IN THE IOWA SUPREME COURT

DONNIE LEE WYLDES JR., Applicant,	:	**Supreme Court No. 24-1123**
v.	:	**WAYNE CO. NO. PCCV022960**
STATE OF IOWA, Respondent.	:	**AMICUS BRIEF IN SUPPORT OF APPLICANT**

BRIEF OF *AMICI CURIAE* CRIMINAL LAW SCHOLARS, SCIENTISTS AND STATISTICIANS IN SUPPORT OF APPLICANT DONNIE LEE WYLDES JR.

Matthew Sease
Iowa Bar No. AT0010484
104 SW 4th Street, Ste. A
Des Moines, IA
msease@seasewadding.com
(515) 883-2222

Donald P. Salzman*
1440 New York Avenue, N.W.
Washington, D.C. 20005
donald.salzman@probonolaw.com
(202) 371-7983

Marley Ann Brumme*
500 Boylston Street
Boston, MA 02116
marley.brumme@probonolaw.com
(617) 573-4861

Hannah Henderson*
500 Boylston Street
Boston, MA 02116
hannah.henderson@probonolaw.com
(617) 573-4878

Counsel for Amici

* Pro Hac Vice Motion Filed Herewith

TABLE OF CONTENTS

TABLE OF AUTHORITIES

CASES

OTHER AUTHORITIES

STATEMENT OF INTEREST OF AMICI CURIAE[1]

Amici are law professors, scientists, and statisticians at America's leading universities who have devoted a substantial part of their teaching, work, research and/or writing to criminal law and procedure, including issues concerning the accuracy and reliability of evidence and equity in criminal outcomes. Their work has been published by major university presses and in leading scientific and law journals. *Amici* are listed in the Appendix.

INTRODUCTION

By ensuring only valid, reliable expert testimony is admitted into evidence, courts play a crucial role in preventing unjust wrongful convictions. "The fair administration of justice requires that science is accurately and effectively communicated to the fact finders" in judicial proceedings. Itiel E. Dror & Nicholas Scurich, *(Mis) use of Scientific Measurements in Forensic Science*, 2 Forensic Sci. Int'l: Synergy 333, 333 (2020).

Firearm and toolmark ("FA/TM") examiners purport to "match" spent ammunition to *one particular firearm*—not a type of firearm, make of firearm, or model of firearm—by visually comparing spent ammunition recovered from a

[1] Both parties have consented to this *amici* brief. No counsel for a party authored this brief in whole or in part, nor made a monetary contribution intended to fund the preparation or submission of this brief. No persons other than *amici*'s counsel made a monetary contribution to this brief's preparation or submission.

crime scene with spent ammunition from a test fire from the firearm suspected to have been used in the crime. FA/TM evidence is premised on the *unproven assumption* that each firearm leaves unique, individualized markings on spent ammunition. Routinely admitted by courts for decades, the scientific community has now sounded the alarm that the "fair administration of justice" is severely threatened by the admission of—or even the unrebutted presentation of—FA/TM evidence. The current scientific consensus is that FA/TM evidence lacks scientific support from well-designed, empirical studies.[2] All studies purporting to demonstrate the validity and reliability of FA/TM evidence have been poorly designed, most having been "developed by insular communities of nonscientist practitioners"—members of the Association of Firearm and Tool Mark Examiners ("AFTE"), i.e., practicing FA/TM examiners rather than scientists—"who did not incorporate effective statistical methods." William A. Tobin, H. David Sheets & Clifford Spiegelman, *Absence of Statistical and Scientific Ethos: The Common Denominator in Deficient Forensic Practices*, 4 Stats. & Pub. Pol'y, at 1 (2017), https://www.tandfonline.com/doi/pdf/10.1080/2330443X.2016.1270175 ("Tobin *et al.*).

[2] *Amici* understand that Mr. Wyldes has briefed the subjective methodology employed by FA/TM examiners and will not address those serious flaws. *See* Post Trial Br. in Supp. of Mot. for New Trial at 20-23, 35-40 (11/9/2023).

Existing FA/TM studies' myriad design flaws render their conclusions statistically—and scientifically—unreliable. Poorly designed studies produce narrowly applicable data from which proponents of FA/TM analysis draw overbroad and improper inferences. For instance, while interested proponents of FA/TM analysis assert it has an error rate of "1%," Susan Vanderplas, Kori Khan, Heike Hofmann & Alicia Carriquiry, *Reply to Response by FBI Laboratory filed in Illinois v. Winfield and Affidavit by Biederman et al.*, University of Nebraska–Lincoln, Department of Statistics: Faculty Publications, at 34 (2022), filed in *U.S. v. Kaevon Sutton* (2018 CF1 009709) (Exhibit 1), the scientific community has overwhelmingly concluded that, because of the design of existing studies, (i) this number is woefully underinclusive and (ii) existing studies have not actually or appropriately shown a "known or potential rate of error." *See Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579, 593-94 (1993); *see also Johnson v. Knoxville Cmty. Sch. Dist.*, 570 N.W.2d 633, 637 (Iowa 1997). The import of a "known or potential rate of error" is obvious: "[k]nowing the error rates in a particular forensic domain is a vital measurement needed to ascertain the weight of the evidence. The appropriate weight of the evidence cannot be known without some sense of the rates at which the technique errs." Dror & Scurich, *supra*, at 333; *see also* Thomas D. Albright, *How to Make Better Forensic Decisions*, 119 Proc. Nat'l Acad. of Sci., at 7 (2022), https://doi.org/10.1073/pnas.2206567119 (noting that

the trier of fact "truly needs to know" the "examiner's decision, together with an estimate of the probability that the decision is correct"). Of the few more appropriately designed studies, the most recent—a collaboration between the FBI and the Ames Laboratory—demonstrates that examiners are unable to repeat *their own conclusions* between 21% and 70% of the time. (Ames II, *infra* Section I.B.2.)

Despite this astounding absence of demonstrated validity, FA/TM evidence is often presented to juries in broad categorical strokes. Testifying examiners report their conclusions by declaring a definite identification (i.e., crime scene ammunition came from the suspected firearm), definite exclusion (i.e., crime scene ammunition did not come from the suspected firearm), or inconclusive (i.e., examiner cannot state one way or another). Such confident, definitive statements, combined with the uncertainty of how often the FA/TM methodology actually results in errors, can easily cause jurors to be misled—or, at a minimum, confused—about the import and weight of FA/TM evidence.

Categorical—but invalid—FA/TM evidence played a critical role in convicting Mr. Wyldes. The firearm examiner, Agent Robert Harvey, examined casings and a bullet collected from the scene and from a gravel road approximately 0.2 miles away, as well as over 18,000 casings collected from several shooting ranges. D0408-13, D0400, Wyldes's Criminal Trial Transcript 174-76, 178, 179-

90, 632 (hereinafter cited T.). Using the AFTE's Theory of Identification, Agent Harvey compared the markings on the crime scene casings and gravel road casings, and determined that they had been fired by the "same firearm," (an "individualization" determination) due to markings left on the fired ammunition by the gun that had fired them. T. 632. Agent Harvey explained that the markings were "very unique," T. 633, testifying, "I had never seen anything quite as severe in all of the casings that I have looked at . . . I thought that these casings were fired from a gun that had a very unique problem, a severe problem," T. 633-636.

Juries often place substantial weight on expert testimony, as they did here. Courts have long recognized this fundamental premise, cautioning that expert testimony can unduly shape jurors' perceptions in criminal trials. *See, e.g.*, *Clemons v. State*, 392 Md. 339, 372 (2006) (warning "[l]ay jurors tend to give considerable weight to 'scientific' evidence when presented by 'experts' with impressive credentials") (citation omitted). Given this tendency, judges must serve as gatekeepers and admit or allow the presentation of only scientifically valid expert testimony.

To prevent future wrongful convictions—and unfair proceedings—based on now-discredited FA/TM evidence, *amici* respectfully urge this Court to acknowledge the power and consequence of this new scientific understanding. Here, science undermines the outcome of Mr. Wyldes's trial and calls his

conviction into doubt, and at the very least renders the FA/TM evidence proffered inadmissible. Amici ask this Court to consider that FA/TM evidence is inadmissible in criminal courts in Iowa because it is not scientifically validated, is inherently unreliable, and does not meet the well-established standards in *Leaf v. Goodyear*. *See Leaf v. Goodyear Tire & Rubber Co.*, 590 N.W.2d 525, 533 (Iowa 1999). Iowa courts consider the *Daubert* factors instructive in assessing "whether the reasoning or methodology underlying the testimony is scientifically valid and . . . whether that reasoning or methodology properly can be applied to the facts in issue." *See Hutchison v. Am. Fam. Mut. Ins. Co.*, 514 N.W.2d 882, 888 (Iowa 1994); *see also Johnson*, 570 N.W.2d at 637 (citing *Daubert*, 590 U.S. at 593–94) (summarizing *Daubert* factors as "whether the theory or technique (1) can be (and has been) tested, (2) has been subjected to peer review and publication, (3) is generally accepted within the relevant scientific community, and (4) has a known or potential rate of error").

ARGUMENT

I. THE SCIENTIFIC COMMUNITY HAS REJECTED THE NOTION THAT FA/TM IS RELIABLE OR SCIENTIFICALLY VALIDATED

The scientific community has re-assessed and forcefully discredited FA/TM because it has not yet been demonstrated to be scientifically valid. Studies on the reliability of FA/TM suffer from myriad methodological flaws rendering them invalid, and the few better-designed studies reveal worrying error rates.

A. Four Reports By Committees Of Experts From The Scientific Community Have Concluded That FA/TM Is Not Scientifically Valid

Between 2008 and 2017, three separate panels of distinguished independent experts from the broader scientific and academic community (not limited to FA/TM)—convened by the National Academy of Sciences ("NAS") and the President's Council of Advisors on Science and Technology ("PCAST")—authored four separate reports raising serious concerns about the scientific validity and reliability of FA/TM evidence (as well as other "pattern-matching" fields).[3] The committees consisted of independent scientists and professors with expertise in physics, chemistry, biology, materials science, engineering, biostatistics, statistics, scientific methodology and study design, and medicine, as well as judges and lawyers—rather than toolmark examiners, whose financial and professional stake in the continued embrace of their discipline is apparent. Each national

---

[3] *See generally*, National Research Council, *Ballistic Imaging* (2008), https://doi.org/10.17226/12162 ("Ballistic Imaging"); National Research Council, Committee on Identifying the Needs of the Forensic Sciences Community, *Strengthening Forensic Science in the United States: A Path Forward* (2009), https://www.ojp.gov/pdffiles1/nij/grants/228091.pdf ("NAS Report"); President's Council of Advisors on Science and Technology, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* (2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf ("PCAST Report"); President's Council of Advisors on Science and Technology, *An Addendum to the PCAST Report on Forensic Science in Criminal Courts* (2017), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_addendum_finalv2.pdf ("PCAST Addendum").

scientific committee heard testimony from forensic scientists, reviewed nearly every available journal article and study involving toolmark examination, and read every article or study submitted by members of the forensic community.[4] These bodies were uniquely qualified to determine whether this field is based on valid, reliable scientific principles or methodologies.

The conclusions of these committees were uniform and devastating: the "fundamental assumptions" underlying toolmark examination, including the claimed uniqueness of all striae, have not been proven; the theory of toolmark identification—i.e., "individualization" or matching any particular tool to a particular mark—is "not a scientific theory"; the method is subjective; and there is insufficient empirical evidence establishing either the scientific validity of the field or even estimating the reliability of toolmark examinations.[5] The committees concluded that FA/TM examination (i) consists of applying a subjective methodology to an unvalidated assumption and (ii) lacks the studies necessary to demonstrate it produces reproducible, repeatable, and valid results.

---

[4] See, e.g., *PCAST Report*, *supra*, at 2, 155-160; *NAS Report*, *supra*, at xx, 2-3; *Ballistic Imaging*, *supra*, at xiii–xvi; *PCAST Addendum*, *supra*.

[5] *See* Ballistic Imaging, *supra*, at 3; NAS Report, *supra*, at 154; PCAST Report, *supra*, at 47, 60, 104, 111, 113.

B. Since The Ballistic Imaging, NAS, And PCAST Reports, The Scientific Community Has Continued To Cast Doubt Upon FA/TM

Since publication of the Ballistic Imaging, NAS, and PCAST Reports, the scientific community has elaborated upon and amplified the criticisms of FA/TM evidence. Numerous publications by authors spanning multiple disciplines have continued to call FA/TM into question, including but not limited to:

- David L. Faigman, Nicholas Scurich, & Thomas D. Albright, *The Field of Firearms Forensics Is Flawed*, Sci. Am. (May 25, 2022), https://www.scientificamerican.com/article/the-field-of-firearms-forensics-is-flawed/;
- Dror & Scurich, *supra*;
- Tobin *et al.*, *supra*;
- Albright, *supra*;
- Alan H. Dorfman & Richard Valliant, *Inconclusives, Errors, and Error Rates in Forensic Firearms Analysis: Three Statistical Perspectives*, 5 Forensic Sci. Int'l: Synergy, at 5 (2022);
- Heike Hoffman, Alicia Carriquiry & Susan Vanderplas, *Treatment of Inconclusives in the AFTE Range of Conclusions*, 19 Law, Probability, and Risk 317 (2020);
- Kori Khan & Alicia Carriquiry, *Shining a Light on Forensic Black-Box Studies*, 10 Statistics & Pub. Pol'y (2023), https://www.tandfonline.com/doi/epdf/10.1080/2330443X.2023.2216748;
- Maria Cuellar, Susan Vanderplas, Amanda Luby & Michael Rosenblum, *Methodological Problems in Every Black-Box Study of Forensic Firearm Comparisons* (forthcoming 2024) (Exhibit 3), https://arxiv.org/abs/2403.17248; and

- Susan Vanderplas, Alicia Carriquiry, & Heike Hofmann, *Hidden Multiple Comparisons Increase Forensic Error Rates*, 121 Proc. of the Nat'l Acad. of Sci. No. 25 (2024), https://www.pnas.org/doi/full/10.1073/pnas.2401326121.

1. Existing FA/TM Studies *All* Suffer From Serious Design Flaws Which Prevent Them From Validating FA/TM As A Discipline

A valid scientific method must be: (i) repeatable, i.e., for FA/TM evidence, the examiner reaching the same conclusion when examining the same evidence; (ii) reproducible, i.e., different examiners reaching the same conclusion when examining the same evidence; and (iii) accurate, i.e., the conclusion is correct. Validation studies are intended to understand the range of conditions under which the method works as required, how well it performs, and to identify conditions under which it is likely to fail.

