Forensic Science International 325 (2021) 110856

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Accuracy and reproducibility of conclusions by forensic bloodstain pattern analysts

R. Austin Hicklin[a,*], Kevin R. Winer[b], Paul E. Kish[c], Connie L. Parks[a], William Chapman[a], Kensley Dunagan[a], Nicole Richetelli[a], Eric G. Epstein[a], Madeline A. Ausdemore[a], Thomas A. Busey[d]

[a] Noblis, Reston, VA, USA
[b] Kansas City Police Department Crime Laboratory, Kansas City, MO, USA
[c] Forensic Consultant, Corning, NY, USA
[d] Indiana University, Bloomington, IN, USA

ARTICLE INFO

Article history:
Received 1 February 2021
Received in revised form 28 April 2021
Accepted 27 May 2021
Available online 3 June 2021

Keywords:
Bloodstain pattern analysis
Forensic science
Forensic identification

ABSTRACT

Although the analysis of bloodstain pattern evidence left at crime scenes relies on the expert opinions of bloodstain pattern analysts, the accuracy and reproducibility of these conclusions have never been rigorously evaluated at a large scale. We investigated conclusions made by 75 practicing bloodstain pattern analysts on 192 bloodstain patterns selected to be broadly representative of operational casework, resulting in 33,005 responses to prompts and 1760 short text responses. Our results show that conclusions were often erroneous and often contradicted other analysts. On samples with known causes, 11.2% of responses were erroneous. The results show limited reproducibility of conclusions: 7.8% of responses contradicted other analysts. The disagreements with respect to the meaning and usage of BPA terminology and classifications suggest a need for improved standards. Both semantic differences and contradictory interpretations contributed to errors and disagreements, which could have serious implications if they occurred in casework.

1. Introduction

Bloodstains are frequently encountered at crime scenes. The forensic discipline of bloodstain pattern analysis (BPA) involves the examination and interpretation of the attributes of bloodstains to determine causal mechanisms [1–4]. In some legal cases, BPA is critical evidence. For example, in the David Camm case [5–7] there were fundamentally contradictory opinions among BPA analysts regarding the classification of the bloodstain pattern that was the key evidence in the case: BPA analysts for the prosecution concluded that the bloodstain pattern on the defendant's clothing was backspatter from a gunshot, but BPA analysts for the defense concluded it was a transfer stain resulting from the defendant assisting his wounded children. BPA differs from many other forensic disciplines (e.g. DNA or latent fingerprint examination) in that it is not focused on source attribution (e.g. who was involved), but rather on addressing what happened at a crime scene [1,2]. For example, BPA conclusions may provide information used in determining whether an incident was suicide or homicide, or whether a claim of self defense is supported (or negated) by the evidence. Although BPA has been admissible as expert testimony for more than 150 years [8], the accuracy and reproducibility of conclusions by BPA analysts have never been rigorously assessed in a large-scale study. A 2009 report from the National Research Council of the National Academies strongly criticized BPA, stating "The uncertainties associated with bloodstain pattern analysis are enormous" and "In general, the opinions of bloodstain pattern analysts are more subjective than scientific" [9]. The National Research Council called for testing of error rates in forensic disciplines, which was echoed in a 2016 report by the President's Council for Science and Technology [10,11]. We conducted this "black box" study [12] to evaluate the accuracy and reproducibility of conclusions made by practicing BPA analysts. Several BPA studies have previously been conducted [13–17], but not with the scale or breadth of the current study.

2. Materials and methods

Bloodstain samples were selected from a pool of 192 samples, which were collected from both controlled collection (123 samples) and operational casework (69 samples). Each participant was

* Correspondence to: Noblis, 2002 E dmund Halley Dr, Reston, VA 20191, USA.
E-mail address: hicklin@noblis.org (R.A. Hicklin).

https://doi.org/10.1016/j.forsciint.2021.110856
0379-0738/

presented with samples via a custom web-based software that presented bloodstain images and recorded test responses. Each participant who completed the study received 150 samples: 30 samples assigned for short text summary conclusions, and 120 samples assigned with multiple classification prompts (mean 4.2 classification prompts per sample) and/or questions (mean 1.2 questions per sample). We obtained a total of 33,005 multiple choice responses and 1760 short text responses from 75 participants, on 192 bloodstain patterns. See SI Appendix 1 for detailed information on Materials and Methods.

2.1. Participation

Participation was limited to practicing BPA analysts. The background survey results (SI Appendix 1.4) illustrate our participants' formal education, training, and experience, representing a diverse group of analysts. The participants were from 14 countries (57% from the U.S.). Analysts generally perform BPA only as one of their responsibilities: nearly half (47%) of the participants perform fewer than five BPA cases per year; 83% have testified in court.

2.2. Prompts

There is no preexisting widely-used BPA conclusion standard that could be adopted for use in this study. The BPA community has developed multiple standards and recommendations for terminology [18–21], detailed in SI Appendix 3.3. In this study, terminology and definitions are based on the standard for terminology developed through the Organization of Scientific Area Committees for Forensic Science (OSAC) and the Academy Standards Board (ASB) [18]. The ASB terminology standard has been recommended by the International Association of Bloodstain Pattern Analysts (IABPA) [22,23] and adopted for use in proficiency tests [24,25], but it is not required for use in casework by BPA analysts.

