NIST Grant/Contractor Report
NIST GCR 25-068

Implications of the Bullet Black Box Study

Report of the Bullet Black Box Working Group

This publication is available free of charge from:
https://doi.org/10.6028/NIST.GCR.25-068

NIST Grant/Contractor Report
NIST GCR 25-068

Implications of the Bullet Black Box Study

Report of the Bullet Black Box Working Group

The Bullet Black Box Working Group

This publication is available free of charge from:
https://doi.org/10.6028/NIST.GCR.25-068

June 2025

U.S. Department of Commerce
Howard Lutnick, Secretary

National Institute of Standards and Technology
Craig Burkhardt, Acting NIST Director and Acting Under Secretary of Commerce for Standards and Technology

NIST GCR 25-068
June 2025

Disclaimers

This publication was produced as part of Cooperative Agreement 70NANB18H232 with the National Institute of Standards and Technology. The contents of this publication do not necessarily reflect the views or policies of the National Institute of Standards and Technology or the US Government.

Certain commercial equipment and materials are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

NIST Technical Series Policies

How to Cite this NIST Technical Series Publication

Bullet Black Box Working Group (2025) Implications of the Bullet Black Box Study. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Grant/Contractor Report (GCR) 25-068.
https://doi.org/10.6028/NIST.GCR.25-068

Contact Information

Dr. Henry J. Swofford, henry.swofford@nist.gov

Abstract

The Bullet Black Box Working Group (BulletBB-WG) was convened to review the results of the NIST-Noblis Bullet Black Box Study, and assess the implications of that study on casework. This report includes the recommendations of the BulletBB-WG on how to address those implications with the goal of improving the practice of the comparison of bullets by forensic firearms examiners. BulletBB-WG's recommendations focus on standard operating procedures in casework, quality assurance, training, proficiency and competency testing, standardization, testimony, admissibility, and use/communication of the Bullet Black Box Study results.

Keywords

Black Box Study; Criteria for Identification; Error; Forensic Firearms Examination; Forensic Science; Quality Assurance; Reproducibility; Task-Relevant Information; Testimony; Training.

Table of Contents

Acknowledgments

We thank the firearms examiners who participated in the Bullet Black Box Study, as well as the coauthors of that study who were not involved in the Bullet Black Box Working Group (Connie Parks, Kensley Dunagan, Brandi Emerick, and William Chapman). The opinions and recommendations expressed in this document are those of the members of the Bullet Black Box Working Group and do not necessarily reflect those of the National Institute of Standards and Technology (NIST) or the United States Government. In no case does identification of any commercial product imply recommendation or endorsement, nor does it imply that the products identified are necessarily the best available for the purpose.

This project was supported by NIST Cooperative Agreement 70NANB18H232.

1. The Bullet Black Box Study and Working Group

The Bullet Black Box Study (hereafter "BulletBB") was conducted by Noblis and NIST to assess the accuracy and reproducibility of bullet comparison decisions by forensic firearm and toolmark examiners (FFTEs). This study particularly focused on the effects on examiners' decision rates when comparing damaged bullets, questioned bullets (in the absence of knowns), or different types of bullets (jacketed hollow-point vs. full metal jacket). In BulletBB, 49 practicing FFTEs conducted 3,156 comparisons of 1,240 distinct comparison sets, including both known-questioned (KQ) and questioned-questioned (QQ) comparisons.

The results of BulletBB have been published in the following journal article, with extensive technical details included in the appendices:

- R. A. Hicklin, C. L. Parks, K. M. Dunagan, B. L. Emerick, N. Richetelli, W. J. Chapman, M. Taylor, R. M. Thompson. Accuracy and Reproducibility of Bullet Comparison Decisions by Forensic Examiners. *Forensic Science International*, 365(112287), 2024. https://doi.org/10.1016/j.forsciint.2024.112287
- Accuracy of Bullet Comparison Decisions by Forensic Examiners Supplemental Information — Appendices A through N. (https://ars.els-cdn.com/content/image/1-s2.0-S0379073824003694-mmc1.pdf)

In the BulletBB journal article, results were reported as they were observed, without consideration of potential implications or recommendations for the firearms community. This document was written to fill that need: here, a group of firearms experts collaborated to build on BulletBB by presenting their consensus views on the implications of BulletBB results, and offering recommendations to the firearms community on how to address those implications.

In order for the results of BulletBB to be made as useful and actionable as possible, Noblis and NIST formed a working group of firearms experts (the Bullet Black Box Working Group, hereafter "BulletBB-WG"), with the stated purpose of (i) reviewing the then-draft BulletBB manuscript, (ii) reviewing comparison sets of bullets from the study that resulted in the highest rates of errors, and (iii) developing a report of findings and recommendations regarding the implications of BulletBB. BulletBB-WG members include experienced firearms examiners and researchers who are prominent in the community, and who have demonstrated insight and effective collaboration in standards development organizations and professional forensic science organizations.

The BulletBB-WG met in multiple virtual meetings and one multi-day in-person meeting.

1.1. The Bullet Black Box Working Group members

The following lists the BulletBB-WG members, with their current affiliations. The opinions presented in this report do not represent the official positions of institutions with which members are affiliated. Practicing examiners are noted as "FFTE"; authors of BulletBB are noted as "BulletBB."

- James Carroll (FFTE) — Crime Laboratory Director, Los Angeles Sheriff's Department; Organization of Scientific Area Committees for Forensic Science (OSAC) Firearms & Toolmarks Subcommittee
- Eric Collins (FFTE) — NIBIN Section Chief/Supervisory Firearm and Toolmark Examiner, Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF); OSAC Firearms & Toolmarks Subcommittee
- R. Austin Hicklin (BulletBB) — Director of the Forensic Science Group, Noblis [Editorial committee]

- Erica Lawton (FFTE) — Firearm/Tool Mark Section Chief, Alabama Department of Forensic Sciences; President of the Association of Firearm and Tool Mark Examiners (AFTE); Vice Chair, OSAC Firearms & Toolmarks Subcommittee
- Jodi Marsanopoli (FFTE) — National Firearms Examiner Academy Program Manager, National Laboratory Center, ATF
- Nancy McCombs (FFTE) — FATE Forensic Instruction; retired California Department of Justice
- Ronald Nichols (FFTE) — President, Nichols Forensic Science Consulting
- Nicole Richetelli (BulletBB) — Forensic Statistical Analyst and Researcher, Noblis [Editorial committee]
- Steve Scott (FFTE) — Retired Tennessee Bureau of Investigation; retired FBI Laboratory
- Andrew Smith (FFTE) — Supervisor, Firearms Unit, San Francisco Police Department; Chair, OSAC Forensic Science Standards Board
- Erich Smith (FFTE) — Physical Scientist, FBI Laboratory
- Travis Spinder (FFTE) — Administrator, Montana Department of Justice
- Henry Swofford — Lead Scientist, Forensic Science Research Program, Special Programs Office, NIST; Chair, OSAC Physics/Pattern Interpretation Scientific Area Committee
- Melissa Taylor (BulletBB) — Senior Forensic Science Research Manager, Special Programs Office, NIST
- Robert Thompson (FFTE, BulletBB) — Associate NIST Researcher; AAFS Academy Standards Board Firearms and Toolmarks Consensus Body; retired Senior Forensic Science Research Manager, Special Programs Office, NIST
- Todd Weller (FFTE) — Weller Forensics, LLC; Chair, OSAC Firearms & Toolmarks Subcommittee
- Xiaoyu (Alan) Zheng — Mechanical Engineer, Sensor Science Division, NIST; OSAC Firearms & Toolmarks Subcommittee

2. Implications of the Black Box Study

This document assumes familiarity with the BulletBB journal article [1] — please refer to the report for details and explanations. Here we summarize the implications of key findings of BulletBB, and detail the BulletBB-WG's recommendations with respect to standard operating procedures (SOPs) in casework, quality assurance (QA), training, proficiency/competency testing, standardization, testimony, admissibility, and use/communication of the BulletBB results.