High-quality study design is needed to achieve these goals. Existing FA/TM studies overwhelmingly suffer from design flaws preventing them from reaching these aims and contributing toward the validation of FA/TM. Leading experts in the scientific community in study design, human cognition, statistics, and other scientific disciplines, have noted the myriad design flaws in existing FA/TM literature, including:

Fundamental Study Design. Three fundamental flaws in the design of nearly all FA/TM studies render them useless for researchers to provide meaningful, generalized results applicable to FA/TM evidence as a whole:

First, some studies have used closed sets with multiple known samples. Closed-set studies are easier because they have a "match" for every test sample in a set allowing examiners to merely look for the closest match in the set. PCAST Report, *supra,* at 108-09. "In a closed set with multiple known samples, we cannot determine how many comparisons were performed for any of the unknown samples, because examiners stop looking once a match is found," preventing an accurate calculation of the error rate. Exhibit 1 at 22. By contrast, in an open set with only one known (a "kit" style set), an examiner could perform only one comparison, making it much easier to calculate the error rate. *Id.*

Second, existing FA/TM studies do not generally report drop-out rates for participants. Research has found that if less than 5% of participants drop out, "there is little threat to the statistical validity of the study, but if more than 20% of participants drop out, the study's validity is severely compromised." Susan Vanderplas, Kori Khan, Heike Hofmann, & Alicia Carriquiry, *Firearms and Toolmark Error Rates*, University of Nebraska–Lincoln, Department of Statistics: Faculty Publications, at 4 (2022), submitted by the defense in *Illinois v. Winfield*, 15 CR 14066-01 (Exhibit 2); *see also* Khan & Carriquiry, *supra*, at 5 ("high rates of missingness [or nonresponse] preclude generalization to a broader population of examiners"). The effect of participant drop-out is unknown for existing FA/TM studies—it could be minimal, or could be significant.

Finally, while a few studies have larger examiner participation, many FA/TM studies suffer from inadequate sample size. Exhibit 3 at 10. In each of the 28 FA/TM studies reviewed, there was no sample size calculation conducted to determine how many firearms, examiners, and bullets/cartridge cases were necessary to meet the study's goals. "The sample size calculation is 'one of the most important parts of any experimental design problem.'" *Id.* (citation omitted). Failure to calculate and implement the appropriate sample size leads to an insufficient number of firearms used in the study, and, as a result, low precision, exacerbating other flaws and compounding the harm. *Id.* at 9.

Participant Sampling. For studies to be generalizable to FA/TM on the whole, participants must be a representative sample from the population at issue—for FA/TM, all qualified examiners in the U.S. Exhibit 2 at 5; *see also* Dorfman & Valliant, *Inconclusives*, *supra*, at *5*; Khan & Carriquiry, *supra*, at 8.

Existing FA/TM studies have not randomly selected participants from the population, instead relying upon self-selected volunteers. Exhibit 2 at 6; Khan & Carriquiry, *supra*, at 9. This leads to inherent biases in the study population; for example, experienced examiners who may have lower error rates than the population of examiners on the whole may be more likely to volunteer "out of a sense of duty to the discipline." Exhibit 2 at 5. In some studies, researchers "impose inclusion criteria that reasonably could be expected to be related to error

rates." Khan & Carriquiry, *supra*, at 3. For instance, the FBI-Ames study only accepted as participants "qualified examiners who were currently conducting firearm examinations, were members of the . . . AFTE, and were employed in the firearms section of an accredited public crime laboratory within the U.S. or a U.S. territory" and excluded FBI employees to avoid a conflict of interest. *Id.* FA/TM examiners qualified by courts as experts, however, often lack one or more of these requirements. *See id.* at 4 (finding that of 60 expert witness CVs reviewed, "with fewer than two of the inclusion criterion used in the FBI/Ames study, the majority of these expert witnesses would have been excluded from participation"). Without a representative participant sample, a study can speak only to the error rate of those participants, not to the discipline as a whole.

Material Sampling. Well-designed studies should also have a representative sample of the ammunition and firearms that an examiner could encounter in casework. A representative sample should be drawn "by first characterizing the full spectrum of firearms, ammunition, and examiners and then taking a random or systemic sample from these." Exhibit 3 at 15. Yet, existing FA/TM studies largely concern a single type of firearm and/or a single type of ammunition. Exhibit 2 at 6. Many existing studies examine firearms of the same make and model manufactured closely in time. *Id*; *see also* Spiegelman & Tobin, *Analysis of experiments in forensic firearms/toolmarks practice offered as support for low*

rates of practice error and claims of inferential certainty, 12 L., Probability and Risk 115, 124 (2012) (noting in one study oft-cited to courts by proponents of FA/TM, only three weapon types were used, "two with sample size 1"), 127 (noting in another oft-cited study, "one type of weapon" and "possibly two types of ammunition" were used) ("Analysis of experiments"). In other cases, the studies concern only firearm makes and models known to mark well. Exhibit 2 at 6-7. Existing FA/TM studies are thus not representative of the discipline on the whole, and findings and error rates may not generalize well to FA/TM on the whole.

Missing Data and Non-Responsive Bias. In addition to those dropping out, participants commonly fail to complete the full study, creating two potential sources of missing data. *Id.* at 7-8; *see* Dorfman & Valliant, *Inconclusives, supra*, at 5. Missing data can create several potential biases—particularly if the participants who are the source of missing data are "systematically different" from those who do complete the study. Exhibit 2 at 7. If missing data is ignored, it could contribute to an underestimate of error rates, such as if examinees do not respond to items they know are likely to be incorrect. *See* Khan & Carriquiry, *supra*, at 17. Researchers in FA/TM studies could use appropriate statistical methods to reduce the potential bias caused by missing data, but none of the existing black-box firearms examination studies have done so. Exhibit 3 at 27. FA/TM studies have not reported their rates of missing data, and it is thus

impossible to assess the magnitude of its effect on the error rates reported by those studies.

Confidence Intervals for Error Rates. In order to draw statistical inferences about the false positive error rate for firearm examiners, there must be a method to measure uncertainty, such as a confidence interval. A confidence interval puts the error rate into perspective. "A confidence interval for the false positive error rate represents the range of plausible values for it that are consistent with the study data." Exhibit 3 at 21. The FA/TM studies reviewed had either invalid or nonexistent uncertainty measures for error rates. *Id.* at 10. For instance, all FA/TM black-box studies reviewed failed to account for "variation among firearms in their likelihood of producing easier/harder to match toolmarks." *Id.* at 22. This failure leads to invalid confidence intervals. *Id.* Studies also failed to account for the study design of multiple examiners and multiple firearms, which leads to statistical dependence. *Id.* at 10. Attempting to interpret the results from FA/TM studies absent this key information on how much uncertainty should be attached to the point estimates of error rates is similar to "learning the estimated percentages from an opinion poll without knowing the margin of error." *Id.* at 25.

Inconclusives. Finally, and critically, FA/TM studies are fundamentally flawed in how they treat inconclusive results. Where an examiner cannot identify a match or definitively rule out a match, AFTE's Theory of Identification permits

the examiner to report results as "inconclusive." Dorfman & Valliant, *Inconclusives*, *supra*, at 1. Because existing FA/TM error rate studies were designed to prescreen and remove test items if they appeared to be inconclusive in nature, *no* answer of inconclusive should be deemed "correct" in these studies. Dror & Scurich, *supra*, at 336. However, most FA/TM studies count an answer of "inconclusive" as a correct answer. *See* Exhibit 3 at 17. Thus, test takers are effectively allowed to skip difficult questions—which are, by definition, more likely to yield wrong answers—by simply answering "inconclusive."[6] This method of scoring unquestionably "misrepresents the reality of evidence in casework" and inherently, artificially depresses the true error rate. Dror & Scurich, *supra*, at 334, 336 ("A priori presuming that inconclusive decisions can never be an error is problematic. If some examiners conclude an identification (or exclusion) whereas other examiners conclude as inconclusive, then at least some of the examiners are mistaken . . . [I]t is obvious [everyone] cannot all be correct when examiners reach different conclusions on identical comparisons."). Other studies simply exclude inconclusives from their analysis altogether, which also artificially deflates potential error rates and renders reported error rates

---

[6] In Ames I (defined below), more than 20% of test takers labeled *every* single different-source cartridge case comparison (the type of comparison that can produce false positives) as inconclusive. David P. Baldwin *et al.*, *A Study of Examiner Accuracy in Cartridge Case Comparisons*, 349 Forensic Sci. Int'l (2014).

uninformative. Dorfman & Valliant, *Inconclusives*, *supra*, at 3; Albright, *supra*, at 5. In either method—whether inconclusives are considered correct or are excluded—examiners may have an incentive "to opt out of the most challenging test cases by responding 'inconclusive', knowing that this will either decrease or have no impact on the study's error rate." Exhibit 3 at 18. At the extreme, these systems would allow an examiner to answer "inconclusive" on every test question and nevertheless receive a perfect score.

* * *

In sum, the existing "'validation studies'" concerning FA/TM "typically result from," among other things "statistical . . . deficiencies in the design and conduct of the experiments, and frequently lead to unjustified inferential extrapolation to universal" application to FA/TM. *Analysis of experiments in forensic firearms/toolmark*, *supra*, at 115. In other words, "[t]he various 'validation studies' may be skilled experiments as forensic proficiency tests for specific examiners (test respondents) in controlled circumstances, but *the same studies as currently exist are inappropriate for extrapolation to universal assumption or otherwise representative of rates of error for the field of*" FA/TM evidence. *Id.* (emphasis added); *see also* Exhibit 2 at 10; Exhibit 1 at 33. Existing studies thus have "conclusions [that] far exceed statistically sound inferences from the experimental evidence." Spiegelman & Tobin, *supra*, at 130; *see also id.* at

118 ("Because most of the studies reviewed by the authors stray to varying degrees from the true scientific method, they frequently contain another characteristic of 'pathological science': wishful data interpretation."); Exhibit 3 at 28 ("Our main finding is that methodological problems are pervasive and consequential, and thus the scientific validity of firearms examination has not been established"). Many of the existing studies are also insufficient to demonstrate validity or reliability because the studies do not release sufficient data for other researchers to assess the quality of the study and utility of the data. Leading statisticians (and the scientific community on the whole) have reached one inescapable conclusion: multiple *well-designed* studies are still needed to demonstrate the general scientific validity and reliability of FA/TM evidence.

2. Information Gleaned From The Better-Designed Studies Demonstrates Alarming Error Rates

Proponents of FA/TM evidence have asserted it is impossible or impractical to conduct studies establishing a discipline-wide error rate in support of the validity of FA/TM and that criticism of existing studies and FA/TM evidence is thus overblown. Not so. While it may be complex to establish a discipline-wide known or potential rate of error using well-designed empirical studies, Spiegelman & Tobin, *supra*, at 130-31, such difficulty should not prevent the FA/TM and scientific communities from attempting to do so—life and liberty are at stake. Indeed, "[u]nlike testimony that results in a determination that the perpetrator of a

crime was of a certain height range, a conclusion that a bullet found in a victim's body was fired from the defendant's gun is likely to lead much more directly to a conviction. That effect is compounded by the fact that a defendant is almost certain to lack access to the best evidence that could potentially contradict (or, of course, confirm) such testimony, which would be bullets fired from other firearms from the same production run." *Abruquah v. State*, 296 A.3d 961, 991 (Md. 2023) (citation omitted).

Contrary to its proponents' assertions, it *is* possible to design FA/TM studies with proper study design. Four existing studies have been better designed, in that they are open-set and test multiple types of firearms and substrates.[7] The results of these studies, however, do not support the validity of FA/TM evidence. They instead demonstrate why more well-designed studies are necessary. In particular, these studies demonstrate alarming error rates in FA/TM evidence—rather than establish or in any way support its validity.

Two studies undertaken by the Ames Laboratory, a Department of Energy national laboratory affiliated with Iowa State University, showed astounding error rates in FA/TM examinations.

[7] Even these studies still suffer from critical flaws, including missingness and improper accounting of inconclusive results.

In the first study ("Ames I"), as assessed by PCAST, researchers identified a positive error rate between 1 in 66 and 1 in 46, i.e., examiners made a false positive or inconclusive identification as frequently as every 1 in 46 examinations—a far cry from the near certainty testifying firearms examiners portray in their testimony. *See* PCAST Report at 11; *see also* Baldwin *et al.*, *supra*.

The results of the second Ames study ("Ames II")[8] are even more troubling. Participants concluded that the results of comparisons were inconclusive 50% of the time with respect to bullets and 40% of the time with respect to cartridge casings—a significant portion of the results. Dorfman & Valliant, *Inconclusives*, *supra*, at 6. The Ames II authors reported inconclusives as "neutral non-errors," which allowed them to report error rates of less than 0.8%. *Id.* at 2. If inconclusives are regarded as potential errors—which they logically are (*see* pp. 21-23, *supra*)—the potential error rate rises to more than ***66%***. *Id.*

Aside from the alarming potential error rate, Ames II highlighted the utter subjectivity and inability of FA/TM to demonstrate the repeatability of its methodology. With respect to bullets, examiners were unable to repeat *their own conclusions* 21% of the time for known matches and 35.3% of known non-

[8] Keith L. Monson *et al.*, *Accuracy of Comparison Decisions by Forensic Firearms Examiners*, 68 J Forensic Sci. 86 (2023).

matches; they were unable to repeat the conclusions of *other* examiners 32.2% of the time for known matches and nearly 70% of the time for known non-matches. Stanley J. Bajic *et al.*, Ames Laboratory-US DOE, Technical Report No. ISTR-5220, *Report: Validation Study of the Accuracy, Repeatability, and Reproducibility of Firearms Comparison* (2020). The cartridge casings results were equally appalling: examiners disagreed with their own conclusions 24.4% of the time for known matches and 37.8% of the time for known non-matches, and disagreed with other examiners 36.4% of the time for known matches and 59.7% of the time for known non-matches. *Id.* In total, "examiners examining the same material twice, disagree[d] with themselves between 20 and 40% of the time." Dorfman & Valliant, *Inconclusives*, *supra*, at 6; *see also* Alan H. Dorfman & Richard Valliant, *A Re-analysis of Repeatability and Reproducibility in the Ames-USDOE-FBI Study*, 9 Stats. & Pub. Pol'y 175 (2022), https://doi.org/10.1080/2330443X.2022.2120137 (Ames II showed "rather weak Repeatability and Reproducibility").

For different reasons, the third arguably better-designed study likewise cannot and does not support the validity or reliability of FA/TM analysis. That study was a small, open-set study of 40 caliber cartridge cases. Mark A. Keisler *et al.*, *Isolated Pairs Research Study*, 50 AFTE J. 56 (2018). A single, limited study of one type of spent ammunition, alone, cannot validate the entire field. Further,

that study is not without a significant design flaw: like many FA/TM studies, the author counted inconclusives as correct answers. *Id.* If the inconclusives were counted as errors or potential errors, the error rate rises to nearly 20%.