We developed three complementary approaches to collect participants' assessments of the mechanism(s) that caused each sample: classification prompts, questions, and short text summary conclusions (detailed in SI Appendix 1.2). Classification prompts used terminology explicitly from the ASB standard [18] (summarized in SI Appendix 3.1), to which participants responded using multiple choice options of *definitive, included*, or *excluded.* A response of *included* indicated there was insufficient support either for a *definitive* decision or for an *excluded* decision. The criteria for these decisions have not been clearly defined in the BPA discipline. Therefore, because of this lack of a standardized sufficiency threshold, *included* is considered as indeterminate in analyses, neither correct nor an error. Questions were added over the course of the study as a means of evaluating the reproducibility of statements made by participants in the short text responses: participants were assigned samples for short text responses early in the study so that other participants would subsequently be assigned questions derived from those text responses. Questions often address reconstruction issues encountered in BPA that go beyond pure pattern classification, such as "Was this the result of two cast-off patterns?" or "Was the decedent standing up when [the] bloodletting event occurred?" Questions were assessed using multiple choice options of *yes*, *possible*, or no; possible is considered as indeterminate in analyses. Classifications were extracted from the short text responses by BPA experts on the study team, to enable evaluating responses against known cause, where possible. The experts also evaluated the short text responses for quality and thoroughness.

2.3. Performance measures, known cause, and operational consequence

Reproducibility of responses was assessed on all samples; accuracy was assessed only when the cause of the bloodstain was known. For each prompt (classification or question), a correct ("known cause") response was generally available for controlled collection samples, and was not available for casework samples. Although the overall mechanism for each controlled sample was known, for prompts considered debatable or semantic we left cause as "unknown." For casework samples we cannot claim certain knowledge of the cause of each bloodstain. We asserted known cause for 47% of prompts (81% of prompts for controlled samples and 0% of prompts for casework samples). Note that asserting the cause of a bloodstain is known does not necessarily imply that a given sample has sufficient information to make a definitive attribution of that cause: since BPA has no standardized criteria for determining the types or quantities of characteristics needed to make a given decision, consensus among analysts is the only available means of assessing whether an indeterminate response (*included* or *possible*) is appropriate. In order to limit the effects of prompts that could be seen as minor or semantic, the BPA analysts on the study team also evaluated each prompt to determine whether an error on that prompt would be highly consequential in an actual case. These are labeled "most consequential", and 22.5% of the prompts were labeled as such (see SI Appendix 1.6).

2.4. Bloodstain pattern samples

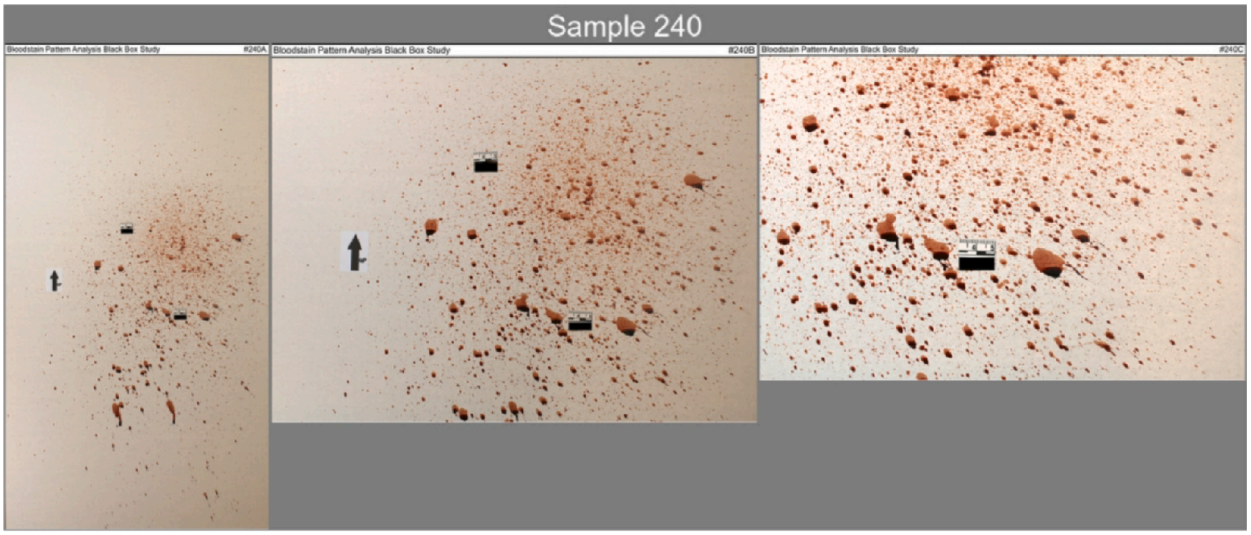
We selected bloodstain patterns in an attempt to be broadly representative of operational casework. In the post-test survey, 72% of the respondents said the difficulty of the samples in the study was similar to casework; 23% said it was harder or much harder than casework. Fig. 1 shows examples of bloodstain patterns from the study that resulted in substantive disagreements among analysts. Note that each sample is shown in multiple images. Sample #240 was caused by *expiration* (see SI Appendix 3 for glossary), created by a team member coughing blood. One of the prompts that participants were provided for this sample was *impact* ("a bloodstain pattern resulting from an object striking liquid blood" [18], such as from a gunshot or club): 16 participants responded *definitive*, 23 *included*, and 4 *excluded* (43 total). Because this sample was known to be caused by *expiration* (Movie S2), the *impact* prompt is false, and those 16 *definitive* responses can be assessed as errors (i.e., contradicting a known cause). The second and third samples in Fig. 1 are from casework, showing examples of prompts that do not have known causes, and therefore are assessed in terms of reproducibility, not error or correctness. For sample #188, participants were provided *impact* as a prompt: 15 responded *definitive*, 11 *included*, and 17 *excluded* (43 total). For Sample #975, in response to the question "Did Pattern A occur after Pattern B?" 8 participants responded *no*, 10 *possible*, and 25 *yes* (43 total). See Data S2 for images and responses for all samples; see [26] for full-resolution imagery for all samples.