These implications are grouped here into the following categories (shown here by section number):

- Section 2.1 Overall study results
- Section 2.2 Variability by make and model of firearm and ammunition
- Section 2.3 Variability by bullet quality
- Section 2.4 Known-questioned (KQ) comparisons vs. questioned-questioned (QQ) comparisons
- Section 2.5 Comparison of bullets from different firearms sharing class characteristics
- Section 2.6 Reproducibility of comparison decisions
- Section 2.7 Variation in decision rates by examiner
- Section 2.8 Variability of NoValue (not suitable) decisions
- Section 2.9 Comparison of different ammunition types
- Section 2.10 Caveats and limitations

Each of the following sections includes pointers to the relevant recommendations made by the BulletBB-WG; the recommendations are detailed in Section 3.

2.1. Overall study results

BulletBB demonstrated that ***rates and distributions of decisions are associated with a variety of factors***, particularly model of firearm, type of ammunition, and quality of bullets. Therefore, we urge that any discussion of the decision rates from this study must specify these factors, and must also note that decision rates can vary significantly by examiner. ***BulletBB was not conducted to derive a single overall error rate that would generalize to all firearms examiners and all bullet comparisons***: the wide variation among all these factors leads us to caution that citing overall rates without differentiating among these factors is likely to be misleading.

In considering the rates reported in BulletBB, it is important to understand that decision rates can be measured in multiple ways, particularly for multi-level conclusion scales.[a] As recommended by OSAC [2] and similar to previous studies [3–5], decision rates can be reported three ways: 1) PRES, which includes all presentations; 2) COMP, which includes all comparisons, excluding NoValue trials, and 3) CALLS, which includes only comparisons with conclusive calls (ID and Excl), excluding trials with NoValue or inconclusive responses (LeanID, Insuff, and LeanExcl). Any references to decision rates should specify the method used (PRES, COMP, or CALLS). Because the variability of NoValue and inconclusive rates is central to the results of the study, the decision rates in the main paper were computed based upon all presentations (PRES), with the COMP and CALLS rates reported in the appendices. We want to emphasize that the proportions of inconclusive responses varied significantly by model of firearm, type of ammunition, and bullet quality, and therefore citing decision rates without consideration of inconclusive rates is likely to be misleading.

[a] Note on decision/conclusion scale: the comparison determinations in BulletBB were taken from the then-current Organization of Scientific Areas Committees (OSAC) draft standard [15], which was a modification of the Association of Firearm and Tool Mark Examiners (AFTE) Range of Conclusions [18]. The quality of each Q was assessed as [OfValue, NoValue]; source conclusions were not assessed for NoValue bullets. Source conclusions were assessed as [Identification (ID), Inconclusive, or Exclusion], with three subcategories of inconclusive [Insufficient Support for Identification (LeanID), Insufficient Support for Either Exclusion or Identification (Insuff), Insufficient Support for Exclusion (LeanExcl)].

After considering these limitations, it is reasonable to note that the results of BulletBB are consistent with the "FBI/Ames2" firearms study [6] and the Best/Gardner study [7]. BulletBB (Appendix H in the journal article) provides a detailed comparison to the FBI/Ames2 black box study, with discussion of limiting the results to the conditions corresponding between the two studies. We note the BulletBB general trends align with other studies in that errors were disproportionately associated with a small number of participants.

The following recommendations are derived from the implications related to overall study results (summarized here; see Section 3 for detailed recommendations):

- Recommendation #1 cautions against citing decision rates from this study without differentiating by firearm manufacturer/model, ammunition type, or bullet quality.
- Recommendation #2 recommends that citations of decision rates from this study should note that decision rates can vary significantly by examiner, and cannot be assumed to apply to any given individual examiner.

2.2. Variability by make and model of firearm and ammunition

BulletBB demonstrated that ***decision rates varied significantly by firearm make/model and by ammunition type***.[b] These differences were particularly notable for the firearm model in the study that contained polygonal rifling: bullets fired from Glock 17 pistols with polygonal rifling resulted in far more inconclusive or unsuitable responses than bullets fired from firearms with conventional rifling (as expected based on previous studies [8]). However, even among makes and models of conventionally rifled firearms there was notable variability in the distribution of decisions, as well as variability when ammunition with different jacket types were fired from the same model of firearm. Overall, the variability in decisions among firearm:ammunition combinations was associated with the comparability of the known bullets (e.g., how consistently class and individual characteristics were reproduced[c] among known bullets from a given firearm). Firearm:ammunition combinations with lower comparability had higher proportions of inconclusive decisions, and lower proportions of mated trials resulting in IDs.

Because the distribution of decisions varied notably among firearm:ammunition combinations, we caution (not just for this study, but in general) against assuming that examiner performance results for a given firearm:ammunition combination are necessarily applicable to a different firearm:ammunition combination. As discussed in Section 2.1, we caution against using overall decision rates (i.e., rates that do not differentiate by firearm manufacturer/model, ammunition type, or bullet quality) in any context implying that those rates would apply to casework in general. Note that for a study (such as this) that includes a variety of different firearms or ammunition, overall decision rates are problematic, because such decision rates are a function of the proportions of each firearm and ammunition, and therefore different proportions of firearms or ammunition could be expected to result in different distributions of decisions. If this study had included a larger proportion of a high-comparability firearm:ammunition combination, the overall rates for (true) IDs on mated comparisons and (true) exclusions on nonmated comparisons would be expected to be higher, and rates of inconclusives would be expected to be lower — the opposite would have been true if this study included a larger proportion of a low-comparability firearm:ammunition combination. For example, if this study had been limited to Winchester White Box

[b] Note that BulletBB results by different types of ammunition distinguished only between FMJ and JHP, and did not distinguish between different jacket compositions or manufacturers.

[c] Note that the term "reproducibility" can be used in two senses: "reproducibility of characteristics" refers to how consistently class and individual characteristics are found among known bullets for a given firearm:ammunition combination; this is distinct from "reproducibility of decisions," which refers to how frequently different examiners reach the same decisions for each response for a given comparison set. In this document we use "comparability of known bullets" as synonymous with "reproducibility of characteristics," so that "reproducibility" is reserved to refer to "reproducibility of decisions."