Finally, FA/TM proponents often point to Maddisen Neuman *et al.*, *Blind Testing in Firearms: Preliminary Results from a Blind Quality Control Program*, 67 Journal of Forensic Sciences 964 (2022), which yielded low error rates and similar rates of inconclusives between case work and proficiency tests as support for reliability of FA/TM field as a whole. However, Neuman et al. had a small sample size (only eleven examiners), the Houston lab at which the examiners all worked is considered a high-quality lab, and examiners were given incentives to discover which were real case samples and which were proficiency tests. Most importantly, this was not a validation study at all but rather a small quality control test to determine whether blind proficiency testing (mixing real case samples and proficiency testing samples into the evaluation with the idea that examiners did not know which was which) was a viable way to mitigate bias. Its purpose differed from validation studies, and its authors acknowledged its limitations.

C. Agent Harvey's Purported "Progressive Deterioration" Theory Is Not A Valid Scientific Method

Methodological flaws that are highly problematic in the abstract can be catastrophic in the real world. As described above (*see supra* p. 16), for a scientific method to be valid, it must be repeatable, reproducible, and accurate.

Agent Harvey's so-called "progressive deterioration theory" falls far short on each of these requirements. Agent Harvey collected over 18,000 casings from locations where Mr. Wyldes had fired a gun in the years before and after the murder, including from a range in Cedar Rapids, a farm in South Dakota, and a farm in rural Wayne County. T. 643. Agent Harvey then examined these over 18,000 casings and was not able to identify any that had been fired by the same firearm as those found at the scene or on the nearby gravel road. T. 647. Instead of concluding that they were indeed fired by different weapons, Agent Harvey concocted his own, untested, unvalidated theory termed "progressive deterioration." T. 632-72. Agent Harvey explained that, despite not being able to match any of the casings to those found at the scene or nearby, the same gun *could* have made markings that looked different, if it had a problem that was gradually worsening. *Id.*

Agent Harvey did not lay out any framework for the theory, nor did he point to any prior instances of this theory's use. Indeed, no validation studies have been conducted regarding this so-called theory, because Agent Harvey simply made it up. The theory was based entirely on subjective conclusions and unfounded assumptions and should be completely discredited.

II. IOWA SHOULD JOIN THE MOVEMENT TO LIMIT BASELESS STATISTICAL CONCLUSIONS IN FA/TM EVIDENCE

Among the chorus of critics of FA/TM are statisticians who have found that, in their present form, studies in many common pattern matching forensic disciplines such as FA/TM and fingerprint comparison, have reported statistically inappropriate conclusions. *E.g.*, American Statistical Association, *Position on Statistical Statements for Forensic Evidence*, at 2 (2019), https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf ("ASA Report"); Spiegelman & Tobin, *supra*, at 115. Opining that two pieces of ammunition were fired from the same gun, for example, "requires knowledge of how common or rare the association is, based on empirical data linked to the case at hand." ASA Report, *supra*, at 2. For instance, saying two faces are similar because they both have two eyes, a nose, and a mouth, is meaningless, because those similarities are shared by the vast majority of the population. Little, if any, such empirical data currently exists, meaning the weight of FA/TM examiners' observations is undetermined. The uncertainty surrounding FA/TM is compounded by the lack of empirical data appropriately establishing a *realistic* potential or actual rate of error. *See supra*, Part I.A; *see also* Karen Kafadar, *The Critical Role of Statistics in Demonstrating the Reliability of Expert Evidence*, 86 Fordham L. Rev. 1617, 1620 (2018) ("Scientific validity and reliability require that a method has been subjected to empirical testing . . . that provides valid estimates of how often the method

reaches an incorrect conclusion." (quoting PCAST Report, *supra*, at 143).) Statisticians attribute the paucity of such data to the poor design of existing FA/TM studies. *E.g.*, Exhibit 2 at 10.

Despite these failings, FA/TM evidence is often presented in broad—and seriously misleading—categorical strokes: there is a definite identification (i.e., a "match"), a definite exclusion, or the results are inconclusive. Dorfman & Valliant, *Inconclusives*, *supra*, at 1. Such statements inherently imply certainty—or at least a very high probability—where there is none. *See* ASA Report, *supra*, at 2; Albright, *supra*, at 9 (examiner statements "foster an illusion of certainty"). As described above, there is *no basis* in existing data to suggest that a FA/TM examiner's conclusion can be made with any degree of probability or certainty: there are no established error rates or empirical data demonstrating the relative frequencies of various characteristics observed on spent ammunition.

This Court should advance the interests of justice and hold that FA/TM evidence is inadmissible because it is unreliable. At a minimum, the Court should hold that FA/TM evidence may not be presented in such a way that implies a statistical or probabilistic basis for a conclusion where there is none.

Given the ever-growing scientific criticism of FA/TM evidence (*see* Part I, *supra*), there has been a nascent nationwide shift toward limiting the admissibility

of FA/TM evidence. Courts have curtailed *how* FA/TM examiners may present their conclusions to the jury.

In *Abruquah,* 296 A.3d at 998, the Supreme Court of Maryland held that the AFTE Theory of Identification was not a sufficiently reliable basis for the unqualified opinion of a prosecution witness that four bullets and one bullet fragment found at the crime scene were fired from the defendant's revolver.

> [B]ased on the record here, and particularly the lack of evidence that study results are reflective of actual casework, firearms identification has not been shown to reach reliable results linking a particular unknown bullet to a particular known firearm . . . In effect, there was an analytical gap between the type of opinion firearms identification can reliably support and the opinion [the expert] offered.
>
> *Id.* at 997.

Likewise, in *Williams v. United States*, 210 A.3d 734 (D.C. 2019), the Court of Appeals found that the trial court committed plain error when it allowed a FA/TM examiner's testimony that the ammunition in question "had all been fired from the same gun," and "fired from" the specific gun recovered in connection with the case. *Id.* at 738. The court unequivocally noted that "the empirical foundation does not currently exist to permit these examiners to opine with certainty that a specific bullet can be matched to a specific gun," and that "these conclusions are simply unreliable." *Id.* at 742; *see also Gardner v. United States,* 140 A.3d 1172, 1184 (D.C. 2016) (holding that "a firearms and toolmark expert may not give an unqualified opinion, or testify with absolute or 100% certainty,

that based on ballistics pattern comparison matching a fatal shot was fired from one firearm, to the exclusion of all other firearms"); *Geter v. United States*, 306 A.3d 126, 132 (D.C. 2023) (concluding that the admission of FA/TM examiner's testimony was plain error under current law and noting that the government's argument ignored "the fact that research does not exist to say that a specific bullet can be matched to a specific gun based on pattern matching") (citations omitted).

III. JURORS TEND TO PLACE GREAT WEIGHT ON SCIENTIFIC TESTIMONY AND COURTS MUST TAKE CARE TO EXCLUDE UNRELIABLE AND INVALID EVIDENCE

Because of the significant weight that juries often place on expert testimony, it is crucial that expert scientific testimony be reliable and scientifically valid. *See, e.g.*, *Daubert*, 509 U.S. at 595 ("Expert evidence can be both powerful and quite misleading because of the difficulty in evaluating it.") (citation omitted); *see also* Dawn McQuiston-Surrett & Michael J. Saks, *Communicating Opinion Evidence in the Forensic Identification Sciences: Accuracy and Impact*, 59 Hastings L. J. 1159, 1188 (2008) (noting that "most jurors begin with an exaggerated view of the nature and capabilities of forensic identification").

"Expert evidence can be both powerful and quite misleading because of the difficulty in evaluating it." Jack B. Weinstein, Rule 702 of the Federal Rules of Evidence Is Sound: It Should Not Be Amended, 138 F.R.D. 631, 632 (1991). The power of flawed forensics to mislead juries has been echoed by numerous scholars

and studies. For example, studies have found that jurors give outsized weight to forensic evidence. *See* Richard H. Underwood, *Evaluating Scientific and Forensic Evidence*, 24 Am. J. Trial Advoc. 149, 166 (2000); *see also* Tom R. Tyler, *Viewing CSI and the Threshold of Guilt: Managing Truth and Justice in Reality and Fiction*, 115 Yale L.J. 1050, 1068 (2006) ("[W]idespread evidence . . . [indicates] people already overestimate the probative value of scientific evidence."). As one study put it, "jurors in this country often accept state forensic testimony as if each prosecution expert witness is the NASA scientist who first put man on the moon." Mark A. Godsey & Marie Alao, *She Blinded Me with Science: Wrongful Convictions and the "Reverse CSI Effect,"* 17 Tex. Wesleyan L. Rev. 481, 495 (2011).

Similarly, studies show jurors struggle to understand basic scientific concepts. "[R]esearch indicates that jurors often do not understand the fundamentals of scientific evidence, and lack the 'ability to reason about statistical, probabilistic, and methodological issues effectively.'" Keith A. Findley, *Innocents at Risk: Adversary Imbalance, Forensic Science, and the Search for Truth*, 38 Seton Hall L. Rev. 893, 948 (2008). As such, this Court should consider how the new scientific consensus on FA/TM analysis—through exclusion of evidence or contrary expert testimony or cross-examination—would change the power of the

evidence the jury heard at the original trial which led to the conviction of Mr. Wyldes.

CONCLUSION

For the foregoing reasons, *amici* respectfully urge the Court to hold that, based on the current limitations described above, firearms identification testimony lacks scientific validation and is inherently unreliable.

Dated: October 1, 2024

Respectfully submitted,

/s/ Matthew Sease
Matthew Sease
Iowa Bar No. AT0010484
104 SW 4th Street, Ste. A
Des Moines, IA 50309
msease@seasewadding.com
(515) 883-2222

Donald P. Salzman*
1440 New York Avenue, N.W.
Washington, D.C. 20005
donald.salzman@probonolaw.com
(202) 371-7983

Marley Ann Brumme*
500 Boylston Street
Boston, MA 02116
marley.brumme@probonolaw.com
(617) 573-4861

Hannah Henderson*
500 Boylston Street
Boston, MA 02116
hannah.henderson@probonolaw.com
(617) 573-4878

* Pro Hac Vice Motion Filed Herewith

CERTIFICATE OF COMPLIANCE WITH TYPEFACE REQUIREMENTS AND TYPE-VOLUME LIMITATION FOR BRIEFS

This brief complies with the typeface requirements and type-volume limitation of Iowa Rs. App. P. 6.903(1)(d) and 6.903(1)(g)(1) because:

This brief has been prepared in a proportionally spaced typeface using Times New Roman in size 14 font and contains 6,473 words, excluding the parts of the brief exempted by Iowa R. App. P. 6.903(1)(g)(1).

/s/ Matthew Sease October 1, 2024

CERTIFICATE OF SERVICE

The undersigned certifies a copy of this combined certificate was served on the 1st day of October, 2024, upon the following persons and upon the clerk of the supreme court.

Erica Nichols Cook
Assistant State Public Defender
Wrongful Conviction Division
6200 Park Avenue, Suite 100
Des Moines, IA 50321
Counsel for Applicant-Appellant
via United States mail

Louis S. Sloven
Assistant Attorney General
Hoover State Office Bldg., 2nd Fl.
Des Moines, IA 50318
Counsel for Respondent-Appellee
via United States mail

/s/ Matthew Sease October 1, 2024

APPENDIX

List of Amici

Maria Cuellar
Assistant Professor of Criminology, Statistics & Data Science
University of Pennsylvania

Thomas D. Albright
Professor
Salk Institute for Biological Studies

David L. Faigman
Chancellor & Dean and
John F. Digardi Distinguished Professor of Law
University of California College of the Law, San Francisco

M. Bonner Denton
Galileo Professor of Chemistry and Biochemistry
University of Arizona

Arturo Casadevall MD, PhD
Chair, Molecular Microbiology & Immunology
Alfred & Jill Sommer Professor and Chair
Bloomberg Distinguished Professor
Professor, Department of Medicine
Johns Hopkins Bloomberg School of Public Health and School of Medicine

Dr. Susan Vanderplas
Associate Professor, Statistics Department
University of Nebraska Lincoln

David Jaros
Professor of Law and Faculty Director of the University of Baltimore School of Law
Center for Criminal Justice Reform

Brandon L. Garrett
L. Neil Williams Professor of Law,
Duke University School of Law
Faculty Director, Wilson Center for Science and Justice

Yvette Garcia Massri
Executive Director
Wilson Center for Science and Justice at Duke Law School

EXHIBIT 1

7-1-2022

Reply to Response by FBI Laboratory Filed in Illinois v. Winfield and Affidavit by Biederman et al. (2022) Filed in US v. Kaevon Sutton (2018 CF1 009709)

Susan VanderPlas
University of Nebraska-Lincoln, svanderplas2@unl.edu

Kori Khan
Iowa State University, kkhan@iastate.edu

Heike Hofmann
Iowa State University, hofmann@iastate.edu

Alicia Carriquiry
Iowa State University.

Reply to Response by FBI Laboratory filed in Illinois v. Winfield and Affidavit by Biederman et al. (2022) filed in US v. Kaevon Sutton (2018 CF1 009709)

Susan Vanderplas, Kori Khan, Heike Hofmann, Alicia Carriquiry

July 1, 2022

Table of contents

1 Preliminaries

1.1 Scope

The aim of this document is to respond to issues raised in Federal Bureau of Investigation[1] and Alex Biedermann, Bruce Budowle & Christophe Champod[2].

1.2 Conflict of Interest

We are statisticians employed at public institutions of higher education (Iowa State University and University of Nebraska, Lincoln) and have not been paid for our time or expertise when preparing either this response or the original affidavit.[3] We provide this information as a public service and as scientists and researchers in this area.

1.3 Organization

The rest of the document precedes as follows: we begin by outlining our main points of agreement with the Federal Bureau of Investigation[4] (hereafter, FBI) and Biedermann, Budowle, and Champod[5] (hereafter, BBC) in Section 2. As a threshold issue, we consider the concept of a general discipline-wide error rate in Section 3 in order to correct statistical misconceptions in Biedermann, Budowle, and Champod[6]. We then describe the statistical concepts underlying our assessment of the discipline of firearms and toolmark examiners in Section 4. Finally, we address specific issues with participant and material sampling (Section 5), study design (Section 6), and the use of inconclusives (Section 7).

[1] *FBI Laboratory Response to the Declaration Regarding Firearms and Toolmark Error Rates Filed in Illinois v. Winfield* (Aff. filed in US v Kaevon Sutton dated May 3, 2022).
[2] *Forensic feature-comparison as applied to firearms examinations: evidential value of findings and expert performance characteristics* (Aff. filed in US v Kaevon Sutton dated April 28, 2022).
[3] Susan Vanderplas et al., *Firearms and Toolmark Error Rates* (Aff. filed in Illinois v Winfield, January 2022).
[4] *Supra* note 1.
[5] *Supra* note 2.
[6] *Supra* note 2.