Supplementary material related to this article can be found online at doi:10.1016/j.forsciint.2021.110856.

3. Results

3.1. Accuracy

Fig. 2 shows the distribution of responses for classifications and questions. For example, of the classification prompts for which the prompt represented the known cause (*true*), 52.8% of responses were *definitive* (agreeing with known cause), but 15.2% of responses were *excluded* (erroneous exclusions, contradicting known cause). Over all 11,634 classification prompts for which there was a known cause, 11.2% of responses contradicted the known cause and therefore were erroneous (weighted average of erroneous *definitive* and erroneous *excluded* responses). For the 2163 questions for which there was a known cause, 11.0% of responses were erroneous (weighted average


Sample 240
Bloodstain Pattern Analysis Black Box Study #240A
Bloodstain Pattern Analysis Black Box Study #240B
Bloodstain Pattern Analysis Black Box Study #240C


Sample 188
Bloodstain Pattern Analysis Black Box Study #188A
Bloodstain Pattern Analysis Black Box Study #188B
Bloodstain Pattern Analysis Black Box Study #188C
Bloodstain Pattern Analysis Black Box Study #188D
Bloodstain Pattern Analysis Black Box Study #188E
Bloodstain Pattern Analysis Black Box Study #188 Label
A
B
C
D

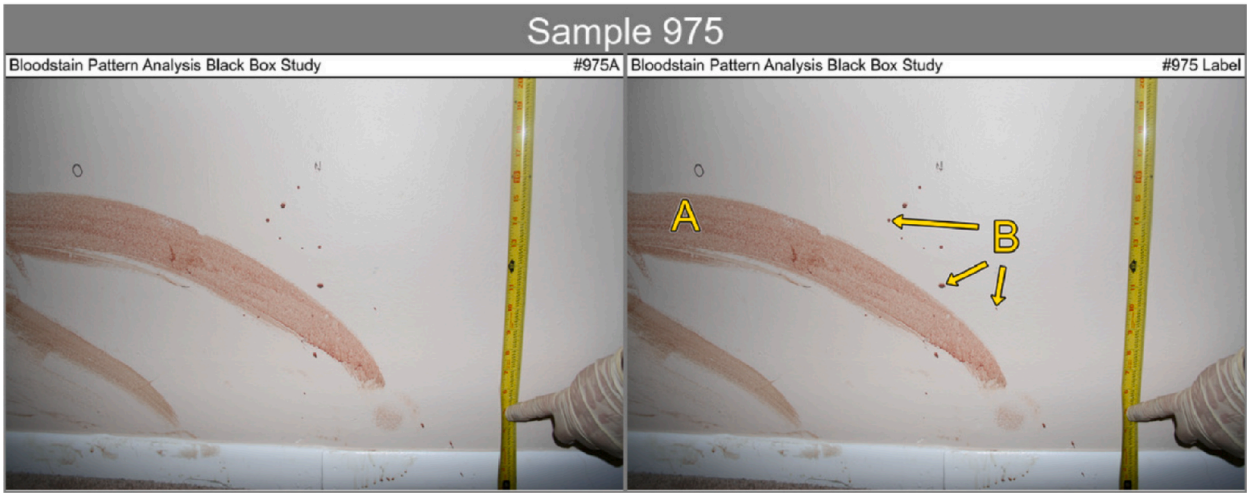

Sample 975
Bloodstain Pattern Analysis Black Box Study #975A
Bloodstain Pattern Analysis Black Box Study #975 Label
A
B

Fig. 1. Examples of bloodstain patterns used in the study. Sample #240 (controlled collection): *expiration* on cardboard, created by a team member coughing blood. Sample #188 (casework): *drip pattern* in basement; victim shot and killed in kitchen; blood *flow* through floor, down a rafter and dripped onto the basement floor. Sample #975 (casework): Three homicide victims within this room; pattern shows characteristics of *spatter* stains altered by a *swipe* resulting in multiple *perimeter stains*. (Descriptions were not provided to participants).

of erroneous *yes* and erroneous *no* responses). These results are similar to [13], which reported that 13% of classifications were erroneous. Responses were indeterminate for 30.1% of classifications and 43.8% of questions. If we consider only determinate responses (*definitive*, *excluded*, *yes*, *no*) on prompts with known cause, 83.0% of responses were correct with respect to known cause ("overall predictive value"). If limited to the most consequential prompts, the error rate for classifications was 9.0% and the error rate for questions was 5.8%; the overall predictive value was 86.6%. (For these and all results, see SI Appendix 2.2 for further explanations and confidence intervals.).

As an alternative to explicit prompts, short text responses provided a means to assess accuracy in a manner comparable to how pattern classifications are reported in operational casework. Of 1760 short text statements, there were 1052 that could be evaluated with respect to known cause (SI Appendix 1.7); of these, 4.8% entirely contradicted known cause and an additional 11.2% partially contradicted known cause (i.e., included both correct and incorrect statements). The BPA experts on the team also evaluated the quality and thoroughness of the short text statements, assessing whether the analysts' observations and conclusions were adequately supported: they determined that 11.3% had errors in reconstruction statements, observations, or unsupported conclusions.