FMJ bullets fired from Ruger P95D pistols, the overall rates for IDs on mated comparisons and exclusions on nonmated comparisons would have likely been much higher (and inconclusives much lower) than if this study had been limited to Glock 17s (or SIG Sauer P229s) using the same ammunition.

The following recommendations are derived from the implications related to variability by make and model of firearm and ammunition (summarized here; see Section 3 for detailed recommendations):

- Recommendation #1 cautions against citing decision rates from this study without differentiating by firearm manufacturer/model or ammunition type.
- Recommendation #3 recommends that sample sets used during initial training, ongoing education, and competency/proficiency tests should include varied comparisons of firearm:ammunition combinations.

Limitations and caveats regarding decision rates by firearm and ammunition:

The study showed trends, not specific rates: the study was large enough to show significant differences by firearm model in conjunction with ammunition type/model, but the study was not large enough to measure decision rates for any specific firearm:ammunition combination with precision.

All of the polygonally-rifled firearms used in the study were the same model (Glock 17)[d], and therefore variability among different models of polygonally-rifled firearms was not assessed.

The study generally had two types of ammunition per firearm model (JHP and FMJ), but did not have enough models of ammunition for each type to determine whether the variability was due to ammunition type, jacket composition, or other aspects of the ammunition (e.g. muzzle velocity or powder charge).

Each of these topics would be appropriate for further research.

2.3. Variability by bullet quality

In BulletBB, "quality" refers to the extent of rifling present for a given bullet (also known as "completeness"). This is distinct from the extent to which features are reproduced among bullets fired from a given firearm:ammunition combination (which BulletBB refers to as "comparability").[e] BulletBB was designed to accommodate a range of bullet quality (from complete bullets to fragments), in light of the fact that the quality of bullets recovered from crime scenes varies significantly and is generally not similar to the high-quality bullets collected as known exemplars. The study results showed that ***lower-quality bullets (i.e., bullets with more damage) tend to have higher rates of inconclusive responses***. The BulletBB-WG notes this is an expected result: as the amount of information (features) available to examiners decreases, the rate of inconclusive decisions increases.

The overall decision rates reported in the study are a function of the distribution of the quality of the questioned bullets. This study takes care to report results with respect to the effects of bullet quality.

The following recommendations are derived from the implications related to variability by bullet quality (summarized here; see Section 3 for detailed recommendations):

[d] The study used three distinct Glock 17 pistols, all Gen1-Gen2, prior to Glock changing to a different rifling method.
[e] See BulletBB Appendix B3.5 for details regarding the assessment of bullet quality.

- Recommendation #1 cautions against citing decision rates from this study without differentiating by bullet quality.
- Recommendation #3 recommends that sample sets used during initial training, ongoing education, and competency/proficiency tests should include a range of bullet quality.

> ***Limitations and caveats regarding rates by quality:***
>
> *The study showed trends by quality, but the study did not include enough very poor-quality questioned bullets to measure the effects of varying quality on decision rates with precision. This topic would be appropriate for further research.*

2.4. Known-questioned (KQ) comparisons vs. questioned-questioned (QQ) comparisons

BulletBB included two types of comparison sets. Known-Questioned (KQ) comparisons are analogous to casework scenarios in which a "questioned" (Q) bullet is compared to bullets from a recovered firearm; in BulletBB, KQsets contained three K bullets that were all fired from a single firearm, and one Q bullet. Questioned-Questioned (QQ) comparisons are analogous to casework scenarios in which a firearm is not recovered, and unknown bullets must be compared to each other (rather than to known exemplars); in BulletBB, QQsets contained two Q bullets. Although there were differences in decision rates between KQ vs. QQ comparisons, after controlling for other factors the distribution of decisions was relatively similar for KQ and QQ comparisons. Since KQsets provide more information to use in making comparisons, one may have expected that QQ comparisons would have a much higher rate of inconclusive responses than KQ comparisons, but the BulletBB results do not support that expectation when considering good-quality bullets. BulletBB includes limited results (small sample size) indicating that when considering poor-quality bullets, QQ comparisons were more likely to be inconclusive than KQ comparisons — this is an intuitive result, in that having multiple Ks available for comparison (rather than a single Q) would be most useful if the Qs were poor quality. The similarity in KQ and QQ decision rates (for good-quality bullets) provides some evidence to support the reliability of bullet comparisons when the firearm is not available.

The BulletBB-WG did not have recommendations derived from the implications related to KQ vs. QQ comparisons.

> ***Limitations and caveats regarding KQ vs. QQ comparisons:***
>
> *The study showed similar decision rates between KQ vs. QQ comparisons for higher-quality bullets, but the study did not include enough very poor-quality questioned bullets to measure the effects of varying quality on decision rates for KQ vs. QQ comparisons with precision. This topic would be appropriate for further research.*

2.5. Comparison of bullets from different firearms sharing class characteristics

For comparisons of bullets from different firearms (i.e., nonmated comparisons), the proportions of decisions were strongly affected by the extent to which the firearms shared class characteristics:

- Comparisons between bullets that differed in caliber almost always resulted in exclusions.[f]
- Comparisons between bullets fired from firearms of the same make, model, and caliber resulted in relatively low exclusion and high inconclusive rates.
- For comparisons of bullets from firearms of different manufacturers or models, rates were strongly affected by whether the number and width of lands and grooves were the same: bullets from different firearms with the same number and width of lands and grooves had notably lower exclusion and higher inconclusive rates.
- All erroneous IDs were on comparisons in which the bullets were the same caliber, and had the same number and width of lands and grooves.

The following recommendation is derived from the implications related to nonmated comparisons (summarized here; see Section 3 for detailed recommendations):

- Recommendation #3 recommends that sample sets used during initial training, ongoing education, and competency/proficiency tests should include challenging nonmated comparisons from firearms of the same make and model, as well as from different model firearms that share class characteristics.

> ***Limitations and caveats regarding comparisons of bullets from different firearms:***
>
> *Some agencies have policies that preclude them from excluding on individual characteristics alone (in other words, if two bullets have the same class characteristics, examiners in some agencies may only be allowed to make ID or inconclusive decisions). Few of the BulletBB participants worked for agencies that prohibited exclusions in this way (see BulletBB Appendix M1, Q#17 & Q#18). However, of the 49 participants, 18 never made exclusion decisions on bullets from the same make/model of firearm even though they were permitted to do so by their agency policy. This may have had an effect on the nonmated exclusion rates, and explain the high nonmated inconclusive rate.*

2.6. Reproducibility of comparison decisions

When participants were given the same comparison sets, they generally reproduced each other's decisions — but not always. Almost all disagreements were differences between a definitive decision (conclusion) and Inconclusive (i.e., ID vs. Inconclusive, or Exclusion vs. Inconclusive): less than 1% of the time did participants make contradictory conclusions (ID vs. Exclusion).

It is important to differentiate contradictory conclusions and disagreements regarding sufficiency for a decision:

- ***Contradictory conclusions*** — If one examiner makes an ID decision and another makes an Exclusion decision, this is a contradiction, and one of the conclusions must be an error (whether or not ground-truth source attribution is available). When ground-truth source attribution is available (as in a study or proficiency test, but not in casework), decisions that contradict known ground truth are demonstrably wrong and should be considered errors (such as an ID on a nonmated comparison).