2 Introduction

Reading the responses submitted to our original affidavit, there are some areas of broad agreement between the anonymous individuals at the FBI, Biederman, Budowle, and Champod, and ourselves:

- There are very good firearms examiners who have a very low false-identification rate.
- Firearms and toolmark examiners are observing real phenomena - the conclusions they draw are based in observable, verifiable markings on the evidence that can provide information about the likely source of the evidence.
- There should be additional research on firearms and toolmark examination focusing on scientific foundations and error rates.

Additionally, we agree with BBC that the current studies are not useful for identifying a domain-wide error rate.

However, we are statisticians. As statisticians, we regularly help other scientists design experiments that are able to make scientifically valid claims about observable phenomena. We have experience working in situations where lives hang in the balance when errors are made: public health, nuclear engineering, and the law, among others. In these situations, it is even more important that experimental designs be as rigorous as possible, and that the conclusions from the studies be interpreted as carefully as possible, because the consequences for being wrong are so serious. It is with this mindset that we approach the topic of error rate studies in firearms and toolmark examination. We make no apologies for the fact that we offer what may seem to be harsh critiques of the state of scientific evidence in this field. Our intent in approaching the discipline in this way is constructive: until the extent of the cancer is identified, treatment cannot begin.

3 Should a Discipline-Wide Error Rate be the Goal?

A fundamental point of contention in BBC is that discipline-wide error rates are not useful or productive. This point seems to be central to their argument, despite not being a focus of our statement. Instead, they argue that the existing validation studies are valuable information regardless of whether they can be generalized to the discipline.[7]

> A domain-wide error rate is, ultimately, a practical impossibility because there is constant variation in (i) the population of examiners (new examiners enter the field, others leave; individual proficiency evolves over time), and (ii) the types of firearms and ammunition manufactured (and subsequently present in general circulation). Thus, it is always possible to argue that existing studies are somehow

[7] Biedermann, Budowle, and Champod, *supra* note 2, pgs 22-23.

> imperfect, which renders the call for a domain-wide, contemporaneously valid error rate ultimately self-defeating. (BBC, pg. 8)

The question of whether a discipline-wide error rate is useful to the court is outside our area of expertise, so we do not address this. We note instead that, it is, in fact, possible to establish valid discipline-error rates with properly designed studies, and we take a moment to address some of BBC's misconceptions about this possibility.

Statistical inference does not require a stable population of examiners or firearms. A common example used to illustrate this fact in introductory statistics courses is a scenario where a company would like to estimate the lifetime of a specific model of light bulb. The company takes a random sample of 30 light bulbs from the production line and measures how long the light bulb takes to burn out. The student is then asked to use the lifetimes of the 30 sample light bulbs to calculate an interval describing the population average lifetime with a certain level of confidence. These calculations are valid even though the company is still manufacturing new light bulbs - that is, the population is not stable.

The perception that a stable population is required to derive inferences is not the only statistical misconception demonstrated by BBC. At the heart of this further confusion are two types of statistics: descriptive statistics, and inferential statistics. **Descriptive statistics** are statistics which describe characteristics of an observed data set, such as "The average height of the men in this room is 5 feet, 9 inches." **Inferential statistics**, by contrast, are statistics which take an observed data set and generalize the information from this data set to a wider set of individuals - the population. Inferential statements might include some discussion of variability, because while the sample value is known, inference to a population involves accounting for the variability inherent in the act of taking a sample from the population. An example would be the statement "We are 95% confident that the average height of a male in the United States is between 5 feet 8.5 inches and 5 feet 9.5 inches."

None of the authors of BBC are statisticians, nevertheless, they state "statisticians' primary focus [is] on inferential statistics" (Biedermann, Budowle, and Champod[8] pg. 22). This is incorrect. There are entire areas of statistical research focused on descriptive statistics. SV and HH specialize in and conduct research on some of these areas. As trained and practicing statisticians, both inferential and descriptive statistics are firmly within all of our areas of expertise.

We assume that BBC meant to imply that we chose to focus on inferential statistics in our initial statement as a matter of preference. We did not- we focused on how validation studies are currently being used. All statements we have reviewed thus far in this case have been inferential statements. For example, Federal Bureau of Investigation[9], page 4 states "In sum, the studies demonstrate that firearm/toolmark examinations, performed by qualified examiners in accordance with the standard methodology, are reliable and enjoy a very low false positive

[8] *Id.*
[9] *Supra* note 1.

rate." A descriptive statement would have read as: "In sum, the studies demonstrated that self-selected participants enrolled in the study enjoyed a low error rate on the test sets they chose to respond to." Similarly, the FBI/Ames study (cited by Federal Bureau of Investigation[10] on page 4) makes the inferential statement "[This] study was designed to provide a representative estimate of the performance of F/T examiners who testify to their conclusions in court."[11] A descriptive statement would read: "This study was designed to provide estimates of the performance of the 173 F/T examiners who participated in the study."

The FBI and BBC cannot have their cake and eat it too– if the use of inferential statements persists, then the problems with study design continue to be a relevant issue (Section 5 and Section 6). The BBC authors argue that we are concealing useful descriptive information by pointing out that the validation studies' designs makes them inappropriate for inference. As previously stated, we made no arguments about descriptive information because no one is using validation studies for that purpose. We take a moment to highlight a few points relevant to using descriptive information in the context of error rate studies.

Descriptive information can be of varying quality. The following three statements are all descriptive statements:

- My son only answered one question incorrectly on his math test.
- My son only answered one question incorrectly on his math test, but didn't answer 30% of the questions.
- My son only answered one question incorrectly, but didn't answer 30% of the questions. The questions he skipped were frequently answered incorrectly by his peers.

In day to day life, a speaker conveying the first statement when the third is true would be considered misleading. Yet, error rate studies currently make claims resembling the first statement, despite having collected sufficient information to make at least one of the other two statements. These statements then, in turn, are conveyed to courts, including this one (see Federal Bureau of Investigation[12] at pg. 4). As this example shows, it is possible to create misleading descriptive statistics. The damage potential is much higher when such statistics are then used for inferential purposes.

With complicated data, misleading descriptive statistics can be created unintentionally. To counteract this, in most other scientific areas, honoring other researchers' requests for de-identified data (data which cannot identify an individual) is considered an essential part of good science. On December 21, 2021 we requested the FBI/Ames study data from Ames lab researchers and were told the FBI has not given Ames researchers permission to share the data. On the same date, we requested the data from the FBI contact, Keith Monson. Our requests have gone unanswered. In any case, whether because the researchers do not have the statistical

[10] *Id.*

[11] Keith L Monson, Erich D Smith & Stanley J Bajic, *Planning, design and logistics of a decision analysis study: The FBI/ames study involving forensic firearms examiners*, 4 Forensic Science International: Synergy 100221 (2022).

[12] *Supra* note 1.

sophistication to take a more nuanced look at their data, or because they do not want to share the data so that others may provide that additional nuance, we are stuck in a situation where the only solution is to describe the shortcomings of the data and studies that are available.

4 Types of Validity

As we will spend the rest of this document discussing validation studies, it is worth taking the time to discuss the different kinds of scientific validity. Different factors in the design of firearms and toolmark studies affect different types of validity. In addition, the consequences for sub-optimal experimental design, study execution, and statistical analysis are different depending on which type of validity is impacted by the sub-optimal choices.

First, let us start off with the notion of **validity** in general. Validity is a measure of how the results of research represent some facet of reality. That is, validity is a mapping between the scientific process of experimentation and analysis of results and the real world. Throughout this section, we'll consider a simple question: How does the amount of water provided influence the growth of plants as measured by the height of the seedling above the ground?

Internal validity[13] is the extent to which the variable manipulated in the experiment (the **independent variable**) can be linked to the observed effect (the **dependent variable**). In our example, the independent variable is the amount of water provided and the dependent variable is the height of the seedlings. **Internal validity** measures how well the experiment can show cause-and-effect or rule out alternate explanations for its findings (e.g. sources of systematic error or bias). Internal validity is often achieved by controlling other factors that may affect the dependent variable. For instance, in our study of water and seed growth, it would be useful to ensure that other factors affecting plant growth (fertilizer, soil quality, light availability) are as consistent as possible so that only the effect of the amount of water is seen in the results.

External validity[14] is the extent to which the experimental results can be generalized beyond the study. That is, given the results of the study, what can we say about the real world? In our example, we would like to be able to say that if our study reveals that seeds grow better when there is more water available, that this would also be true in a garden setting. External validity is always affected by the amount of experimental control we implemented (which affects internal validity) and the number of variables our experiment covers. If we are only varying the levels of water available, for instance, it would be hard for our conclusions to generalize effectively

[13] While Wikipedia is often not reliable for controversial topics, it does contain good information and examples for many statistical concepts. We link to it throughout this section because it is easily accessible, unlike the statistical textbooks which would provide more respectable citations but might require a library request. The page on internal validity contains a number of good illustrations of how internal validity is established and/or threatened by experimental design considerations. Internal validity, Wikipedia (2022), https://en.wikipedia.org/w/index.php?title=Internal_validity&oldid=1089044842 (last visited Jun 20, 2022).

[14] External validity, Wikipedia (2021), https://en.wikipedia.org/w/index.php?title=External_validity&oldid=1060911552 (last visited Jun 20, 2022).

to a garden where e.g. temperature fluctuations may also impact seed growth. When trying to ensure both internal and external validity, experimenters must experimentally manipulate many different factors, ensuring that all combinations of the factors are tested. While this is tedious but feasible in some settings, it is more difficult in other settings where we have less experimental control - for instance, we cannot *assign* sex to people for the purposes of experimentation, but we can ensure that we test individuals of both sexes. When human beings are involved in experiments as participants, external validity is partially dependent on whether our sample matches our population on various dimensions of interest: in tests of examiner error rate, for instance, we probably do not need to ensure that our sample participants' height is a match to the wider population, but we should ensure that the sample's experience is representative of the wider population of firearms and toolmark examiners.

External validity is closely related to the notion of statistical **inference**, which is the ability to make broad statements about a population represented by an experimental sample.

A subset of external validity, **construct validity**[15] is the extent to which an experiment (method, study design, analysis, etc.) measures the real-life thing of interest. For instance, if we are more broadly interested in plant health in our seedling study, we would need to establish that seedling height is a good measure of overall plant health, at least over the range of time we are studying[16]. Showing construct validity requires that there is an unbroken link between the experiment and the real-world phenomenon. Construct validity can be threatened when participants are aware they are being observed (the Hawthorne effect), when there is bias in the experimental design (intentional or unintentional), when participants are aware of researcher expectations and desires, and when there are confounding variables that are not measured or assessed in the experiment. One critique of the closed-set study design[17] is that it under-estimates the false identification rate (in addition to a complete inability to estimate the false elimination rate)[18]; this is a critique based on the study's construct validity (and as a result, its external validity).

An additional concept contained within external validity is **ecological validity**[19]: the extent to which the study's procedures, measurements, and other design variables relate to the real-world context. That is, does a study performed in a laboratory setting generalize to the

[15] Construct validity, Wikipedia (2021), https://en.wikipedia.org/w/index.php?title=Construct_validity&oldid=1060911505 (last visited Jun 20, 2022).

[16] For instance, it is possible that during the germination and initial sprouting period, plant height is a good measure of health, but that after the initial plant is established, we might need to consider e.g. plant color, number of leaves, root depth, and so on as well. If this is the case, it is important that any statements about the broader construct are careful to identify the time period for which those observations might be valid.

[17] A closed-set study is one in which every unknown to be examined corresponds to a provided known sample. In closed-set studies, examiners can rely on the closest matching known sample to make an identification, even if in a casework situation with the same unknown and known sample, the examiner would return a different result.

[18] Heike Hofmann, Susan Vanderplas & Alicia Carriquiry, *Treatment of inconclusives in the AFTE range of conclusions*, 19 Law, Probability and Risk 317–364 (2021), https://doi.org/10.1093/lpr/mgab002.

[19] Ecological validity, Wikipedia (2022), https://en.wikipedia.org/w/index.php?title=Ecological_validity&oldid=1078684982 (last visited Jun 20, 2022).

outside world? For firearms and toolmark error rate studies, experimenters must establish that the study procedures are a good representation of the process of firearms and toolmark examination in casework - if, as in some historical studies, participants evaluated a low-quality photograph of a bullet through a microscope for the study, but need to evaluate actual fired ammunition in casework, the study might potentially lack construct validity. Mock-jury studies often provide individual participants with written transcripts, but this probably does not adequately mimic the experience of sitting on a jury, listening to testimony, observing the different participants in the trial, and then deliberating in a room with other individuals to reach consensus. Experimenters performing such studies may want to follow up the written transcript study with a study involving videos of a mock trial (to assess the effect of sitting through the trial) and then perform an additional group study where participants must deliberate as if on a jury in order to demonstrate that results have good ecological validity.

Another type of validity is **statistical validity**: the extent to which the statistical calculations and tests which summarize the experiment's results are believable. Statistical validity requires that sampling procedures, measurement procedures, and the statistical calculations are all appropriate for the experimental design and for the variables under investigation. This type of validity affects both internal and external validity, because the relationship between the independent and dependent variables is determined through statistical calculations (internal validity) but the ability to make statements about the population (external validity) is also a result of statistical calculations and statistical inference.

It is worth noting that almost any experiment conducted will not have perfect internal, external, statistical, construct, and ecological validity. However, if multiple experiments have been conducted on the same basic topic, it is important to assess whether the total set of experiments collectively demonstrates each type of validity. This is what is required to produce **convergent validity**, an idea mentioned by BBC (pg. 21). As we demonstrated in our initial statement, and will demonstrate again in this response, because the validation studies which currently exist have consistent flaws, it is not possible to take the total set of validation studies and argue that they have convergent validity.

5 Participant and Material Sampling: Threats to External Validity

> One of the primary concerns with error rates provided by "well-designed" studies is that even well designed, well-executed studies cannot compensate for sampling bias in the participant pool. That is, no matter how well the experiment is laid out, if the participants are not a representative sample from the population (in this case, all qualified firearms examiners in the United States), the results of the study do not generalize to that population. (Vanderplas et al., 2022)

In our initial statement, we identified sampling bias as a threat to external validity that we could not bound numerically through statistical measures. That is, we do not have enough

information to assess whether studies conducted to date have a representative sample of firearms and toolmark examiners. The FBI and BBC both remarked upon our "pessimistic" view of e.g. treatment of participant dropout rates, claiming it was incredibly unlikely every non-response would be an error. We stated this in our initial statement. Our calculations served the intended purpose of providing an upper bound for the possible error rate. Currently, the calculation of error rates are assuming that no additional errors would have been made -which is also unlikely given the number of missing responses. This effectively calculates a lower bound for the error rate. However, unlike us, the researchers putting forth these estimates do not explicitly state their assumptions or that they have calculated a bound. As a result, casual observers (and the court) are left to assume that the error rate is the lower bound. This is misleading.