3.2. Consensus

Between five and 49 participants responded to each prompt (mean 33.2 participants per classification prompt; 26.8 per question). The responses for each prompt can be seen as votes in a decision space, as shown in Fig. 3, which plots each prompt in terms of the proportion of each response type. For example, a classification prompt of *spatter* (with a true known cause) with 46 responses (22 *definitive*, 12 *included*, 12 *excluded*) is plotted as a blue point in the classifications chart at (48%, 26%). This provides a means to evaluate results in terms of consensus, which serves multiple purposes: it provides an understanding of the collective behavior of analysts, it serves as a proxy for the known cause when the cause for a bloodstain is unknown, and it indicates the collective judgment of analysts regarding whether the samples contain sufficient information to make a given decision. These results show that consensus was limited, and errors were widely distributed across prompts: only 3% of prompts received unanimous responses (i.e., 100% on the x- or y-axis), 33% of prompts had at least 75% consensus (293 classification prompts and 54 questions), and 81% of prompts had a majority consensus (649 classification prompts and 190 questions). If there were strong consensus, we would see clumps at the top left and bottom right (and potentially bottom left) of Fig. 3, with few points in between. If we assess just the *excluded* vs. *not excluded* decision (combining *definitive* and *included* as *not excluded;* SI Appendix 2.4), 13.9% of responses to classification prompts had unanimous consensus (i.e., were at 0% or 100% on the y-axis), and 63.2% had at least a 75% supermajority consensus (≥75% or ≤25% on the y-axis).

Majority responses were rarely incorrect. For classification prompts, responses with a 95% supermajority never contradicted known cause; 75% supermajority responses contradicted known cause on four prompts (1.0% of 392 classification prompts with known cause); a majority contradicted known cause on an additional five prompts (nine total, 2.3%); and a plurality contradicted known cause on an additional eight prompts (17 total, 4.3%). On questions, a 75% supermajority never contradicted known cause, a majority (or plurality) contradicted known cause on one question (1.0% of 96 questions with known cause). When limited to the most consequential prompts, the majority was always correct.

3.3. Reproducibility

In addition to accuracy (agreement with known cause) and consensus (agreement with a majority of other analysts), analysts can also be assessed in terms of reproducibility: how frequently they reproduce one another's decisions when each response for a given prompt is compared to all other responses for that prompt. One advantage of reproducibility and consensus is that they can be assessed for all samples, including those from operational casework;

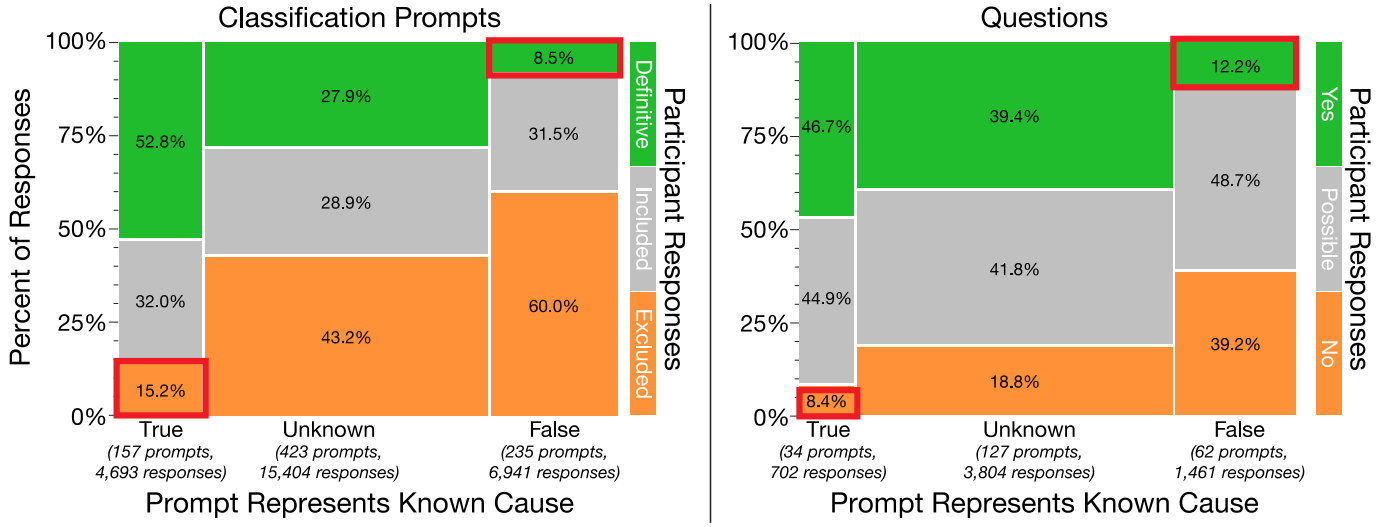


Fig. 2. Mosaic plots of responses. The columns indicate whether the prompts represent known cause of the bloodstain: prompts that are consistent with known cause are labeled *true*; prompts contrary to known cause are labeled *false*; prompts for which the cause is unknown are labeled *unknown*, which we do not assess in terms of accuracy or error. Responses are color-coded, with proportions shown in the y-axis. Erroneous responses (i.e., contradicting known causes) are outlined in red. For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