[f] Of the comparisons of bullets fired by firearms of different calibers, all but three resulted in exclusions; the remaining three were all *Inconclusive* (two *LeanExcl*, and one *Insuff*) and involved damaged bullets.

- ***Sufficiency for a decision*** — If one examiner makes an ID decision and another makes an Inconclusive decision on the same bullets, this is a disagreement regarding whether there is sufficient information to make an ID decision; similarly, examiners who disagree regarding Exclusion vs. Inconclusive decisions are disagreeing on the sufficiency of information to make an Exclusion decision. When experts disagree regarding sufficiency, it is an oversimplification to say that they can or should be judged as correct or incorrect. If the judgments of multiple examiners are available, a given decision can be assessed as contrary to consensus, but that should not be conflated with judging a decision to be an error. Note that in BulletBB each comparison set was only seen by one to four participants (mean 2.3, median 3), limiting the ability to assess consensus on decisions for each comparison set.

The fact that examiners do not always reproduce each other's decisions has implications in multiple aspects of casework:

- ***Verification*** — The extent to which examiners reproduce decisions has direct implications for verifications in casework. Verification can be implemented either as blind verification (in which the verifier is NOT aware of the decision of the initial examiner), or non-blind verification (in which the verifier IS aware of the decision of the initial examiner). Agencies differ in how verification is implemented: in the background questionnaire responses from the participants in this study, 12% stated that in their agencies blind verification was always required, 27% indicated non-blind verification was always required, and the remaining 61% required verification only for some decisions. The fact that most agencies only verify some decisions raises the concern that some decision types are reported by labs based on the decision of a single examiner, which may be of further concern given that errors are disproportionately associated with a few individual examiners. An additional justification for more rigorous verification is that BulletBB reported no erroneous IDs were reproduced and most erroneous exclusions were not reproduced, suggesting that these errors would generally have been detected if blind verification were routinely performed.
- ***Documentation*** — The fact that examiners do not always reproduce each other's decisions also underscores the importance of documentation in casework, both for the initial examiner and verifier. The working group members reviewed the comparison sets that resulted in high error rates in the study. In some cases, the potential causes for the erroneous decisions were apparent (e.g., erroneous exclusions when JHP and FMJ bullets were compared) or could be inferred (e.g., differing decisions on comparisons in which the appearance of individual characteristics was inconsistent among the 3 Ks) — but in many cases the working group members were unable to ascertain how the decision was reached. Although this was a function of the BulletBB study design (which did not require documentation), this also serves as an illustration of why documentation of comparisons is critical for understanding the bases of decisions — particularly when examiners reach different decisions.
- ***Training*** — Variability in examiners' decisions has implications for FFTE training (both initial training as well as ongoing education), in that training should instill an understanding of the extent to which such variation occurs, the circumstances in which disagreements occur, and potential means of minimizing and mitigating such variation.
- ***Testimony*** — Variability in examiners' decisions has implications for testimony in that it would be problematic for testimony to state or imply that examiners always reach the same decisions, particularly with respect to sufficiency judgements.
- ***SOPs*** — Variability in examiners' decisions has implications for agencies' policies and procedures in that agencies may wish to assess whether enhancements or more specific interpretation criteria may help increase consistency in comparison decisions among examiners, particularly with respect to making assessments of sufficiency and determining when to make inconclusive decisions. In order to understand the extent of reproducibility of decisions in casework, agencies may wish to collect

operational reproducibility information, as recommended by the Texas Forensic Science Commission [9] and NIST [10].

The following recommendations are derived from the implications related to reproducibility of comparison decisions (summarized here; see Section 3 for detailed recommendations):

- Recommendation #3 recommends that sample sets used during initial training and ongoing education should include samples that have previously resulted in disagreements among examiners.
- Recommendation #4 makes recommendations regarding how the reproducibility of comparison decisions should be reflected during testimony.
- Recommendation #5 recommends the standardization of verification policies, in order to improve the consistency of verification among agencies.
- Recommendation #6 recommends 100% verification for all comparison decisions.
- Recommendation #7 recommends considering the variability of examiners when selecting or assigning verifiers.
- Recommendation #8 recommends capturing operational reproducibility data.
- Recommendation #9 makes recommendations regarding documentation of examinations.
- Recommendation #10 recommends that agencies review their policies and procedures with respect to the reproducibility of comparison decisions.

Limitations and caveats related to reproducibility:

The reproducibility results in BulletBB can be used as a proxy to estimate the effects of blind verification. These results should not be assumed to apply to non-blind verification, because a non-blind verifier could be influenced by the initial examiner's decision and therefore the trials cannot be considered independent.

2.7. Variation in decision rates by examiner

Decision rates varied notably among the participants in the study. The participants were all practicing FFTEs, but as a small number of volunteers, they should not be assumed to be representative of the overall population of practicing FFTEs.

The participating examiners exhibited varying tendencies towards specific decisions, showing different levels of accuracy, effectiveness, and error. The results show that examiner performance cannot be assessed using a single measure: performance is multidimensional and should consider rates of errors (FID[g] and FEX), rates of "correct" decisions (TID and TEX), and rates of inconclusive or NoValue decisions. For example, most of the participants who made no errors (FIDs or FEXs) had fewer correct decisions (TIDs and TEXs) and more inconclusive decisions than average — in other words, they were more accurate but less effective, and assessing examiners solely on error rates would be a misleading oversimplification.

A few participants made a disproportionate number of errors: two participants had more than half of all FIDs, and three participants had nearly half of all FEXs. Other forensic examiner studies have reported similar "outlier examiners" (such as [4,11,12], in each of which one participant made more than half of the erroneous IDs). The fact that any practicing FFTEs had such high error rates is an obvious concern. An

[g] BulletBB uses the following abbreviations for ID and Exclusion decisions with respect to ground truth source attribution: a "TID"(true ID) is an ID on a mated comparison set; a "FID"(false ID) is an erroneous ID on a nonmated comparison set; a "TEX"(true exclusion) is an exclusion on a nonmated comparison set; a "FEX"(false exclusion) is an erroneous exclusion on a mated comparison set.

additional concern relates to examiners who are outliers due to unusually high rates of inconclusive decisions.

Review of some BulletBB comparisons that resulted in erroneous ID responses indicates that some participants may not have a good understanding of what constitutes sufficient agreement for ID in striated marks when the quantity of striae is low. Given this, training exercises that focus on quality of agreement and consecutiveness of striae may be considered.

Agency policies addressing the problem of such outlier examiners should consider a two-part response:

- ***Prevention*** — Policies aimed at limiting the possibility of new FFTEs becoming outliers through training, testing, and continuing education. We note that these policies are recommended elsewhere in this document, but their necessity is underscored when considering the importance of preventing new examiners from becoming outliers.
- ***Detection*** — Policies aimed at identifying practicing examiners that operate outside the norm in terms of accuracy or effectiveness through verification, proficiency tests, continuous training, and handling of underperformance.