In the case of participant sampling, however, we cannot create upper and lower bounds for possible error rates. This does not mean that participant sampling concerns are not important to consider, however: biased sampling procedures are a consistent source of potential bias that affects every national validation study conducted in the US to date.

5.1 Voluntary Participation and Validity Concerns

We specifically identified that because studies use voluntary participants, the study participants are likely to differ from the wider population of firearms and toolmark examiners in important ways, but in ways that we cannot statistically quantify.

The FBI correctly identified that there is no way to compel participation from participants in research studies conducted according to current federal guidelines.

> Since 1945, many organizations have adopted codes stating that voluntary and informed consent of human subjects in research is essential. The importance of this concept has been codified in the Code of Federal Regulations, which specifically requires that researchers obtain informed consent when using human subjects (45 C.F.R. § 46.101-122). These rules are binding on all federal agencies and contractors. (FBI 5)

While we cannot speak to whether this type of participation meets the requirements set out in 45 CFR 46.103, we note some error rate studies mention that participants were compelled to participate by their employer[20]. However, we agree that there are reasons why research studies have to make do with voluntary participation.

[20] "In order to get a broad cross-section of the latent print examiner community, participation was open to practicing latent print examiners from across the fingerprint community. A total of 169 latent print examiners participated; most were self-selected volunteers, while the others were encouraged or required to participate by their employers." (Bradford T. Ulery et al., *Accuracy and reliability of forensic latent fingerprint decisions*, 108 Proceedings of the National Academy of Sciences 7733–7738 (2011), https://www.pnas.org/doi/full/10.1073/pnas.1018707108 (last visited Jun 20, 2022))

With a self-selected sample, it becomes even more critical to take steps to ensure the participants are representative of the population of interest. Interestingly, Federal Bureau of Investigation[21] mentions that clinical trials are conducted on volunteers. This comparison is not perfect[22], but the FBI's reliance on clinical trials is crucial because the sampling design in validation studies is so egregious relative to medicine (and other fields). The National Institute for Health (NIH) is our country's medical research agency. The NIH has very strict funding requirements: researchers are required to establish that their sample will be representative of the population, inclusive of minority groups, and otherwise will meet the very high bar set for experimental design and composition[23]. When working with volunteer participants, researchers use strategies like case matching, where two individuals are matched on every dimension that is feasible within the total set of volunteers and then these two individuals' performance on the drug vs. placebo is compared. In other studies, the full set of volunteers is not included in the study; instead, a demographically representative sample of the wider population is chosen from among the volunteers (within practicable constraints). Stated more broadly, medical researchers take care to ensure that the study design provides for both external and internal validity, working within the constraints of a population of volunteers. This additional care to ensure both internal and external validity is missing in FTE validation studies, which is why we raised the issue of representative samples in the first place.

Unlike medical trials, validation trials do not typically take steps to ensure the population is representative. Some studies make an effort to at least not exclude participants, such as the Ulery et al.[24] study: "In order to get a broad cross-section of the latent print examiner community, participation was open to practicing latent print examiners from across the fingerprint community."

However, many FTE studies arbitrarily adopt inclusion criteria requiring that participants be active examiners employed by a crime lab, currently conducting firearms examinations, members of AFTE, etc. For example, the FBI/Ames study cited by the FBI[25] has a number of inclusion criteria. It is not clear how the inclusion criteria were applied because the technical report[26] of the study's inclusion criteria disagrees with a peer-reviewed paper's[27] description of the inclusion requirements with the use of "and" and "or" for the listed conditions.

- "Only respondents who returned a signed consent form and were currently conducting firearm examinations and were members of AFTE, or else were employed in the firearms

[21] *Supra* note 1.

[22] Examiners control their response to the black-box studies, where most people do not have conscious control over biological responses to e.g. drugs or vaccines, and we pointed out this distinction in our original response

[23] National Institutes of Health, *Inclusion of Women and Minorities as Participants in Research Involving Human Subjects | grants.nih.gov*, NIH Grants & Funding (2022), https://grants.nih.gov/policy/inclusion/women-and-minorities.htm (last visited Jun 18, 2022).

[24] *Supra* note 20.

[25] See Federal Bureau of Investigation, *supra* note 1 page 4

[26] Stanley Bajic et al., *Validation Study of the Accuracy, Repeatability, and Reproducibility of Firearm Comparisons*, 127 (2020).

[27] Monson, Smith, and Bajic, *supra* note 11.

section of an accredited crime laboratory within the U.S. or a U.S. territory were accepted into the study."[28]

- "Participation was limited to fully qualified examiners who were currently conducting firearm examinations, were members of AFTE, and were employed in the firearms section of an accredited public crime laboratory within the U.S. or a U.S. territory."[29]

There is never any justification given for the inclusion criteria, and there is some evidence these inclusion criteria are not representative of practicing F/T examiners. For example, we collected 60 unique expert witness curriculum vitae for F/T examiners from Westlaw Edge. If we use some of the criteria listed for the FBI/Ames study in Monson, Smith, and Bajic[30] only 63% were current AFTE members, 65% were employed by a public agency, and only 38% were both current AFTE members and employed by a public agency. In other words, 62% of these examiners would have been excluded from the FBI/Ames study using less than half of the inclusion criteria defined in that study. More problematically, there is also evidence that some inclusion criteria that have been used have been associated with reduced error rates in other disciplines. For example, Heidi Eldridge, Marco De Donno & Christophe Champod[31] reports that palmar print examiners employed outside of the U.S. disproportionately account for false positives. The FBI/Ames study explicitly excludes F/T examiners employed outside of the U.S.

These sources of bias discussed in this section are subtle, and require a close reading of the study's methods section. While many scientific journals rely on peer review to identify and correct these issues, the review which takes place in trade journals such as the AFTE journal do not necessarily catch and correct issues with the description and presentation of study results. However, all journals rely on the study's authors to describe the study recruitment and selection methods clearly and in detail. This does not typically happen in validation studies.

Statistically, what is required for external validity is to argue that the sample is **representative** of the population characteristics[32]. This burden falls on the experimenters; it is up to them to make the affirmative argument that the sample is representative of the population. We have suggested that polling AFTE members might reach a set of participants who are more invested in the discipline and that individuals who have the time and/or lower caseloads to participate in studies might not be representative of the wider population of firearms examiners in part because these are things that were not addressed by study authors when describing the

[28] Bajic et al., *supra* note 26.

[29] Monson, Smith, and Bajic, *supra* note 11.

[30] *Id.*

[31] *Testing the accuracy and reliability of palmar friction ridge comparisons–a black box study*, 318 Forensic Science International 110457 (2021).

[32] Contrary to the selected quote in BBC pg. 19, we state this explicitly in our original statement. The suggestion that a full census of the population of examiners is necessary is because such a census would make it easier for researchers to make the representative argument about an individual study. The census would need to consist of demographic characteristics: training, experience, gender, education; tracking this same demographic information in the validation studies would allow researchers to compare the two sets of values and make the argument that the sample is representative of the wider population.

participant selection in the study. In order to make the argument that the sampled participants are representative, study authors need to track participation, compute demographic summaries of the sample which may be relevant (geography, age, training level, case load, professional memberships), and compare these to the wider population. To support this, it might be helpful if accrediting organizations maintained a register of people who have certification in each discipline to assist with having some statistics of the population to compare against.

5.1.1 Statistical Language and Logic

Both the FBI and BBC raised the issue of hypothetical language which was used in our initial affidavit, reproduced here to provide context.

> there are many potential lurking covariates that would meaningfully affect the error rates estimated by the studies. For instance, it is possible that experienced examiners are more likely to volunteer to participate in these studies out of a sense of duty to the discipline: these examiners might have lower error rates due to their experience, which would lead to an estimated error rate that is lower than the error rate of the general population of all firearms examiners (including those who are inexperienced). In fact, in studies which differentiate between trainee and qualified examiners, we find a higher error rate among trainees (Duez et al. 2018). (Vanderplas et al., 2022, pg. 5)

> There are many variables which might be expected to increase likelihood of volunteering for a study and also change the expected error rate: education, experience, confidence, amount of time available for study participation. (Vanderplas et al., 2022, pg. 5)

BBC specifically called out these statements:

> This critique is a rhetorically subtle formulation because it uses a true statement (here: higher error rate among trainees) to create a doubt for which no direct evidence is provided. That is, Vanderplas et al. (2022) give no evidence for whether experienced examiners are actually more inclined to participate than less experienced examiners. (BBC pg. 25)

And the FBI also responded:

> The Statement fails to cite any evidence to support this claim. In fact, less experienced examiners were commonly represented as participants in numerous studies. Several studies listed in Table 1 have queried the experience level of participant examiners, and those analyses concluded that experience level did not significantly affect performance. If sampling bias had affected the outcome of one or more of these studies, one would expect the rate of reported false positives to vary considerably. (FBI pg. 7)

It should be noted that the rhetorical device employed in our original statement is common in statistics; it is not intended to mislead. However, it does make the implicit assumption that the reader is familiar with scientific logic. The presence of a confounding variable (a variable whose effect on the response cannot be separated from the explanatory variable) is sufficient to remove our ability to make a causal statement about the association between two variables (e.g. the explanatory variable causes the change in the response variable)[33]. Thus, statisticians acknowledge the presence of a confounding (or "lurking") variable (in this case, an examiner's experience, duty, education, confidence, and available time) that might co-vary with the dependent variable (in this case, the likelihood that an examiner self-selects into a study). These statements are almost always hypothetical because the presence of such a variable precludes decisive statements[34]. In this case, the presence of such lurking variables without the ability to compare the volunteer sample's demographics to the wider demographics of the population makes it logically difficult to argue that results from a self-selected sample can be generalized to the population.

In addition, we have asked for the information which would allow us to make these hypothetical statements more concrete by applying statistical techniques for correcting estimates affected by drop-out rates. Unfortunately, our requests have been rebuffed: it is common for forensic scientists to decline or ignore requests to share study data with other researchers. This is contrary to the widespread understanding of the requirements of ethical science[35] as well as the norms for research practice in many other disciplines (even disciplines which collect human-subjects data subject to federal protection). As statisticians, we commonly post our (de-identified) data on sites such as GitHub or FigShare for archival purposes as well as to enable other researchers to access the data, statistical computations, and manuscript preparation records[36].

An additional point of contention here is that the FBI states that "those analyses concluded that experience level did not significantly affect performance". The FBI is overstating their

[33]Section 4.1 Summary, Nathan Tintle et al., Introduction to statistical investigations (2015).

[34]One easy example of a lurking variable is that the number of baby births are correlated with the number of storks in European countries. It would be relatively easy to falsely draw the conclusion that storks are associated with babies, but this ignores the lurking variable of the geographic size (and population size) of the country. Causation cannot be inferred when there are lurking variables or when the study is observational in nature. Alex Mayyasi, *Do Storks Deliver Babies?*, Priceonomics (2014), https://priceonomics.com/do-storks-deliver-babies/ (last visited Jun 23, 2022).

[35]Howard Bauchner, Robert M. Golub & Phil B. Fontanarosa, *Data Sharing: An Ethical and Scientific Imperative*, 315 JAMA 1238–1240 (2016), https://doi.org/10.1001/jama.2016.2420 (last visited Jun 23, 2022); Clifford S. Duke & John H. Porter, *The Ethics of Data Sharing and Reuse in Biology*, 63 BioScience 483–489 (2013), https://doi.org/10.1525/bio.2013.63.6.10 (last visited Jun 23, 2022); Michael W. Ross, Martin Y. Iguchi & Sangeeta Panicker, *Ethical aspects of data sharing and research participant protections*, 73 American Psychologist 138–145 (2018); Carol Tenopir et al., *Data Sharing by Scientists: Practices and Perceptions*, 6 PLOS ONE e21101 (2011), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021101 (last visited Jun 23, 2022).

[36]One example of this is the GitHub repository for our paper on inconclusives in the AFTE range of conclusions, available at https://github.com/heike/inconclusives. All of the data and code are available for anyone to access, in addition to the full set of edits to the manuscript draft over time.

claim here, as well as selecting only studies which support their conclusion. Chapnick et al. (2021)[37] found that error rates for trainees were higher than those for qualified examiners. Baldwin (2014)[38] explicitly did not examine trainee examiners:

> Although it might be desirable to understand how non-practicing or untrained participants might perform under the same circumstances as trained examiners, there are important statistical reasons for not including trainees. The expected rates of error are low enough that dividing our participant pool into subgroups that are trained and not trained would add cost to the study without adding enough participants to allow a precise measurement of error rates for this group of trainees. It was deemed more important to measure the error rates for trained practicing examiners accurately and precisely than to measure the effect of another variable with much less precision and accuracy. (Baldwin 2014, pg. 7)

The only other mention of experience in Baldwin (2014) involves a finding of a weak correlation between the number of inconclusive calls and years of training:

> There are mild inverse correlations between the number of inconclusive/nonresponse calls made with the known different-source cases, and the reported number of years of training (correlation = -0.1393) and number of years of experience (correlation = -0.1034); that is, there is a weak tendency for examiners with more training or experience to make fewer inconclusive calls. (Baldwin 2014, pg. 16)

This is not a conclusion that experience does not affect error rates; while the findings reported here are not evaluated for statistical significance, and may not rise to meet that bar, they do explicitly highlight the possibility that experience is associated with an examiner's rate of reporting inconclusive results. In addition, there is no statistical test of whether error rates are related to experience anywhere else in the Baldwin paper.

Finally, the FBI's final response to our hypothetical, "If sampling bias had affected the outcome of one or more of these studies, one would expect the rate of reported false positives to vary considerably.", is false. This statement likely stems from a misunderstanding of the difference between random error and bias. Sampling error is the error in an estimate due to the difference between one sample and the next - that is, who is and is not included in the study - due to random sampling. Random sampling ensures that over many different samples, we still produce **unbiased** estimates because the sampling method itself is not biased. The problem is that when the sampling method itself is biased (and, in many cases, biased in the same structural way), we have no statistical guarantees that the resulting estimates are similarly unbiased. In fact, we have reason to suspect that the structural biases might be similar across different

[37] Chad Chapnick et al., *Results of the 3D Virtual Comparison Microscopy Error Rate (VCMER) Study for firearm forensics*, 66 Journal of Forensic Sciences 557–570 (2021), https://onlinelibrary.wiley.com/doi/abs/10.1111/1556-4029.14602 (last visited Dec 6, 2021).

[38] David P. Baldwin et al., *A Study of False-Positive and False-Negative Error Rates in Cartridge Case Comparisons:*, (2014), http://www.dtic.mil/docs/citations/ADA611807 (last visited Jan 29, 2020).

studies because the sampling bias is of the same type in each study, which might well lead to a bias in one direction for the collective set of studies.