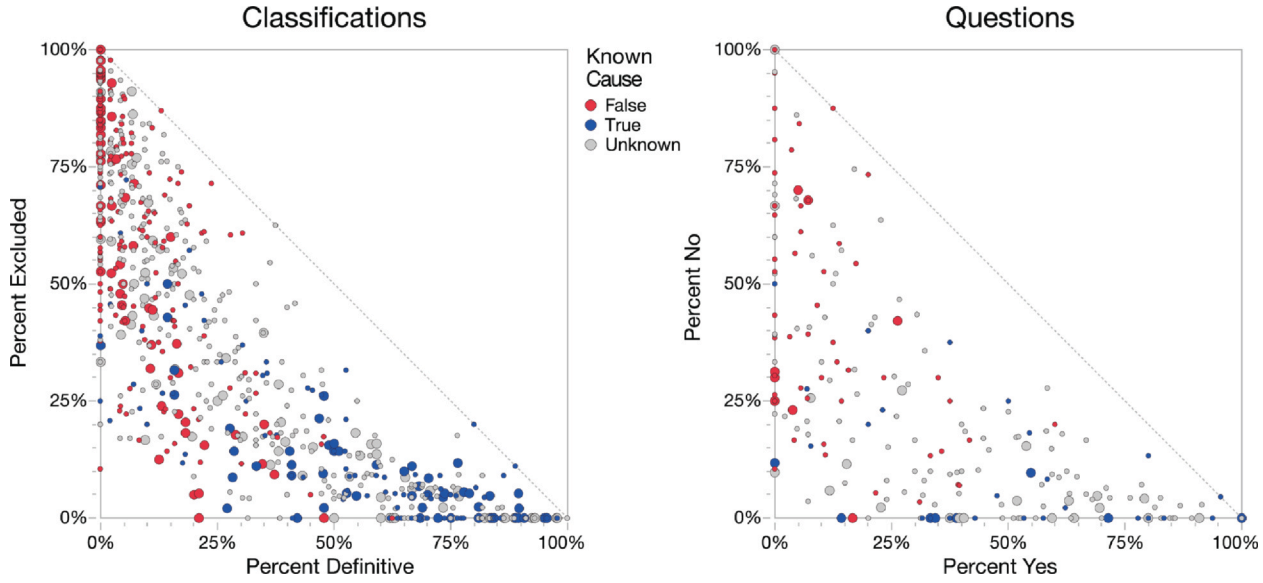


Fig. 3. Consensus on classifications and questions. Each prompt (classification or question) is shown as a circle, positioned based on proportion of responses. Prompts assessed as "most consequential" are shown as larger circles. Responses were unanimous (i.e., superimposed on the 100% corners) on 26 classification prompts and 5 questions. (27,038 responses on 815 classification prompts; 5967 responses on 223 questions).

accuracy, by contrast, requires known cause. Both reproducibility and consensus reflect on the overall reliability of BPA: imperfect reproducibility or consensus limits precision and places an upper bound on accuracy.

In some casework situations, two or more BPA analysts render classification opinions on the same bloodstain evidence, such as during technical review or in court. For this reason, we evaluated the reproducibility (inter-analyst variability) of responses (SI Appendix 2.5). Fig. 4 summarizes the reproducibility of responses. For example, for every participant who responded *excluded* to a classification prompt, 65.3% of other participants also responded *excluded* to that prompt (agreement rate), 25.3% responded *included*, and 9.4% responded *definitive* (contradiction rate). Across all classification prompts and questions, the overall agreement rate (OAR; the proportion of other participants who had an identical response to a given prompt) was 54.6%; the overall contradiction rate (OCR; the proportion of other participants who had a diametrically opposed response) was 7.8%. If limited to the most consequential prompts, OAR was 56.3% and OCR was 6.2%. Contradictions were distributed broadly across prompts: 549 of the 815 classification prompts and 146 of the 223 questions resulted in contradictions.

When reproducibility of responses is conditioned on known cause (SI Appendix 2.5; Fig. S10), erroneous responses were often reproduced: 17.7% of erroneous *definitives* were reproduced, as were 34.1% of erroneous *excluded* responses, 24.2% of erroneous *yes* responses, and 22.5% of erroneous *no* responses. These results suggest that if two BPA analysts both analyze a pattern (such as occurs operationally during technical review) they cannot always be expected to agree, and if they do agree they may both be wrong.

3.4. Semantic issues

Many of the disagreements—and some of the errors—may be attributed to semantic differences rather than contradictory interpretations. Such semantic issues include inadequate delineation between some pattern types (such as between *splash* and *drip*

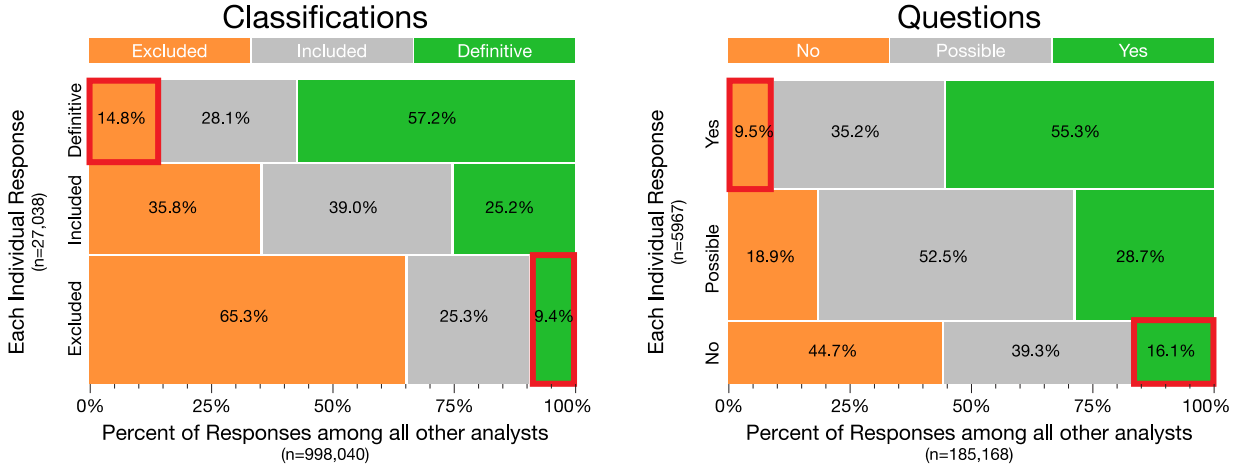


Fig. 4. Reproducibility of responses. Each percentage represents the probability of a second analyst providing a given response, conditioned on the first analyst's response. Contradictory responses are outlined in red. We calculate these probabilities by comparing each response for a given prompt to all other responses for that prompt. Counts are of all pair-wise combinations of responses from different analysts on each prompt. For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