The following recommendations are derived from the implications related to variation in decision rates by examiner (summarized here; see Section 3 for detailed recommendations):

- Recommendation #2 cautions that citations of decision rates from this study should note that decision rates can vary significantly by examiner, and can be disproportionately affected by outlier examiners.
- Recommendation #3 recommends considerations for the sample sets used during initial training, ongoing education, and competency/proficiency tests (which may assist in the prevention and detection of outlier examiners).
- Recommendation #6 recommends 100% verification for all comparison decisions (which may assist in the detection of outlier examiners).
- Recommendation #7 recommends considering the variability of examiners when selecting or assigning verifiers (which may assist in the detection of outlier examiners).
- Recommendation #8 recommends capturing operational reproducibility data (which may assist in the detection of outlier examiners).
- Recommendation #11 makes recommendations regarding policies to detect practicing examiners with anomalous rates of decisions.

2.8. Variability of NoValue (not suitable) decisions

BulletBB showed inconsistency among participants in assessments of NoValue. Most participants never assessed any comparison sets as NoValue; overall, few NoValue responses were received, and those that were reported were rarely reproduced.

The BulletBB instructions asked participants to assess the value of the bullets for comparison according to the OSAC definition of NoValue as "A judgement that the bullet is inadequate to form any source conclusion due to insufficient quality or quantity of features, size, damage, or clarity of the item (i.e., any object that does not bear any class, subclass and/or individual toolmarks of value for source conclusion)."[h] Given the definition states "inadequate to form ANY source conclusion" some BulletBB-WG members were surprised that any NoValue assessments were received, given that even the poorest-quality bullets

[h] This definition was taken from the OSAC draft standard [15], which as of the time of writing has been proposed but not approved as a standard by the AAFS Academy Standards Board

in the study had at least enough of the rifling present to assess the direction of twist, and land or groove width.

The following recommendations are derived from the implications related to variability of NoValue decisions (summarized here; see Section 3 for detailed recommendations):

- Recommendation #6 recommends 100% verification for NoValue decisions.
- Recommendation #12 recommends standardizing value/suitability assessments.
- Recommendation #13 recommends documenting the basis for NoValue decisions.

2.9. Comparison of different ammunition types

Mated comparisons in which different types of ammunition were fired from the same firearm had a high rate of erroneous exclusions: ***almost all the erroneous exclusions in the study were on QQ comparisons with different types of ammunition (FMJ vs. JHP)***.

Note that the AFTE Procedures Manual states, "The discipline recognizes that an elimination of a firearm by other than class characteristics is possible but that such an elimination is an exceptional situation" ([13], reiterated by SWGGUN in [14]).

The following recommendations are derived from the implications related to comparisons of different types of ammunition (summarized here; see Section 3 for detailed recommendations):

- Recommendation #3 recommends that sample sets used during initial training, ongoing education, and competency/proficiency tests should include a variety of different bullet jacket types and compositions.
- Recommendation #14 recommends that agency policies should have specific guidelines pertaining to exclusion results from the examination of samples consisting of different ammunition types.

> ***Limitations and caveats regarding comparison of different ammunition types:***
>
> *For each caliber, the study generally included JHP ammunition from one manufacturer and FMJ ammunition from another manufacturer, and therefore the study could not evaluate the effects of different ammunition type (e.g., JHP, FMJ) vs. different ammunition manufacturer/model/lot (e.g., Winchester USA White Box, Remington UMC, Hornady Critical Duty) vs. different jacket material (e.g., copper, copper-zinc alloy, steel, synthetic). This topic would be appropriate for further research.*

2.10. Caveats and limitations

BulletBB should be considered an exploratory, hypothesis-generating study: an observational study rather than an experimental study. Although the size of the study limits the precision of the measured rates, the results indicate relative effects, and factors for further detailed study.

When discussing the results of this study, the following caveats and limitations should be emphasized:

- Although all of the participants were practicing FFTEs, the study was based on a small number of volunteers, and therefore should not be assumed to be representative of the overall population of practicing FFTEs.

- Because errors were disproportionately made by a few participants, those participants had an outsized effect on the measured rates — it is not known how prevalent such examiners are in the overall FFTE population.
- Although the study team made efforts to ensure that the comparison samples included in the study were broadly representative of operational casework, the samples included in the study should not be assumed to be representative of any specific case or of the casework encountered by a specific agency.
- The study shows trends, but the small sample size and limited number of participants limit the precision of any specific rates. Rates reported in this study are not intended to be taken as precise measures of operational decision rates.
- These results do not account for operational "safety nets" implemented by laboratories/agencies, such as consultation, verification, conflict resolution, technical review, and other quality assurance procedures.
- Participants were instructed to conduct the comparisons with the same diligence employed in their operational casework. However, since the participants knew they were being tested, some examiners may have taken the study less seriously or more seriously than casework. The BulletBB results are anonymous, and cannot be attributed to specific examiners; therefore, there was no professional consequence if some participants were more rushed or more/less cautious than they might have been in casework.
- Participants may or may not have followed the procedures and policies they use in casework, some of which (such as requirements for detailed documentation) may affect how examiners make decisions.
- Not all firearms examiners use the same conclusion scale in casework. BulletBB used the five-level AFTE/OSAC scale [15,16], but that was only used operationally by 37% of participants; 63% of participants use a three-level conclusion scale that does not differentiate between subcategories of inconclusive.[i]
- The study reports results for bullets, which cannot be assumed to apply to cartridge cases.

[i] The three-level scale consists of (Identification, Inconclusive, Exclusion) — the five-level AFTE/OSAC scale subdivides Inconclusive into three subcategories (Insufficient Support for Identification, Insufficient Support for Either Exclusion or Identification, Insufficient Support for Exclusion) — these are abbreviated in BulletBB as (LeanID, Insuff, LeanExcl).

3. Recommendations

The following recommendations are limited to those unanimously supported by the BulletBB-WG: there were no objections among working group members to any of these recommendations.

Recommendation #1 **We strongly caution against citing decision rates from this study that do not differentiate by firearm manufacturer/model, ammunition type (manufacturer, material, and/or design), or bullet quality (extent of damage). Citing specific overall rates from this study would be problematic and potentially misleading because such rates are an average across an arbitrary mix of these factors, the proportions of which may vary significantly from casework at any given agency.**

- **When citing decision rates from this study, the citation should refer to the firearms and ammunition used in the study. Decision rates from a study with a given set of firearm:ammunition combinations should not be assumed to necessarily apply to other firearm:ammunition combinations.**
- **When citing decision rates from this study, the citation should refer to the quality (extent of damage) of the bullets being compared. Decision rates are affected by bullet quality.**
- **When citing decision rates from this study, the citation should note how the denominator is calculated: using all presentations (PRES), all comparisons but omitting NoValue responses (COMP), or limited to conclusive calls (CALLS, omitting inconclusive and NoValue responses).**

Recommendation #2 **When citing decision rates from this study, the citation should note that decision rates can vary significantly by examiner. Therefore, the decision rates from a study with a given pool of examiners may differ from rates from a study with a different pool of examiners, and cannot be assumed to apply to any given individual examiner. Decision rates can be disproportionately affected by outlier examiners.**

Recommendation #3 **Sample sets used during initial training, ongoing education, and competency/proficiency tests should include the following:**[j]

- **Comparisons of samples using a variety of firearm models, incorporating both polygonal and conventional rifling.**
- **Comparisons of samples using a variety of different bullet jacket types (e.g., full metal jacket (FMJ), jacketed hollow-point (JHP), total metal jacket (TMJ), coated/washed/plated bullets, etc.) and jacket composition (e.g., copper-zinc alloy, steel, synthetic, etc.).**
- **For each firearm model used, comparisons of firearm:ammunition combinations with a range of comparability of class and individual characteristics among known bullets. These sample sets must include firearms that reliably reproduce toolmarks (good comparability of known bullets), as well as some that either produce few toolmarks overall (minimal**

[j] Note that the *AFTE Training Manual* [19] recommends these types of exercises.

markings) and those that do not reproduce toolmarks reliably (poor comparability of known bullets).