5.1.2 Assessment of Significance

While participant selection and inclusion bias is one of the biggest issues we identified, in that we cannot easily bound the effect it has on error rates, it is by no means the only issue with existing FTE studies. If the only issue with the studies that are typically cited in court in support of firearms and toolmark analysis as a discipline were that it included self-selected volunteers who may meaningfully differ from the population, then it would be reasonable to interpret the results of these studies with that caveat in mind. However, the situation as it currently stands is one of a rowboat: if there is only one small hole in the rowboat, the boat can stay afloat while its occupants bail it out; if there are many holes in the rowboat of varying sizes, it is much more likely that the boat will sink. So it is with the error rates from these studies: there are many flaws in the studies, and while we can bound the effect on the error rates for some flaws, the overall effect is that the studies are sinking.

5.2 Material Sampling

In our original statement, we argue that as with examiners, we need to be able to make the claim that firearms studies cover a representative set of ammunition and firearm combinations in order to suggest that such studies are broadly generalizable. We are not the first group of statisticians to highlight this issue: the problem is mentioned in the 2009 NRC report[39], follow-up experiments have been proposed for several different previously published studies[40], and the 2016 PCAST report[41] described the necessary characteristics for studies establishing foundational validity, including

> "The studies must involve a sufficiently large number of examiners and must be based on sufficiently large collections of known and representative samples from relevant populations to reflect the range of features or combinations of features that will occur in the application." (PCAST pg. 52)

The FBI response misstates our position as much more extreme than the reality:

[39] Strengthening forensic science in the United States: A path forward, (National Research Council (U.S.) ed., 2009).

[40] C. Spiegelman & W. A. Tobin, *Analysis of experiments in forensic firearms/toolmarks practice offered as support for low rates of practice error and claims of inferential certainty*, 12 Law, Probability and Risk 115–133 (2013), https://academic.oup.com/lpr/article-lookup/doi/10.1093/lpr/mgs028 (last visited Oct 23, 2018).

[41] President's Council of Advisors on Science and Technology, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature Comparison Methods*, (2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf (last visited Mar 7, 2019).

> The Statement claims that existing firearms error rate studies cannot reflect an accurate error rate because they fail to encompass the full range of firearms and ammunition available to the public and are thus not representative of samples encountered during casework. (FBI pg. 8)

Instead, they argue that it would be better to focus on manufacturing methods:

> No single study (or even numerous studies) can fully capture all firearms and ammunition that currently exist in the United States. However, the more relevant variable to study is the manufacturing processes used to create these firearms that impart the class and individual characteristics analyzed during an examination. (FBI pg. 8)

We largely concur, despite the attempts to paint our position as so extreme as to require that studies exist for all combinations of firearms and ammunition which currently exist in the United States. However, it is important for those conducting such studies to identify the manufacturing method and to list the types of weapons a study might be reasonably applied to on the basis of similar manufacturing. That is, the authors of a study should be responsible for outlining the reasonable scope of generalization for a study, and this should be explicitly stated in the discussion of the study's results.

BBC have similar objections to our desire to see better combinatorial studies, but for different reasons.

> "In Section 5 of Vanderplas et al. (2022, at pp. 6–7), the authors mention that existing studies cover only a limited number of firearms and ammunition types, thus preventing the possibility to generalize. . . " (BBC pg. 25-26; additional quotes from our affidavit are provided)

In their response, they highlight their insistence that there is no average examiner and no average combination of firearm and ammunition.

> "Second, as much as there is no"average" examiner, there is no "average" combination of firearm and ammunition. Instead, there are many firearm and ammunition categories (or types) for which a single average error rate could not meaningfully reflect examiner performance. It would be a too optimistic figure for reputedly difficult firearm and ammunition types, and too conservative one for less challenging comparison pairs. However, it would be exaggerated to require that an expert has previously seen (i.e., worked with) all possible combinations of firearms and ammunition." (BBC pg. 26)

The critique of our position by BBC represents a fundamental misunderstanding as to why we want to see a broad set of studies on manufacturing methods and ammunition types: it is not to determine an average combination of firearm and ammunition, or an average error rate, or to ensure that experts have worked with all possible combinations of firearms and ammunition.

Statistics tends to be only peripherally concerned with averages: instead, we study variability and its effect on the different estimates we compute. When we indicate that part of the scientific foundation for firearm and toolmark studies is that we understand the ways in which marks might vary based on firearm manufacturing method and/or type of ammunition, it is because we want to be able to assess the external validity of the error rate studies across the wide range of conditions found in case work. While we addressed the issue of a general discipline-wide error rate as being within the range of statistics in an earlier section, this further illustrates the misconceptions that BBC have about the use of statistics. It is precisely because of the variability in difficult firearm and ammunition types vs. less challenging comparison types that we need broad studies: we recognize that variability and want to scientifically establish the consequences for error rates.

If we return momentarily to the hypothetical plant study we proposed in the validity section, we are essentially arguing that it is important to understand not only how plant growth changes with watering, but to ensure that those same findings hold across different temperature ranges and soil types commonly encountered in spring gardens. Without systematic manipulation of those variables across some studies that rely on the same principles of plant biology and development, we cannot ensure that our study's findings generalize well to new conditions.

Just as we want to ensure that validation studies can be generalized to the population of examiners and do not contain systematic biases that might over- or under-estimate the error rate of firearms and toolmark comparisons, we also want ensure that error rate studies are conducted on types of firearms (or manufacturing methods) and ammunition which are likely to be compared in casework. That is, our concerns about firearm manufacture and ammunition materials boil down to concerns about the *external validity* of error rate studies. At the risk of making another hypothetical statement, if error rate studies are conducted on combinations of firearms and ammunition which are known to mark well[42], then there is a risk that the error rate studies under-estimate the error rates which might be encountered in casework, where not all combinations of ammunition and manufacturing method are idealized. When there has not been any systematic attempt to assess the impact of these factors on error rates or on the visual information available to examiners that would be expected to influence error rates, this issue of external validity remains unresolved.

[42]We are not experts on the intricacies of different types of ammunition, but it is well known (and oft referenced in scientific publications in the field) that some types of ammunition do not "mark" well due to coatings or other material treatments of the ammunition surface. Examples of studies which investigate or discuss the phenomena of "marking well" include Nicole Groshon, *The effects of: Lacquered ammunition on the toolmark transfer process*, 2020, https://indigo.uic.edu/articles/thesis/The_Effects_of_Lacquered_Ammunition_on_the_Toolmark_Transfer_Process/13475034/files/25862940.pdf (last visited Jun 20, 2022);Valentina Manzalini et al., *The effect of composition and morphological features on the striation of .22LR ammunition*, 296 Forensic Science International 9–14 (2019), https://www.sciencedirect.com/science/article/pii/S0379073818310624 (last visited Jun 20, 2022); Deion P Christophe, *Approaching Objectivity in Firearms Identification: Utilizing IBIS BULLETTRAX-3D's Sensor Capturing Technology*, 2011, https://shareok.org/bitstream/handle/11244/324663/ChristopheDP2011.pdf?sequence=1 (last visited Jun 20, 2022).

It is also worth noting that we are not the first statisticians to suggest thorough study of the discipline is necessary, nor the first to be accused of making impossible requests.

> "without understanding the proper design of experiments, modelling and sampling procedures, numerous articles in the firearms/toolmarks domain literature assert, and several judges have mistakenly observed or implied, that assessing rates of examiner error are impossible because every firearm ever made cannot be tested."[43]

There are multiple means by which such external validity might be achieved, lest we be accused of failing to offer constructive solutions to the problems we have identified[44]. First, of course, would be to conduct error rate studies that consider ammunition and/or weapon type as a variable of interest and manipulate that variable as part of the experimental design, then test whether error rates are different across different types of ammunition and manufacturing methods. This would be the most direct way to address this premise, because error rates would be directly tied to the manufacturing method and ammunition type. Unfortunately, most validation studies cover only one design and one or two types of ammunition, as shown in Table 1[45]. Those studies which are conducted over multiple types of ammunition and/or firearms do not break down responses by firearm and ammunition type.[46]

Another way to address this premise would be to conduct several studies assessing the number, quality, and/or variety of individual or accidental markings suitable for comparison across multiple types of ammunition and/or manufacturing methods. This method would not specifically address error rates, but it would be reasonable to argue that if the type and quantity of individual markings suitable for comparison was similar across ammunition and/or manufacturing methods that the error rates for such comparisons should also be similar because the fundamental information available to the examiners would be expected to be similar. Note that this requires an additional degree of abstraction (ammunition/manufacturing -> markings -> error rates), but that the scientific logic still holds, even if the connection is more tenuous. An additional complication with this option is that we are not aware of an objective method

[43] Spiegelman and Tobin, *supra* note 48.

[44] As in BBC, pg. 8, "The dismissive attitude towards existing error rate studies, i.e., their wholesale rejection, is not helpful in that it offers no constructive advice on how the data could be used with properly acknowledged limitations."

[45] One type of ammunition and one primary type of weapon (with several known non match comparison weapons of similar manufacture) in Jaimie A Smith, *Beretta barrel fired bullet validation study*, 66 Journal of forensic sciences 547–556 (2021); one type of firearm and one type of ammunition in Baldwin et al., *supra* note 46; one type of firearm and two types of ammunition in Alfred Biasotti, *A statistical study of the individual characteristics of fired bullets*, 4 Journal of Forensic Sciences 34 (1959); one type of firearm and one type of ammunition in James E. Hamby et al., *A Worldwide Study of Bullets Fired From 10 Consecutively Rifled 9MM RUGER Pistol Barrels—Analysis of Examiner Error Rate*, 64 Journal of Forensic Sciences 551–557 (2019), https://onlinelibrary.wiley.com/doi/abs/10.1111/1556-4029.13916 (last visited Jan 29, 2020).

[46] Tasha P. Smith, G. Andrew Smith & Jeffrey B. Snipes, *A Validation Study of Bullet and Cartridge Case Comparisons Using Samples Representative of Actual Casework*, 61 Journal of Forensic Sciences 939–946 (2016), https://onlinelibrary.wiley.com/doi/abs/10.1111/1556-4029.13093 (last visited Dec 12, 2021); Keisler, M. A., Hartman, S. & Kil, A., *Isolated Pairs Research Study*, 50 AFTE Journal 56–58 (2018).

for assessing the quantity of accidental information present in a fired cartridge case or bullet, nor for assessing how much individualizing information is necessary to make an informed comparison. Introducing an additional degree of subjective assessment for the marking quality would introduce additional variability that may mask the coupling between the ammunition and firearm combination and the error rates in black-box studies. However, there are certainly exploratory studies which assess the quality of markings for different types of ammunition in a specific firearm.[47] It might also be reasonable to assess the quantity of individual characteristics using an automatic system, such as NIBIN or IBIS[48] and make the argument that if a computer system can make the distinction it is reasonable for a human examiner to do so as well[49].

5.3 Consecutive Manufacturing

Another concern we originally raised in our affidavit was that of the use of consecutively manufactured firearms for error-rate studies.

> Several studies used consecutively manufactured barrels and/or slides to increase the difficulty of the comparisons, since these types of samples create the greatest potential to produce toolmark patterns and/or subclass characteristics that are similar in appearance although produced from two different sources. (FBI pg. 3)

Our concern is one of external validity. We agree that consecutively manufactured barrels may provide a higher degree of challenge in some circumstances, but this additional difficulty comes with a cost: it is harder to generalize results to the broad class of firearms of X type when you have only tested e.g. 10 consecutively manufactured barrels. Instead, the results of such a study can only be generalized to a specific point in time. This is one facet of an oft-discussed tradeoff in experimental design: you can increase experimental control, randomize subjects to treatment conditions, and take other precautions to ensure that your experiment is providing the answer to your experimental question (internal validity), but many of these control measures reduce the ability to generalize the results to wider settings because the experimental control doesn't mirror natural conditions.[50] This paradox is also mentioned by Spiegalman & Tobin[51] in their 2013 assessment of the state of firearms validation and error rate studies.

[47] Manzalini et al., *supra* note 50; Groshon, *supra* note 50; Brian Mayland & Caryn Tucker, *Validation of Obturation Marks in Consecutively Reamed Chambers*, 44 AFTE Journal 167–169 (2012), https://afte.org/uploads/documents/paid_for_download_products/44_2_2012_Spring.pdf (last visited Jun 27, 2022).

[48] Jan De Kinder, Frederic Tulleners & Hugues Thiebaut, *Reference ballistic imaging database performance*, 140 Forensic Science International 207–215 (2004), https://www.sciencedirect.com/science/article/pii/S0379073803005371 (last visited Jun 20, 2022); Christophe, *supra* note 50.

[49] Of course, it is much easier to test a computer algorithm's ability to make these comparisons, with the added benefit that such algorithms do not usually provide inconclusive decisions.

[50] Donald T. Campbell, *Factors relevant to the validity of experiments in social settings.*, 54 Psychological Bulletin 297–312 (1957), http://doi.apa.org/getdoi.cfm?doi=10.1037/h0040950 (last visited Jun 19, 2022).

[51] Spiegelman and Tobin, *supra* note 48.

Table 1: Firearms studies listed by the FBI along with gun manufacturer and ammunition type, where specified. Note that studies using the same weapons have been grouped together, deviating from the otherwise chronological ordering. Proficiency tests with various firearms AND bullets (e.g. not systematically manipulated) were excluded from this table.