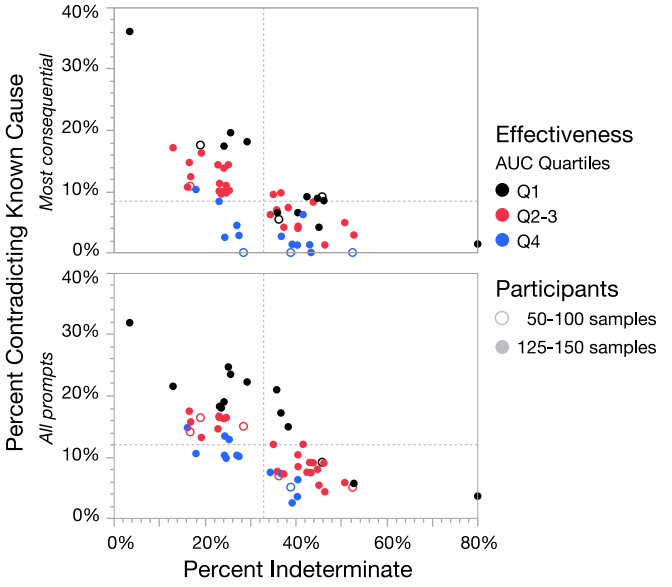


Fig. 5. Comparison of participants in terms of accuracy (y-axis), decisiveness (x-axis), and effectiveness (color: see text). Dashed lines represent means. Top panel (most consequential prompts): rates calculated on 3968 responses to 133 prompts (mean 71 responses per participant); bottom panel (all prompts): rates calculated on 11,810 responses to 488 prompts (mean 211 responses per participant). (N = 56 participants; omits the 19 participants who completed fewer than 50 of the 150 assigned samples; the 12 participants with fewer than 100 samples are shown as open circles.). For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

patterns, or whether *cessation* is a subcategory of *cast-off*), and some definitions are ambiguous (such as the minimum quantity necessary to classify a pattern as a *pool* or *saturation stain*, or the types of external factors that should be considered *alteration*). Even for the simplest bloodstain patterns there is notable disagreement: when provided a classification prompt of *spatter* on samples consisting of a single drop of blood on a non-porous horizontal surface (a *drip stain*), out of 105 responses 42 were *definitive* and 46 were *excluded*. In the responses to classification prompts, errors were disproportionately associated with some pattern types, which may in part be explainable by semantic issues. Participants erroneously excluded more than 35% of *splash*, *projected*, and *satellite* patterns (SI Appendix 2.9; Table S25). Participants often incorrectly concluded that *splash* patterns were *drip patterns* (34% erroneous definitive rate), and often incorrectly concluded that *satellite stains* were *impact* patterns (30%). These results indicate that there was not general agreement among participants on the delineation between *splash* and *drip patterns*; the high rates of exclusion on *projected* and *satellite* patterns may be attributable to a combination of semantic issues and differences in interpretation. Such semantic issues sometimes limited our ability to define known cause for a specific classification or question, even when video of the event was available. In the post-study survey (SI Appendix 1.4b) participants indicated that "several examples stretched the semantic interpretation of definitions," in particular regarding *projected* and *impact* patterns; this was bolstered by analysis of the short text responses. We report results for both all prompts and the most consequential results in order to limit the effects of semantic issues. Although some semantic disagreements would presumably be unlikely to have significant consequences in actual casework, their prevalence obscures the extent of serious disagreements. This lack of agreement on the meaning and usage of BPA terminology and classifications illustrates the need for improved standards.

3.5. Comparing participants

Fig. 5 compares the participants in terms of accuracy. When calculating rates for each participant, we limit analyses to the 56 participants who completed at least 50 samples. Each chart shows three interrelated dimensions: accuracy (measured as the proportion of erroneous responses, y-axis), decisiveness (measured as the proportion of indeterminate responses, x-axis), and effectiveness (color). We assess the effectiveness of participants in terms of the receiver operating characteristic area under the curve (AUC), which uses the participants' responses as predictors of known cause: two decision thresholds are modeled by considering indeterminate responses as positive or negative (SI Appendix 2.7). High AUC values result not only from a high number of positive (*definitive or yes*) responses on true prompts and negative (*excluded* or *no*) responses on false prompts, but also from a high number of non-negative (*definitive, included*, *yes*, or *possible*) responses on true prompts and non-positive (*excluded*, *included*, *no*, or *possible*) responses on false prompts. The most effective analysts are shown in blue closest to the bottom-left corner of each graph in Fig. 5. Variation within a quartile can be seen as differences in risk aversion among participants, shown as an inverse relationship between accuracy and decisiveness within a given color band: participants who tend to give more determinate responses (shifted left on the x-axis) are more likely to make mistakes (shifted up on the y-axis), whereas others in the same color band make fewer mistakes at the expense of being less definitive. See SI Appendix 2.6 and Fig. S11 for further details and analogous results from performance metrics assessing consensus and reproducibility.