- **Comparisons of samples with a range of bullet quality (from pristine to damaged bullets, and from intact bullets to those with separated and/or fragmented jackets). These should include bullets with a range of quality and quantity of information selected specifically to evaluate how examiners make suitability decisions.**
- **Comparisons of challenging nonmated comparisons from firearms of the same make and model — as well as from different model firearms that share the same direction of twist, number of lands and grooves, and width of lands and grooves.**

The following special case or unusually difficult samples should be included in sample sets used during initial training and ongoing education:

- **Comparisons of a variety of error-prone and difficult-to-interpret samples, so that examiners receive feedback in a nonpunitive training environment to further develop and maintain their expertise.**
- **Comparisons of samples that have previously resulted in disagreements among examiners, so examiners have a detailed understanding of making decisions when it is debatable whether there was sufficient information to make a given conclusion (ID vs. Inconclusive, or Exclusion vs. Inconclusive).**
- **Comparisons that focus on quality of agreement and consecutiveness of striae, to assist examiners in developing a good understanding of what constitutes sufficient agreement for ID in striated marks when the quantity of striae is low.**

Recommendation #4 **The study results demonstrate that examiners do not always reach the same decisions for a given comparison. This should be reflected during testimony:**

- **Statements in testimony should not indicate an expectation that examiners always reach the same decisions.**
- **When this topic is raised in testimony, examiners should be encouraged to elaborate on the differences between contradictory opinions (ID vs. Exclusion) and sufficiency judgements (ID vs. Inconclusive, or Exclusion vs. Inconclusive).**
- **When this topic is raised in testimony, it should be made clear that almost all of the disagreements that occurred were regarding whether there was sufficient information to make a given conclusion (ID vs. Inconclusive, or Exclusion vs. Inconclusive) — the study indicated that examiners very rarely contradicted each other (ID vs. Exclusion).**

Recommendation #5 **The firearms community should work to standardize verification policies, in order to improve the consistency of verification among agencies.** [k]

Recommendation #6 **Agency policy should dictate 100% verification for all comparison decisions (identifications, inconclusives, exclusions), as well as assessments of NoValue. Blind verification is preferred when possible; at least a portion of the verifications should be blind.**

[k] As proposed in the OSAC *Standard for Verification of Source Conclusions in Toolmark Examinations* [20].

Recommendation #7 **The variability of examiners should be considered when selecting or assigning verifiers:**

- **Examiners should not be permitted to select their own verifiers.**[l]
- **Verification assignments should be randomized or rotated among all technical staff, when practical.**
- **Verifiers should be examiners who were not consulted in the primary examination, when possible.**
- **The technical reviewer and verifier should be different examiners, when possible.**

When labs do not have the staff to conduct verification as recommended here, alternative methods should be considered, such as agreements with neighboring agencies or other service providers.

Recommendation #8 **Agencies should consider capturing the proportion of primary examiner decisions with respect to verifier decisions, in order to capture operational reproducibility data.**[m] **This should include a) instances in which the primary examiner and the verifier disagree but reach a common decision after discussion, and b) instances in which the primary examiner and the verifier cannot reach a common decision and a conflict resolution process is used.**

This information may be used to instill an understanding of how the examiners in an agency make decisions, which may potentially help build consistency in decisions, and flag areas in which improvements to standard operating procedures might be warranted — as well as tracking trends for discussion and conflict resolution outcomes.

Recommendation #9 **The level of documentation for an examination should be adequate for an independent examiner to understand the basis for the decision.**[n]

- **Verifications should be documented prior to discussions between the primary and verifying examiner.**
- **Decisions that rely on the interpretation of individual characteristics should always be documented with photographs, with annotations when the areas/characteristics used to support the decision are not readily apparent.**

Recommendation #10 **Agencies may wish to review their policies and procedures and assess whether enhancements or more specific details and interpretation criteria may help increase consistency in comparison decisions among examiners, particularly with respect to making assessments of sufficiency and determining when to make inconclusive decisions.**

[l] As recommended in the OSAC *Standard for Verification of Source Conclusions in Toolmark Examinations* [20].
[m] As recommended by the Texas Forensic Science Commission [9] and NIST [10].
[n] As recommended in the ANAB *Accreditation Requirements for Forensic Testing and Calibration (AR 3125)* [21], and the AFTE *Procedures Manual* [13]

Recommendation #11 **In order to detect practicing examiners with anomalous rates of decisions, we recommend the following:**

- **A verification policy should be designed to detect unsupportable decisions, patterns of inconclusive decisions in which definitive conclusions would be supported, or a pattern of nonconsensus decisions — with follow-up action.**
- **Current examiners should undergo continuous training exercises which include exercises with challenging comparisons, the results of which would be used to point out areas for further examiner training, if appropriate.**
- **Training should be designed to present novel comparisons to examiners that may challenge their understanding of the conditions for reaching a certain decision.**
- **Agency management must take indications of underperformance (e.g., excessive conflict resolution, poor performance on training exercises or tests, etc.) seriously and respond appropriately.**

Recommendation #12 **The firearms community should work to standardize how value/suitability assessments are made, in order to improve the consistency of these determinations among examiners.**

Recommendation #13 **The basis for all value/suitability decisions should be documented. All examinations that result in a NoValue or not suitable decision (i.e., examinations that do not proceed to a comparison decision) should undergo verification or technical review.**

Recommendation #14 **Agency policy should have specific guidelines pertaining to exclusion results from the examination of samples consisting of different ammunition types (e.g. FMJ vs. JHP). Such guidelines should have specific criteria, including but not limited to requirements for documentation, quality assurance, and technical review.[o]**

[o] See [13] and [14] for examples.

4. Conclusions

BulletBB was conducted to provide information for agency/laboratory managers, forensic practitioners, policy makers, and the legal community regarding the relative effects of a variety of factors on the accuracy and reproducibility of decisions made by FFTEs when comparing bullets. In this document, the BulletBB-WG is building on the results of BulletBB to create a set of actionable recommendations that we believe will be of value to the firearms examination community in improving and enhancing the practice of bullet comparisons.