Study	Year	Type	Consec	Gun	Ammo
Brundage	1998	Bullet	Yes	Ruger P85 9mm	Winchester
J. Hamby et al.	2019	Bullet	Yes	Ruger P85 9mm	Winchester
Bunch & Murphy	2003	Cartridge	Yes	Glock Luger 9mm	Unspecified
DeFrance & Van Arsdale	2003	Bullet	Yes	Smith & Wesson .357 Magnum	Unspecified 158 grain jacketed soft-point
E. Smith	2005	Both	No	Ruger P89	Remington UMC 115 grain, copper-jacketed
Lyons	2009	Extractor	Yes	Colt 1911 A1, Caspian Arms Extractors	Speer Lawman .45 Auto 230 grain FMJ
Fadul	2011	Bullet	Yes	Glock EBIS	Federal 9mm
Mayland & Tucker	2012	Chamber	Yes	Kel-Tec, Hi-Point, Ruger	Winchester, Remington, Federal 9mm Luger 115 grain FMJ
Fadul et al.	2012	Slides	Yes	Ruger	Unspecified 9mm
Cazes & Goudeau	2013	Slides	Yes	Hi-Point 9mm C-9	Winchester 9mm Luger 115 grain FMJ
Stroman	2014	Cartridge	No	Smith & Wesson	Independence .40 S&W, 180 grain FMJ
Baldwin et al. (aka Ames I)	2014	Cartridge	No	Ruger	Remington 115 grain FMJ
Smith et al.	2016	Both	No	Taurus, Sig Sauer, Glock	92 UMC CC, 92 UMC Bu, 92 WIN BEB CC, 92 WIN BEB Bu, 92 Hi-Shok/Hydra-shok Bu, 92 American Eagle CC, 92 Speer GD CC, 92 Speer GD Bu
Duez et al.	2018	Cartridge	No	Colt Ruger P95 DC Taurus PT 24/7	PMC
Keisler et al.	2018	Cartridge	No	Glock 22,23,27 HK USP Compact S&W 40V, 40VE	CCI 40 S&W 180-grain gold dot
Kerkhoff et al.	2018	Cartridge	No	Glock (x39) Sig Sauer (x1)	Various
J. Smith	2021	Bullet	Yes	Beretta	Federal 9mm FMJ
C. Chapnick et al.	2021	Cartridge	No	Various 9mm Luger, 40 S&W, and 45 Auto	Unspecified
Law & Morris	2021	Cartridge	No	Various 9mm Luger	Federal American Eagle 124 grain FMJ
Bajic et al. (aka Ames II)	2021	Bullet	Some	Ruger, Beretta	Wolf Polyformance 9mm Luger 115 grain FMJ
Bajic et al. (aka Ames II)	2021	Cartridge	Some	Jimenez, Beretta	Wolf Polyformance 9mm Luger 115 grain FMJ

As not all studies conducted use consecutively manufactured firearms, this is one of the less critical threats to external validity. Its inclusion here serves primarily to highlight the difference between the statistical concept of good experimental design and that of firearms and toolmark examiners, whose gaze is much more narrowly focused on the process of toolmark creation.

6 Study Design: Threats to Internal and External Validity

Our concerns about the design of firearms and toolmark error rate studies are also related to concerns about validity, but study design impacts both internal validity and external validity. Before we discuss the nuances of experimental design and appropriate, scientifically supported conclusions, we want to quickly address some broad claims about the importance of good experimental design in validation studies.

The FBI maintains that the various study designs which have been conducted since Daubert (which is a much earlier time than we expected given the sea change that has occurred in forensics since the 2009 National Research Council and 2016 PCAST reports) provide meaningful ways to assess examiners' abilities.

> Since the *Daubert* decision in 1993, there have been 25 firearm/toolmark error rate studies conducted. They include black box studies with open set designs, studies with partially open set designs, and closed set study designs. These various experimental designs have provided meaningful ways to assess the ability of examiners to make accurate source conclusions. (FBI 2)

While we will not entirely discount the idea that there may be some amount of usable data in some of the poorly designed studies, we do feel that it is important to state in strong terms that the design flaws in many of these studies are significant enough to threaten the study's external validity. This means that they are **not meaningful for assessing the broad capability of FTEs to make accurate source conclusions**.

6.1 Closed and Open Set Studies

Study designs which are closed-set and involve multiple knowns threaten internal validity, as the study design is such that it does not allow us to estimate the number of comparisons performed by the examiner (and thus, an overall error rate cannot be calculated). In addition, these designs introduce constraints that allow conclusions based on factors unrelated to the firing process. As a result, closed-set, multiple-known studies produce a biased error rate that reflects other factors in addition to the examiners' proficiency in making evidence based conclusions.

> The community of researchers and practitioners appears to have taken this concern to heart. In a recent review of selected studies between 1998 and 2021, Monson et al. (2022, pg. 2) find that the closed set design is mainly used in studies prior to

> the publication of the PCAST Report (seven out of twelve summarized pre-PCAST studies). In turn, only two of six post-PCAST studies summarized by Monson et al. (2022, pg. 2) use the closed set design. (BBC pg. 24)

We acknowledge that most studies conducted since the PCAST report have used improved designs, however, we still feel the need to emphasize the issues involved in closed-set designs because some expert witnesses (and the FBI) still cite these studies and argue that they are useful when estimating examiner error rates.

The FBI uses the term a "partially open set study" to indicate a study with multiple knowns and one unknown.

> A "partially open" test design is an inter-comparison design where there are some unknowns having no matching pair. (FBI pg. 2)

This is what we would call an open-set design; the FBI is conflating two different experimental design considerations: whether or not every unknown sample has a known in the set, and whether there are multiple knowns included in the set. This distinction is important, because it speaks to how we derive the number of comparisons made by the examiner:

- In an open set with multiple known samples, if there is a match between the unknown and one of the knowns (which is not guaranteed), the examiner does not have to examine the correspondence between the unknown and any remaining, unexamined knowns. This means that we do not know how many elimination comparisons were completed by the examiner. If there is no match between the unknown and any of the knowns, then we can assume the examiner compared the unknown to all of the known samples. We can arrive at an upper bound and a lower bound for the number of comparisons performed, but we cannot precisely estimate the overall error rate, the sensitivity, or the specificity.
- In a closed set with multiple known samples, we cannot determine how many comparisons were performed for any of the unknown samples, because examiners stop looking once a match is found. Because examiners tend to assume that studies are closed-set even when not directly told that this is the case, it is possible to use logical deduction to reduce the potential for error in these studies.
- In an open set with only one known (a "kit" style set), we know that the examiner could only perform one comparison. These studies make the calculation of the error rate much easier by removing any statistical guesswork and/or ambiguity from the error rate calculation process.
- No one has attempted (nor should attempt) a closed-set study with only one known, because this would be reductive to the point of providing no information.

The number of comparisons made by the examiner is essential when calculating the error rate for the study, since the total comparisons is the denominator of that ratio. The unfortunate term "partially open" suggests that the FBI does not fully understand that the open-set issue

is only part of the design problem; the inter-comparison designs which include multiple knowns are in fact a large issue as well.

Fundamentally, the problem with closed-set studies is that they under-estimate the false elimination rate (because examiners know that the unknown matches one of the knowns) and also under-estimate the rate at which examiners provide inconclusive decisions. This is a threat to the internal validity of the study (in that error rates cannot be calculated properly) and the external validity of the study (because information is present in the test which is not present in case work). The problem with inter-comparison designs (designs with multiple knowns) is that they threaten the internal validity of the study, because we cannot calculate the number of comparisons completed by the examiner.

6.2 Human-in-the-Loop Study Design and Analysis

One argument put forth by BBC suggests that we cannot validate tests which require subjective human judgement in the same way as chemical and medical laboratory tests are validated. This is fundamentally wrong.

> In such validation studies, many test items with known ground truth status are processed and the number of correct and incorrect responses are recorded, leading to standard performance metrics such as sensitivity and specificity. Results of such validation studies can then serve as an indication of the performance with which a test can be expected to operate when applied by consumers (assuming, again, they properly operate the test). Consequently, there can be discussion about whether the performance characteristics of a candidate test are "good enough" to be deployed in a particular context of application. (BBC pg. 16-17)

> Arguably, there is no generic and human-independent performance measure for feature comparison in forensic firearm examination, akin to performance characteristics used for traditional laboratory testing procedures. (BBC pg. 18)

First, there is nothing in the description of validation studies in general which would seem to not apply to firearms and toolmark examination, other than the idea that a consumer is the one operating the test. If we consider a "standard" chemical test such as a home pregnancy test, the examiner is analogous in this case to the test strip (which is a slightly dehumanizing comparison, but we will work within the analogy set up by BBC). The goal of any entity regulating the use of such tests, whether the court or the FDA, would be to determine whether the test is reliable in discriminating between the possible states of nature the test is designed to discriminate between: pregnant or not pregnant, same source or different source. If there is variability in the test's performance under different circumstances (or different examiners), then it is important to know that at the outset, before the test is approved for general use - that variability will factor into the overall error rate, leading to a range of possible error rates (which is something that statistical calculations are designed to handle: after all, statistics

is the study of variability). So while we agree that there is variability in the performance of different examiners, we do not agree that it is useless to consider a discipline-wide error rate for the comparison of different types of marks on the basis that there is additional variability due to the human "in-the-loop". We would expect that impression-based marks would potentially need to be considered separately from striation-based marks (because the necessary features for comparison are very different), but unlike BBC, we do not consider a general summary statistic about the error rate of evaluating one type of marks to be a useless measure. In fact, their insistence that discipline-wide error rates are useless is at odds with a number of statements from researchers in the discipline that are found in error rate studies as well as in reports such as those issued by PCAST and NAS. The error rate of a technique is at the heart of any scientific evaluation of that technique.

Even if we concede that the human-in-the-loop nature of firearms examination makes it unlike validation of a chemical test, that does not mean that error rates are invalid or that studying the performance of humans in a general sense is not important. Many medical imaging procedures also require a human to make a qualitative and even binary decisions (cancer or benign lump? appendicitis or not?) that include the presence of inconclusive results (when, e.g. the appendix cannot be identified on a scan of the abdomen).[52] The medical community still actively studies the error rates of these diagnostics and the performance of the human examiners, calculates discipline-wide error rates and diagnostic utility rates (including inconclusives as negative outcomes), and is actively investigating the clinical use of algorithms that support human decision-making.[53]

There are other ways in which comparing pattern forensics black-box studies to medical studies is useful. Like FTE studies, medical studies are typically conducted on volunteers, however, there are significant differences in the statistical and scientific rigor in medical studies that are worth examining:

- Medical studies take great pains to ensure that the volunteers selected for a study are demographically representative of the population[54].
- There are strict guidelines for preregistration of study designs.[55]
- Study results for preregistered designs must be reported even if the conclusions from the study are not statistically significant. This requirement is intended to combat the "file drawer problem", an area of potential bias that we did not even start to address in our

[52] Jacob L. Jaremko et al., *Incidence and Significance of Inconclusive Results in Ultrasound for Appendicitis in Children and Teenagers*, 62 Canadian Association of Radiologists Journal 197–202 (2011), https://doi.org/10.1016/j.carj.2010.03.009 (last visited Jun 9, 2022).

[53] Nan Wu et al., *Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening*, 39 IEEE transactions on medical imaging 1184–1194 (2020).

[54] NIH funding guidelines now require that studies proposed ensure inclusion of women and minorities in proportions that allow generalization to the relevant population under investigation (National Institutes of Health, *supra* note 31).

[55] David T. Mellor & Brian A. Nosek, *Easy preregistration will benefit any research*, 2 Nature Human Behaviour 98–98 (2018), https://www.nature.com/articles/s41562-018-0294-7 (last visited Jun 18, 2022).

initial affidavit. The file drawer problem is a well known phenomena in many other areas of science, however, and it is reasonable to expect that forensic science is not exempt.
- Study results are reported and analyzed accounting for participant drop-out biases
- Collected data (in anonymized form) are published along with the study so that other scientists can repeat the analysis for themselves.

What is remarkable about the comparison to medical studies is that none of the conventions for appropriate scientific rigor in medicine are observed in studies of firearms and toolmark examiner error rates. Granted, study preregistration is not a convention observed in all disciplines, but if we accept the analogy to clinical trials because of the serious consequences of the results, it stands to reason that validation studies should be observing this level of experimental and scientific rigor.

6.2.1 The Use of Objective Assessment Tools

> Although research is currently underway on computer-based methods for comparing questioned and known items, and assigning probative value to comparisons, in the current state of forensic practice such methods are not yet widely employed for case-specific evaluations, if at all. Instead, automatic comparison methods are mainly used for investigative purposes, such as the screening of large databases and retrieving specimens with similar features and ranking these specimens according to their degree of similarity with respect to a searched item. (BBC pg. 10)

Many of the problems identified with participant sampling become less problematic for external validity if objective methods are used which reduce the variability of examiner conclusions by providing quantitative information that is similar across examiners, reliable for decision-making, and the result of audit-able, explainable calculations. We firmly believe that this is the best path forward for pattern-based forensic evidence, and we have been actively involved in developing, implementing, and validating algorithms intended for direct item-to-item comparisons[56]. These algorithms are different from database searches such as NIBIN and IBIS that are designed to return the N closest matches from the database in that they provide a direct measure of feature similarity between two specified samples.

One issue raised by both the FBI and BBC, as well as other expert witnesses, is that researchers at the Center for Statistics and Applications in Forensic Evidence (CSAFE) has used data from error rate studies in our own research. One reason we have been able to make use of this data is that because we design algorithms, we can be sure that some of the biases which exist in validation studies do not exist in our research. This distinction is illustrative of the

[56] Eric Hare et al., *Automatic matching of bullet land impressions*, 11 The Annals of Applied Statistics 2332–2356 (2017); Susan Vanderplas et al., *Comparison of three similarity scores for bullet LEA matching*, Forensic Science International 110167 (2020), http://www.sciencedirect.com/science/article/pii/S0379073820300293 (last visited Feb 10, 2020); Joe Zemmels, Heike Hofmann & Susan VanderPlas, cmcR: An implementation of the 'congruent matching cells' method (2022).

differences between algorithm validation studies and examiner validation studies. Consider, for instance, our use of data from closed-set studies[57] when developing an algorithm for assessing the similarity of different bullets. We obtained several test sets used in the study and, using a digital microscope, created 3D scans of the surface of the fired bullets. Then, we developed statistical methods to calculate features from those 3D scans; these features were fed into an algorithm that takes two scans, computes the features, and evaluates the similarity of the two features, eventually boiling down all of that data into a number between 0 and 1, where 0 indicates extreme dissimilarity and 1 indicates extreme similarity between the two scans. We know that our algorithm is not capable of using any of the information about the fact that the scans are from a closed-set study, because we can see exactly what features are being computed and how those features are combined to arrive at the final similarity score. That is, our algorithm is audit-able and fundamentally transparent in a way that the examiner's conclusion is not. We know exactly what information was used to train the algorithm, and how generalizable the algorithm is to data outside of the training set (for instance, its performance on a different model of firearm with similar manufacturing techniques)[58]. Because our algorithm does not depend on examiner responses to the validation study, but instead depends only on the 3D scans of bullets sent to examiners, we can use the bullet scans without compromising our algorithm's internal or external validity.

In addition, some CSAFE researchers who are not part of this discussion have used validation study data in order to demonstrate the use of statistical analysis techniques in forensics contexts. We are not the extremists that BBC and the FBI have painted us as: we will continue working within the system to improve statistical analysis methodology at the same time as we push for better study designs and the use of objective assessment methods. We see this as the most pragmatic approach to improve the discipline as a whole: while we will continue to argue that error rates derived from FTE validation studies are not sufficiently reliable for use, we will also push for the adoption of better statistical analysis methods in the academic forensic evaluation literature.

6.3 Are Tests Like Casework? An Assessment of External Validity

One of BBC's arguments against the calculation of a general domain-wide error rate is that existing studies fall short of mimicking casework and may not apply to a particular case:

> [Black box] studies give only a snapshot of the performance of a selected number of examiners in conducting a particular task under more or less controlled experimental conditions. The experimental nature of these studies implies that, by definition, they fall short of mimicking casework conditions to at least some extent and may not apply to the circumstances in a particular case. (BBC pg. 18)

[57] Hamby et al., *supra* note 54.
[58] Vanderplas et al., *supra* note 65.

