fundamental Theories of Physics* (An International Book Series on The Fundamental Theories of Physics: Their Clarification, Development and Application), **1993**. Vol 53. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-2217-9_12.

[8] C. Aitken, F. Taroni, A. Biedermann, Statistical interpretation of evidence: Bayesian analysis, in: Encyclopedia of Forensic Sciences, Academic Press, 2013, pp. 292–297, https://doi.org/10.1016/B978-0-12-382165-2.00195-1.

[9] C. Kruse, The Bayesian approach to forensic evidence: Evaluating, communicating, and distributing responsibility, Soc. Stud. Sci. 43 (5) (2013) 657–680, https://doi.org/10.1177/0306312712472572.

[10] J. Curran, T. Champod, J. Buckleton, Forensic Interpretation of Glass Evidence, CRC Press (2000), https://doi.org/10.1201/9781420042436.

[11] G Zadora, A Martyna, D. Ramos, C. Aitken. Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data. *John Wiley & Sons Inc*, **2014**. DOI: 10.1002/9781118763155.

[12] M Kwan, KP Chow, F Law, P Lai. Reasoning About Evidence Using Bayesian Networks. In: I Ray, S Shenoi. (eds) Advances in Digital Forensics IV. Digital Forensics. IFIP — The International Federation for Information Processing, **2008.** Vol 285. Springer, Boston, MA. DOI: 10.1007/978-0-387-84927-0_22.

[13] R. Corzo, T. Hoffman, P. Weis, J. Franco-Pedroso, D. Ramos, J. Almirall, The Use of LA-ICP-MS Databases to Calculate Likelihood Ratios for the Forensic Analysis of Glass Evidence, Talanta 186 (2018) 655–661, https://doi.org/10.1016/j.talanta.2018.02.027.

[14] A. van Es, W. Wiarda, M. Hordijk, I. Alberink, P. Vergeer, Implementation and Assessment of a Likelihood Ratio Approach for the Evaluation of LA-ICP-MS Evidence in Forensic Glass Analysis, Sci. Justice 57 (3) (2017) 181–192, https://doi.org/10.1016/j.scijus.2017.03.002.

[15] P. Vergeer, A. van Es, A. de Jongh, I. Alberink, R. Stoel, Numerical Likelihood Ratios Outputted by LR Systems Are Often Based on Extrapolation: When to Stop Extrapolating? Sci. Justice 56 (6) (2016) 482–491, https://doi.org/10.1016/j.scijus.2016.06.003.

[16] C. Aitken, G. Zadora, D. Lucy, A Two-Level Model for Evidence Evaluation, J Forensic Sciences 52 (2) (2007) 412–419, https://doi.org/10.1111/j.1556-4029.2006.00358.x.

[17] G. Zadora, Evaluation of Evidence Value of Glass Fragments by Likelihood Ratio and Bayesian Network Approaches, Anal. Chim. Acta 642 (1–2) (2009) 279–290, https://doi.org/10.1016/j.aca.2008.10.005.

[18] Z. Wen, J. Curran, S. Harbison, G. Wevers, Bayesian mixture modeling for glass refractive index measurement, Sci. Justice 61 (4) (2021) 345–355, https://doi.org/10.1016/j.scijus.2021.05.002.

[19] S. Ryland, R. Kopec, P. Somerville, Evidential Value of Automobile Paint, Part 2 - Frequency of Occurrence of Topcoat Colors, J Forensic Sciences 26 (1) (1981) 64–74, https://doi.org/10.1520/JFS11330J.

[20] National Research Council. Strengthening Forensic Science in the United States: A Path Forward. *The National Academies Press,* **2009**. Washington, DC. DOI: 10.17226/12589.

[21] AS Association. American Statistical Association Position on Statistical Statements for Forensic Evidence. **2019**. https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf.

[22] Standard Guide for Interpretation and Reporting in Forensic Comparison of Trace Materials (in development). https://www.nist.gov/system/files/documents/2020/04/02/ChSAC-Mat_Interpretation_Document_MARCH2020_0.pdf.

[23] T. Trejos, S. Koch, A. Mehltretter, Scientific Foundations and Current State of Trace Evidence—a Review, J Forensic Chemistry 18 (2020), 100223, https://doi.org/10.1016/j.forc.2020.100223.

[24] E.S. Berner, G.D. Webster, A.A. Shugerman, J.R. Jackson, J. Algina, A.L. Baker, E. V. Ball, C.G. Cobbs, V.W. Dennis, E.P. Frenkel, L.D. Hudson, E.L. Mancall, C. E. Rackley, O.D. Taunton, Performance of four computer-based diagnostic systems, N. Engl. J. Med. 330 (25) (1994) 1792–1796.

[25] E.S. Berner, R.S. Maisiak, C.G. Cobbs, O.D. Taunton, Effects of a decision support system on physicians' diagnostic performance, J. Am. Med. Inf. Assoc. 6 (5) (1999) 420–427.

[26] RM Neal. MCMC Using Hamiltonian Dynamics. In: S Brooks, A Gelman, G Jones, X Meng (eds.) Handbook of Markov Chain Monte Carlo, **2011**. Chapman and Hall/CRC. DOI: 10.1201/b10905.

[27] Stan Development Team. **2015**. Stan Modeling Language User's Guide and Reference Manual, Version 2.9.0.