In general, the participants exhibited a continuum of performance: errors were widely distributed among participants, and all participants who completed more than 50 samples made multiple errors. However, two participants showed notably anomalous results. One participant (top left of both charts in Fig. 5) contradicted known cause on 36% of responses on highly consequential prompts (32% on all prompts), but was indeterminate on only 4% of responses—that participant was responsible for 5.7% of all errors in the study. Another participant (bottom right of both charts in Fig. 5) contradicted known cause on only 1% of responses on highly consequential prompts (4% on all prompts), but was indeterminate on 80% of responses. Both of the anomalous participants currently conduct bloodstain pattern analysis as part of their employment, work in a laboratory environment, conduct fewer than 5 BPA cases per year, have testified in court as BPA experts, have at least a master's degree, did not complete a formal program of BPA instruction/supervision, and are not certified by the International Association for Identification (IAI); one is from the US. A total of five participants share these background attributes—note the others who share these attributes did not show problematic performance. Human subjects research protections do not permit revealing further information that could be used to identify these individuals.

In order to further characterize the performance of the participants, we developed a novel procedure for detecting and reporting any associations between participants' performance and their background attributes (SI Appendix 2.8). Performance was assessed for 54 of the participants (omitting the two outliers, and the participants who completed fewer than 50 samples each) with respect to 25 background attributes using two complementary approaches: variable importance analysis and attribute-specific significance testing. Variable importance analysis (VIA) was conducted by considering all attributes simultaneously and coupling linear regression and random forest analysis to yield importance scores. In addition, significance testing was conducted for each attribute individually using the Kruskal-Wallis test to yield p-values and q-statistics. Using

these importance scores, p-values, and q-statistics, we set association thresholds and a reporting criteria hierarchy to determine which (if any) of these background attributes exhibited sufficient support to indicate an association with performance. For the majority of background attributes (including length of training, educational degree, certification, or length of experience), we found no support for associations with performance. The exceptions were country of practice and extent of advanced workshop training, for which we found limited support for an association with performance (SI Appendix 2.8).

4. Discussion

Our results show that conclusions by BPA analysts were often erroneous and often contradicted other analysts. Such errors could have serious implications if they occurred in casework, as would conflicting conclusions among BPA analysts if those resulted in conflicting testimony in court. Many of the disagreements among BPA analysts—and some of the errors—may be attributable to semantic differences. The results show that there is often a lack of agreement on the meaning and usage of BPA terminology and classifications, suggesting a need for improved standards. Unless there is general consensus on what criteria are necessary and sufficient to make a given decision we cannot expect high rates of reproducibility among analysts.

The results here are intended to provide estimates for use in decision making, improving procedures and training, and future research. These results should not be taken to be precise measures of operational error rates: the error rates reported here describe the proportion of erroneous results for this particular set of samples with these particular participants; these rates cannot and should not be assumed to apply to all BPA analysts across all casework. The discipline of bloodstain pattern analysis is not solely defined by pattern classification but rather it includes multiple other aspects that were not evaluated within this study. The study differed from operational casework in that analysts were asked to provide responses based solely on photographs, analysts were not provided case-relevant facts that may have aided in making conclusions, and the means of reporting conclusions were different from the manner in which BPA analysts typically reach conclusions. These results do not account for operational quality assurance measures, such as technical review or verification.

In conducting this study and performing analyses, the authors developed a number of detailed recommendations, which we suggest may be considered by the BPA community in general, by standards bodies, and by laboratory management. These recommendations fall into the following broad categories: methodology, terminology, implications for casework, and lessons learned. Please see SI Appendix 1 for detailed recommendations.

5. Significance statement

The analysis of bloodstain pattern evidence left at crime scenes relies on the expert opinions of bloodstain pattern analysts. This is the first large scale rigorous evaluation of the accuracy and reproducibility of practicing bloodstain pattern analysts' conclusions. Our results show that conclusions were often erroneous and often contradicted other analysts. The disagreements with respect to the meaning and usage of BPA terminology and classifications suggest a need for improved standards. Both semantic differences and contradictory interpretations contributed to errors and disagreements, which could have serious implications if they occurred in casework.

Funding

This project was supported by Award No. 2018-DU-BX-0214, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice.

CRediT authorship contribution statement

R. Austin Hicklin: Conceptualization, Investigation, Resources, Methodology, Formal analysis, Data curation, Validation, Writing - original draft, Writing - review & editing, Visualization, Supervision. **Kevin R. Winer:** Conceptualization, Investigation, Resources, Writing - review & editing. **Paul E. Kish:** Conceptualization, Investigation, Resources, Writing - review & editing. **Connie L. Parks:** Investigation, Resources, Data curation, Validation, Writing - original draft, Writing - review & editing. **William Chapman:** Investigation, Resources, Writing - review & editing, Supervision. **Kensley Dunagan:** Data curation, Writing - review & editing. **Nicole Richetelli:** Methodology, Formal analysis, Validation, Writing - original draft, Writing - review & editing, Visualization. **Eric G. Epstein:** Methodology, Formal analysis, Validation, Writing - original draft, Writing - review & editing, Visualization. **Madeline A. Ausdemore:** Methodology, Formal analysis, Writing - review & editing. **Thomas A. Busey:** Methodology, Formal analysis, Validation, Writing - review & editing, Visualization.

Data availability

All data is available in the main text, supplementary materials, or archived at OSF (24); because participants were assured of anonymity, results by participant are summarize or deidentified to prevent reidentification.

Declaration of Competing Interest

Authors declare no competing interests.

Acknowledgements

We thank the BPA analysts who participated in this study, as well as Kyle Dalrymple for software development, and Niki Osborne, Nikki Blackwell, and LeeAnn Singley for their insights during study design. The opinions, findings, and conclusions or recommendations expressed in this program are those of the authors and do not necessarily reflect those of the Department of Justice.

Supporting Information

Supplementary data and information associated with this article can be found in the online version at doi:10.1016/j.forsciint.2021.110856. Other Supplementary Materials for this manuscript include the following: Detailed appendices; Participant Instructions; Proofsheets (low-resolution summary images) and responses for all 192 samples; Response Data (sample descriptions, classification prompts and questions with responses, short text responses, deidentified survey responses, summary results by participant); Example videos of creation of controlled collection samples.