References

1. Hicklin RA, Parks CL, Dunagan KM, Emerick BL, Richetelli N, Chapman WJ, et al. Accuracy and Reproducibility of Bullet Comparison Decisions by Forensic Examiners. Forensic Sci Int. 2024;365(112287). https://doi.org/10.1016/j.forsciint.2024.112287
2. OSAC Human Factors Committee. Human Factors in Validation and Performance Testing of Forensic Science (OSAC Technical Series 0004). 2020. https://www.nist.gov/system/files/documents/2020/05/22/OSACTechSeriesPub_HF in Validation and Performance Testing of Forensic Science_March2020.pdf. https://doi.org/10.29325/OSAC.TS.0004
3. Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. Accuracy and reliability of forensic latent fingerprint decisions. Proc Natl Acad Sci U S A. 2011;108(19). https://doi.org/10.1073/pnas.1018707108
4. Hicklin RA, McVicker BC, Parks C, LeMay J, Richetelli N, Smith M, et al. Accuracy, Reproducibility, and Repeatability of Forensic Footwear Examiner Decisions. Forensic Sci Int. 2022;339(111418). https://doi.org/10.1016/j.forsciint.2022.111418
5. Hicklin RA, Eisenhart L, Richetelli N, Miller MD, BelCastro P, Burkes T, et al. Accuracy and Reliability of Forensic Handwriting Comparisons. Proceedings of the National Academy of Sciences. 2022;119(32):e2119944119. https://www.pnas.org/doi/abs/10.1073/pnas.2119944119https://doi.org/10.1073/pnas.2119944119
6. Monson KL, Smith ED, Peters EM. Accuracy of comparison decisions by forensic firearms examiners. J Forensic Sci. 2023 Jan 1;68(1):86–100. https://doi.org/10.1111/1556-4029.15152
7. Best BA, Gardner EA. Reproducibility of Bullet Comparison Conclusions. AFTE Journal. 2024;56(1):63–6.
8. Fadul Jr TG, Hernandez GA, Stoiloff S, Gulati S. An Empirical Study to Improve the Scientific Foundation of Forensic Firearm and Tool Mark Identification Utilizing Consecutively Manufactured Glock EBIS Barrels with the Same EBIS Pattern. NIJ Report. 2013;
9. Texas Forensic Science Commission. Final Report on Complaint No. 21.27; National Innocence Project, University of Colorado Law School (Houston Police Dept Crime Lab; Firearms/Tool Marks). 2024 Apr.
10. Swofford H, Lund S, Iyer H, Butler J, Soons J, Thompson R, et al. Inconclusive decisions and error rates in forensic science. Forensic Science International:Synergy. 2024;8:100472. https://doi.org/10.1016/j.fsisyn.2024.100472
11. Busey TA, Heise N, Hicklin RA, Ulery BT, Buscaglia J. Characterizing missed identifications and errors in latent fingerprint comparisons using eye-tracking data. PLoS One. 2021;16(5 May). https://doi.org/10.1371/journal.pone.0251674
12. Hicklin RA, Richetelli N, Taylor A, Buscaglia J. Accuracy and reproducibility of latent print examiner decisions on comparisons resulting from searches of an automated identification system. Forensic Sci Int. 2025 Apr;(370). https://doi.org/10.1016/j.forsciint.2025.112457
13. AFTE. Technical Procedures Manual. 2015.
14. SWGGUN. Elimination Factors Related to FA/TM Examinations. 2016.
15. OSAC Firearms & Toolmarks Subcommittee. Standard Scale of Source Conclusions and Criteria for Toolmark Examinations (OSAC Proposed Standard; v. 1.0). 2019; https://www.nist.gov/system/files/documents/2020/03/24/100_fatm_roc_and_criteria_standard_asb_mar2019_OSAC Proposed.pdf
16. AFTE Criteria for Identification Committee. Theory of identification, range striae comparison reports and modified glossary definitions. AFTE Journal. 1992;24(3):336–40.

17. AFTE. Glossary, 6th Edition. 2013.
18. AFTE Criteria for Identification Committee. Theory of identification, range striae comparison reports and modified glossary definitions. AFTE Journal. 1992;24(3):336–40.
19. AFTE. Training Manual (Version 1.0). 2021 Feb.
20. OSAC Firearms & Toolmarks Subcommittee. Standard for Verification of Source Conclusions in Toolmark Examinations (Proposed Standard). 2020.
21. ANSI National Accreditation Board (ANAB. Accreditation Requirements for Forensic Testing and Calibration (AR 3125). 2023.

Appendix A. Glossary, Abbreviations, and Acronyms

This section defines terms and acronyms as they are used in this report. Quoted definitions are from the AFTE 2013 Glossary [17]. For further explanation, please see the BulletBB journal article [1].

AFTE	Association of Firearm and Tool Mark Examiners.
ANAB	American National Standards Institute (ANSI) National Accreditation Board
Ammunition	"One or more loaded cartridges consisting of a primed cartridge case, propellant, and with or without one or more projectiles. Also referred to as fixed ammunition or live ammunition (slang term)." [17]
Bullet	"A non-spherical projectile for use in a rifled barrel." [17]
Cartridge	"A single unit of ammunition consisting of the cartridge case, primer, propellant, and with or without one or more projectile(s). Also applies to a shotshell." [17]
Cartridge Case / Case	"The container for all the other components which comprise a cartridge. Serves as a gas seal during the firing of a cartridge." [17]
Class characteristics	"Measurable features of a sample which indicate a restricted group source. They result from design factors and are determined prior to manufacture." [17]
Comparison set (Cset)	In this study, a set of bullets provided to participants for comparison, containing either 2 *Questioned bullets* (QQset) or 1 *Questioned bullet* and 3 *Known Bullets* (KQset).
Conclusion	In this study, a decision of *ID* or *Excl*, sysnoymous with *definitive decision*. (Compare to *Inconclusive*)
Consensus	In this study, the extent to which examiners concur on a given decision for a given comparison set — particularly when a majority or supermajority of responses concur on one decision.
Decision rate	In this study, the proportion of bullet comparisons that result in a given decision.
Definitive decision	In this study, a decision of *ID* or *Excl*, sysnoymous with *conclusion*. (Compare to *Inconclusive*)
Error	(With respect to examiners' decisions) In this study, a definitive source decision (*ID* or *Excl*) that is verifiably false with respect to known ground truth. *Insuff* and *NoValue* decisions cannot be assessed in terms of accuracy (or correctness, or error) with respect to ground truth. (Compare to *Incorrect*)
Exclusion (Excl)	In the instructions for this study, *Excl* was defined as An expert opinion that two items of toolmark evidence were marked by different tools. This is the strongest statement of non-association expressed in forensic firearm and toolmark examination. An Exclusion is justified when the observed characteristics of the items in question provide extremely strong support for the proposition that they were marked by different tools and extremely weak or no support for the proposition that the two were marked by the same tool. This conclusion is based on 1) demonstrable differences in class, subclass or individual characteristics, 2) task-relevant information, and 3) the cumulative results of training and casework examinations that have either been performed, peer reviewed, and/or published in peer-reviewed journals. In other words: You are certain that these were NOT fired from the same firearm.
False exclusion (FEX) / Erroneous exclusion	In this study, an exclusion on a mated Cset. (Contrary to ground truth)
False ID (FID) / Erroneous ID	In this study, an ID on a nonmated Cset. (Contrary to ground truth)
FMJ	"Full Metal Jacket Bullet – A projectile in which the bullet jacket encloses the entire bullet, with the usual exception of the base. Also called full jacketed, full patch, full metal case, metal cased, metal patched, and ball ammunition." [17]
Grooves	"Depressed or cut channels in the bore of a firearm barrel to impart rotary motion to a projectile." [17]
Ground truth	In this study, definitive knowledge that a comparison is either mated or nonmated.
Identification (ID)	In the instructions for this study, Identification was defined as An expert opinion that two items of toolmark evidence were marked by the same tool. Source identification is the strongest statement of association expressed in forensic firearm and toolmark examination. An Identification conclusion is justified when, in the examiner's opinion, the observed similarities in characteristics of the items in question provide extremely strong support for that proposition and negligible support for the proposition that the two items were marked by different tools. In this context, "negligible support" means that based on 1) known empirical research and validation studies, and 2) the cumulative results of training and casework examinations that have either been performed, peer reviewed, or published in peer-reviewed scientific literature, it is the examiner's opinion that it is extremely unlikely any firearms or tools other than those identified are capable of producing marks exhibiting sufficient agreement for identification.