References

[1] T. Bevel, R.M. Gardner, Bloodstain Pattern Analysis with an Introduction to Crime Scene Reconstruction, CRC Press, Boca Raton, 2008.
[2] S.H. James, P.E. Kish, T.P. Sutton, Principles of Bloodstain Pattern Analysis: Theory and Practice, CRC Press, Boca Raton, 2005.

[3] H. MacDonell, Bloodstain Pattern Interpretation, Laboratory of Forensic Science, Corning, 1982.
[4] H. Macdonell, Bloodstain pattern interpretation, Wiley Encyclopedia of Forensic Science, John Wiley & Sons, Ltd, Chichester, UK, 2009, pp. 359–395, https://doi.org/10.1002/9780470061589.fsa066
[5] David R. Camm v. State of Indiana (812 N.E.2d 1127, Court of Appeals of Indiana), 2004.
[6] David R. Cammv State of Indiana (908 N.E.2d 215, Supreme Court of Indiana), 2009.
[7] David R. Cammv. State of Indiana (957 N.E.2d 205, Court of Appeals of Indiana), 2011.
[8] State of Maine v. Knight (43 Me. 11, Supreme Judicial Court of Maine), 1857.
[9] National Research Council, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington, D.C., 2009.
[10] President's Council of Advisors on Science and Technology, Report to the President. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Executive Office of the President, Washington, D.C., 2016.
[11] 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.
[12] J.L. Mnookin, Of black boxes, instruments, and experts: testing the validity of forensic science, Episteme 5 (2008) 343–358, https://doi.org/10.3366/E1742360008000440
[13] M.C. Taylor, T.L. Laber, P.E. Kish, G. Owens, N.K.P. Osborne, The reliability of pattern classification in bloodstain pattern analysis, part 1: bloodstain patterns on rigid non-absorbent surfaces, J. Forensic Sci. 61 (2016) 922–927, https://doi.org/10.1111/1556-4029.13091
[14] M.C. Taylor, T.L. Laber, P.E. Kish, G. Owens, N.K.P. Osborne, M.C. Taylor, T.L. Laber, P.E. Kish, G. Owens, N.K.P. Osborne, M.C. Taylor, T.L. Laber, P.E. Kish, G. Owens, N.K.P. Osborne, The reliability of pattern classification in bloodstain pattern analysis – part 2: bloodstain patterns on fabric surfaces, J. Forensic Sci. 61 (2016) 1461–1466, https://doi.org/10.1111/1556-4029.13191
[15] T. Laber, P. Kish, M. Taylor, G. Owens, N. Osborne, J. Curran, Reliability Assessment of Current Methods in Bloodstain Pattern Analysis (NIJ/NCJRS Report 247180), 2014. https://www.ncjrs.gov/pdffiles1/nij/grants/247180.pdf.
[16] B. Meneses, B. Gestring, A Preliminary Study: Evaluating Error Rate Associated with Bloodstain Pattern Analysis, Cedar Crest College, 2009.
[17] S.K.Y. Yuen, M.C. Taylor, G. Owens, D.A. Elliot, The reliability of swipe/wipe classification and directionality determination methods in bloodstain pattern analysis, J. Forensic Sci. 62 (2017) 1037–1042, https://doi.org/10.1111/1556-4029.13298
[18] AAFS Standards Board, ASB Technical Report 033: Terms and Definitions in Bloodstain Pattern Analysis, 2017.
[19] International Association of Bloodstain Pattern Analysts (IABPA), Suggested IABPA Terminology List, Int. Assoc. Bloodstain Pattern Anal. Newsl. 12 (1996) 15–17. https://static1.squarespace.com/static/543841fce4b0299b22e1956a/t/54be8e3ee4b0121e42a6c0d7/1421774398257/IABPA+Terminology+1996.pdf.
[20] International Association of Bloodstain Pattern Analysts (IABPA), Suggested IABPA Terminology List, 2004. https://static1.squarespace.com/static/543841fce4b0299b22e1956a/t/54be8822e4b06fad9ba9d473/1421772834653/BPATerminology.pdf.
[21] Scientific Working Group on Bloodstain Pattern Analysis (SWGSTAIN), Scientific Working Group on Bloodstain Pattern Analysis: Recommended Terminology, Forensic Sci. Commun. 11 (2009). https://archives.fbi.gov/archives/about-us/lab/forensic-science-communications/fsc/april2009/standards/2009_04_standards01.htm.
[22] International Association of Bloodstain Pattern Analysts, Published Standards Related to BPA, J. Bloodstain Pattern Anal. 35 (2020) 30. ⟨https://iabpa.org/docs/JBPA_Vol_35_No_3_September_2020_v1.pdf.
[23] International Association of Bloodstain Pattern Analysts, Standards and Technical Reports Published by the Academy Standards Board (ASB/ANSI - BPA Consensus Body), (2021). https://www.iabpa.org/asb_ansi_-_bpa_consensus_body.php.
[24] Forensic Testing Program, Bloodstain Pattern Analysis Test No. 18-5601/2/5 Summary Report, 2018.
[25] Forensic Testing Program, Bloodstain Pattern Analysis Test No. 20-5601/5 Summary Report, 2020. https://cts-forensics.com/reports/20–5601.5_Web.pdf. (Accessed 10 December 2020).
[26] Noblis, Bloodstain Pattern Analysis Black Box Study Dataset (V1, Dec 2020), 2020. https://doi.org/10.17605/osf.io/2ckhw.