	In other words: You are certain that these were fired from the same firearm.
Inconclusive	In this study, a comparison decision of *LeanID*, *InSuff*, or *LeanExcl*. (Compare to *Definitive decision*)
Incorrect	(With respect to examiners' decisions) In this study, a leaning decision that is verifiably false with respect to known ground truth. *Insuff* and *NoValue* decisions cannot be assessed in terms of accuracy (or correctness, or error) with respect to ground truth. (Compare to *Error*)
Individual characteristics	"Marks produced by the random imperfections or irregularities of tool surfaces. These random imperfections or irregularities are produced incidental to manufacture and/or caused by use, corrosion, or damage. They are unique to that tool to the practical exclusion of all other tools." [17]
Insufficient Support for Either Exclusion or Identification (Insuff)	In this study, a subcategory of *Inconclusive*. In the instructions for this study, Insufficient Support for Either Exclusion or Identification was defined as "An expert opinion that the observed similarities and differences in characteristics of the items neither substantially support nor substantially refute the proposition that the bullets were marked by the same source tool. An Insufficient Support for Either Exclusion or Identification conclusion is justified when, in the examiner's opinion, there is agreement of all discernible class characteristics, but, due to an absence, insufficient agreement and/or disagreement, or lack of reproducibility of individual characteristics, no other conclusion can be reached. In other words: You are inconclusive, and not leaning toward identification or exclusion."
Insufficient Support for Exclusion (LeanExcl)	In this study, a subcategory of *Inconclusive*. In the instructions for this study, Insufficient Support for Exclusion was defined as "An expert opinion that the observed similarities and differences in characteristics of the items in question are insufficient for conclusive exclusion but provide substantial support for the proposition that the two items were marked by different tools and weak support for the proposition that they were marked by the same tool. An Insufficient Support for Exclusion conclusion is justified when, in the examiner's opinion, there is agreement of all discernible class characteristics and some disagreement of individual characteristics similar to that which has been demonstrated by known nonmatches, but insufficient for Exclusion. In other words: You are leaning toward an exclusion but do not have enough differences to support that conclusion."
Insufficient Support for Identification (LeanID)	In this study, a subcategory of *Inconclusive*. In the instructions for this study, Insufficient Support for Identification was defined as "An expert opinion that the observed similarities and differences in characteristics of the items are insufficient for conclusive identification but provide substantial support for the proposition that the two items were marked by the same tool and weak support for the proposition that the two were marked by different tools. An Insufficient Support for Identification conclusion is justified when, in the examiner's opinion, there is agreement of all discernible class characteristics and some agreement of individual characteristics but not exceeding the best KNM [known nonmatch] and is therefore insufficient for Identification. In other words: You are leaning toward an identification but do not have enough similarities to support that conclusion."
Jacket / Bullet jacket	"The envelope enclosing the core of a projectile that is typically of metallic construction." [17]
JHP	"Jacketed Hollow Point Bullet – A bullet having a metal jacket enclosing a lead alloy core. The entire bullet is enclosed except for the nose, which has a cavity." [17]
Known bullet (K)	A representative bullet collected by discharging a firearm under controlled settings for comparison or analysis. Also known as a test fire. (Compare to *Questioned bullet*)
KQset	Known-Questioned comparison set — In this study, a comparison set containing one questioned bullet and three known bullets.
Land	"The raised portion between the grooves in a rifled barrel." [17]
Land and groove impressions	"Impressed areas on the bearing surface of a bullet caused by a bullet engaging with the rifling in the barrel of a firearm." [17]
Mated	In this study, a Cset in which all bullets are known to have originated from the same source, based on definitive (ground truth) *a priori* knowledge.
NIST	National Institute of Standards and Technology.
No Value / Unsuitable for Source Conclusion (NoValue)	In the instructions for this study, *NoValue* was defined as A judgement that the bullet is inadequate to form any source conclusion due to insufficient quality or quantity of features, size, damage, or clarity of the item (i.e., any object that does not bear any class, subclass and/or individual toolmarks of value for source conclusion).
Nonconsensus	In this study, a decision on a given comparison set that differs from the majority of responses on that comparison set.
Nonmated	In this study, a Cset in which the bullets are known to have originated from different sources, based on definitive (ground truth) *a priori* knowledge.
Of Value for Source Conclusion (OfValue)	In the instructions for this study, Of*Value* was defined as

	A preliminary judgement that the bullet has potentially sufficient class, subclass and/or individual characteristics for further evaluation, examination, or comparison with other known-source or questioned-source bullets for potential source conclusion.
Polygonal rifling (PR)	"Lands and grooves having a rounded profile instead of the traditional rectangular profile. Polygonal rifling is often seen in hammer forged barrels." [17]
QQset	Questioned -Questioned comparison set — In this study, a comparison set containing two questioned bullets.
Quality	In this study, the extent of damage to a bullet, defined in terms of the amount of rifling available for comparison, and assessed as a quality grade of A through D based on participant responses.
Questioned bullet (Q)	In casework, a bullet of unknown source or origin. In this study, a bullet of varied quality collected to simulate the quality and attributes of discharged bullets found in casework. (Compare to *Known bullet*)
Reproducibility	In this study, the rate of agreement among multiple examiners on the same data.
Rifling / General rifling characteristics	"The number, width, and direction of twist of the lands and grooves in a barrel of a given caliber firearm." [17]
Subclass characteristics	"Features that may be produced during manufacture that are consistent among items fabricated by the same tool in the same approximate state of wear. These features are not determined prior to manufacture and are more restrictive than class characteristics." [17]
True exclusion (TEX)	In this study, an exclusion on a nonmated Cset. (Consistent with ground truth)
True ID (TID)	In this study, an ID on a mated Cset. (Consistent with ground truth)