There is at least one study[59] that used blind proficiency testing, which mimics casework better than most studies in that 1) it is truly blind, that is, the participants are not aware that they are being tested[60], and 2) the study incorporates the verification protocols used at Houston Forensic Science Center (HFSC), which are not usually incorporated into the error rate calculations in FTE studies. In addition, this study is free from some of the participant selection biases present in other studies by virtue of the fact that examiners were essentially compelled to participate as part of continued employment, and thus sampling and selection biases were not a concern. As with most things, however, there are trade-offs: the more narrow the study's participants, the lower our ability to generalize results to a wider population. This study only covered the Houston Forensic Science Center, so it is difficult to generalize the results outside of examiners at HFSC, where different protocols would be used and examiners would be expected to have different training and mentoring opportunities.

A similar statement is found in the FBI's response:

> Another important point is that these studies capture the participants' conclusions without the benefit of the verification process and other quality control measures utilized during actual casework. These measures include independent examination of the evidence by another qualified examiner (i.e., verification) before a report may be issued. They also include administrative and technical review of an examiner's report. These quality control measures would likely lower the error rates reported in these studies even further. (FBI pg. 4)

We would love to see more error rate studies conducted using blind proficiency tests; such studies clearly have better external validity in some respects, even if they often cannot be generalized outside of the laboratory where they were conducted. We recognize that not all laboratories have the resources of HFSC, and that such testing is expensive; as a result, it is still beneficial to the discipline to have error rate studies which serve as estimates of examiner error without the benefit of verification processes, because such estimates are usually derived from examiners across multiple laboratories and thus can, under the right sampling procedures, be generalized to a wider population of examiners. If the data from these proficiency tests were made available to the community in an anonymized way, it might even be possible to assess the effect of the verification process on the error rates, which would be useful information for interpreting studies without that verification process (it might be possible to estimate e.g. the magnitude of the reduction in error based on a verification process similar to that used at HFSC).

[59] Maddisen Neuman et al., *Blind testing in firearms: Preliminary results from a blind quality control program*, 67 Journal of Forensic Sciences 964–974 (2022), https://onlinelibrary.wiley.com/doi/abs/10.1111/1556-4029.15031 (last visited Jun 18, 2022).

[60] note that this definition of "blind" is more strict than that sometimes used by forensic scientists, in which a blind test means that the person being tested doesn't know the answers (cite Bunch & Murphy).In experimental design, the notion of "blind" testing refers to participants and experimenters not knowing who was assigned to each treatment group because such knowledge might influence the test evaluation. In order for the same aim to be achieved in forensic tests, we must instead ensure that the examiner does not know that they are being tested so that we can more accurately measure how they respond to case work.

If the circumstances of a particular case are such that error rate studies are not applicable, as suggested by BBC, then that is something that should be brought up when the firearms and toolmark expert is testifying. While it is unlikely that a specific error rate or numerical adjustment could be identified, this would at least allow the judge and/or jury to identify a starting point and a direction in which the error rate might be revised.

Our prior statement, and this statement, address the general discipline of firearms and toolmark examination. We focus on assessing the question of whether firearms and toolmark evidence has broad scientific support, with the conclusion that while there is some scientific evidence to support the idea that firearms and toolmark examination is useful for assessing questions of source, the quality of that evidence falls well short of that required for "broad scientific support" due to fundamental issues with internal and external validity in the validation studies which exist to date.

6.4 Nonresponse Bias

It is common for studies involving human subjects to involve some degree of drop-out or nonresponse. Individuals may agree to participate in a survey and then fail to actually engage (drop out) or they may leave some survey questions unanswered (item nonresponse). There are many statistical methods to handle these problems.[61]

In order to begin to address these problems, researchers first have to acknowledge them. In every study we have reviewed, the limitations due to nonresponse and drop-out bias are not acknowledged. No study utilizes common statistical methods for assessing the impact of nonresponse and drop-out bias[62]. More troubling, these studies do not release any data to facilitate other researchers filling in these gaps.

As the holders of the data, the researchers conducting validation studies are the ones who bear the burden of addressing the missingness in their analyses. Choosing the correct methods depends on exploring the patterns of missingness in the data. Instead, currently, these researchers ignore the problem and proceed with inappropriate statistical analyses- despite the availability of existing appropriate methods that could be used.

The authors of BBC and the FBI responses do not refute these statements. Instead, they attempt to distract from the issue.

> This assertion is a further example of the use of a true statement (here: the existence of non-responses) for suggesting conclusions based on assumptions for which actual evidence is lacking. That is, Vanderplas et al. (2022) provide no basis to believe

[61] There are, in fact, entire areas of statistical research devoted to such methods. For some examples, see Roderick JA Little & Donald B Rubin, 793, Statistical analysis with missing data (2019) and Jae Kwang Kim & Jun Shao, Statistical methods for handling incomplete data (2014).

[62] Angela M Wood, Ian R White & Simon G Thompson, *Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals*, 1 Clinical Trials 368–376 (2004), https://doi.org/10.1191/1740774504cn032oa (last visited Jun 23, 2022).

> that all non-respondents would render erroneous answers; an error rate based on such an extreme assumption is hypothetical and not conducive of advancing a constructive discourse over what the potential of error could realistically be. In line with our discussion throughout this document, we reiterate that (i) the imperfection of existing studies and related data is not contested, (ii) imperfect data should not be dismissed entirely (provided that limitations are properly acknowledged), but interpreted within the relevant scope (e.g., limiting conclusions to those examiners who properly responded), and (iii) even if data were perfect (in strict statistical terms), the resulting domain-wide error rate would characterize an abstract question and, hence, be of limited practical usefulness. (BBC pg. 27)

As we have discussed, limitations are *not* being acknowledged. We are also not arguing imperfect data needs to be dismissed entirely. Instead, we assert the simple fact: researchers are inappropriately using methods developed for completely observed data for data which are far from completely observed. Deflecting again from this issue, the authors of BBC take umbrage with our suggestion that the nonresponse is likely leading to underestimates of the error rates.

> The Statement claims that "[g]iven what we know about why people drop out of black box studies; we would expect that studies with non-response bias underestimate the error rate." It is unclear what the Statement "knows" about why people drop out of black box studies, as it cites no data that supports this claim. (FBI pg. 11)

Research into testing and assessment in the educational setting has consistently indicated that "intuition and empirical evidence" support that "[E]xaminees are more likely to omit items when they think their answers are incorrect than items they think their answer would be correct."[63] If an examinee is proficient enough to know when they are likely to be incorrect, then this type of behavior will lead to an underestimate of error rates if missingness is ignored.

We rely on what is known about testing more generally to suggest a direction of bias because the data from validation studies are typically not shared. To our knowledge, no FTE validation study has released any data capable of being analyzed by a third party. However, a recent study for palmar prints by Eldridge, De Donno, and Champod[64] did release some data. While the released data does not contain sufficient information to apply methods to adjust for missingness, it does allow for the beginning of an exploration of the patterns of missingness. For example, Eldridge, De Donno, and Champod[65] identified two factors that were at associated with higher false positive error rates among examiners. These factors were being a non-active examiner and

[63] Robert J Mislevy & Pao-Kuei Wu, *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing*, 1996 ETS Research Report Series i–36 (1996) pg. 16. See also, Steffi Pohl, Linda Gräfe & Norman Rose, *Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models*, 74 Educational and Psychological Measurement 423–452 (2014) and Shenghai Dai, *Handling missing responses in psychometrics: Methods and software*, 3 Psych 673–693 (2021).

[64] *Supra* note 39.

[65] *Id.*

being employed by an agency outside of the United States. We explored both characteristics and their relationships with missingness. Our analyses indicated that being employed by an agency outside of the United States was also associated with a higher likelihood of examiners failing to respond to over 50% of their assigned comparisons. In other words, a group of examiners who were disproportionately likely to make false positives were also disproportionately likely to skip comparisons. Thus, for this study, there is evidence that the false positive error rate calculated ignoring the missingness is an underestimate.

7 Inconclusives

7.1 The Importance of Both Identification and Elimination

> When courts choose to consider the known or potential rate of error as a factor bearing on reliability, the key concern for admissibility is the frequency of false identifications. (FBI pg. 1-2)

The FBI is not alone in their assertion that false identifications are important. Such claims are made by expert testimony[66] and even in the PCAST report, the criteria for foundational validity of a forensic discipline are the sensitivity rate and the false-positive rate.[67] We agree that the false positive (false identification) rate is important, but there are fundamental issues with the focus only on identifications when we look at the structural setup of evaluating examiner conclusions, summarized in Figure 1.

If examiners are only able to spot similarities, then there should be only one threshold: either the samples under comparison are sufficiently similar, or they are not. This results in a binary classification problem - one which neatly matches the true state of the evidence: either the two items were from the same source, or they were from different sources.

If examiners can spot similarities and differences, but only focus on similarities, then they are ignoring available evidence which might be exculpatory, either because of training biases to look for similarities or because identifying differences is a harder cognitive problem. In this case, the system is set up to evaluate examiners based on whether they can identify both similarities and differences, with a middle category of inconclusive for examiners to use when there is insufficient evidence in either direction. Using such an evaluation system while focusing only on one type of error is problematic from the standpoint of objectively evaluating examiners' claims about the scientific nature of their discipline.

The FBI's discussion of the concept of the "Best Known Non Match" suggests that they are looking only at similarities:

[66] Todd Weller in People v. Ross, 68 Misc. 3d 899, 129 N.Y.S.3d 629, 2020 N.Y. Slip Op. 20153 (N.Y. Sup. Ct. 2020)

[67] President's Council of Advisors on Science and Technology, *supra* note 49, pg. 159.

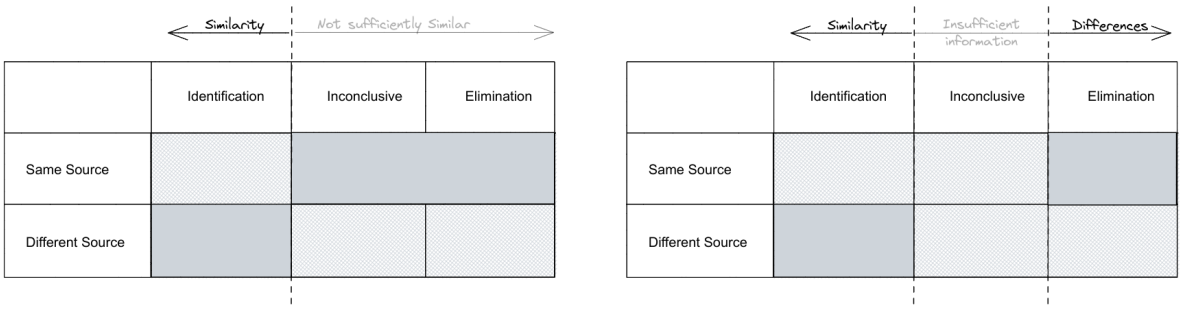

Figure 1: If examiners only spot similarities, then the classification scheme on the left is appropriate and examiners should confine themselves only to claiming to be able to make identifications, grouping inconclusives and eliminations together as having insufficient similarity to make an identification. If examiners spot similarities and differences, then it is important to evaluate the error rate of both false identifications and false eliminations, as it speaks to the fundamentals of the claims examiners make about their abilities.

> The ability to assess pattern agreement develops during training when an examiner evaluates the "best" agreement between two specimens known to have originated from different sources — "the Best-Known Non-Match." (FBI pg. 3)

while BBC suggest that there is not even agreement on what different examiners might consider similarities and/or differences:

> different examiners may assign different evidential values to observed features, and they may even disagree about what exactly constitute similarities and differences (in accidental characteristics) for a given pair of compared items. (BBC pg. 10)

We bring up the issue of how errors are counted in part to point out that even the basic criteria underlying subjective assessment of firearms and toolmark evidence are not agreed upon by examiners, and in part because there is a fundamental mismatch between the evaluation criteria examiners appear to use and the way the errors assessed in the community. This issue is at the heart of HH, SV, and AC's paper on inconclusives[68]. While BBC identify statements made in this paper as inconsistent with statements in our affidavit, we would like to highlight the difference in context: in the Law, Probability, and Risk paper, we were examining specifically the use of inconclusives in error rate studies; in our affidavit we were examining the utility of error rate studies when evaluating the discipline of firearms and toolmark examination. The latter is a much broader question which requires consideration not only of study design, but also of sampling and general statistical procedures. We are accustomed to the nuances of data

[68] Hofmann, Vanderplas, and Carriquiry, *supra* note 26; It is customary in statistics to cite the print edition once the paper has been released; this is why in the responses the paper is given the year 2020 and in our citation it is listed as 2021. The paper was released online before the official release of the print edition.

collection and analysis, including framing the question under investigation in such a way that it can be precisely answered within the bounds of the data which has been collected.

7.2 Probative Value of Inconclusives

> "... a typical item of evidence (or observation made by a scientist) may not only occur when one hypothesis (i.e., one version of a contested event) is true, but also when an alternative hypothesis is true." ... "We note that what is of crucial importance for our discussion throughout this document is that, in general, for evidence to have probative value with respect to two competing hypotheses, the probability that the evidence would arise under one hypothesis must be different from the probability of that evidence to arise under the respective alternative hypothesis. In essence, we would like to have evidence that is (much) more typically encountered if one version of a contested event is true rather than some alternative version. Evidence that has this property is said to have discriminative capacity – i.e., it has (probative) value." (BBC pg. 10-11)

Using this definition, we have previously shown[69] that inconclusives have probative value - they are much more likely to occur when evidence is from different sources than when evidence is from the same source. While we acknowledge that there is considerable disagreement between experts in the area of inconclusives, we firmly believe that the treatment of inconclusives as correct decisions by FTEs and error rate studies is incorrect based on the logic that underlies most scientific studies: statistical hypothesis testing.

In a statistical hypothesis test, we start out with a conclusion that we want to disprove, called the null hypothesis (H_0 in mathematical notation). The null hypothesis might be "Plant growth is not associated with increased moisture", or it might be "the two items originate from different sources". Then, a statistical experiment is conducted and evidence is assembled, with the assumption that the null hypothesis is true. A probability is calculated which rests on the assumption that H_0 is true; if that probability is sufficiently small, then we conclude that we are unlikely to have observed our data if H_0 is true, and that there is evidence to support the alternative.

On the left side of Figure 1, it is possible to see how this plays out in firearm and toolmark assessment. We start by assuming that the two pieces of evidence come from different sources. As the FBI has indicated, examiners are trained to look for similarities, suggesting that as similar features accumulate, the conclusion moves from "different sources" to "same source" - that is, the accumulation of similarities between the two items causes the examiner to reject the null hypothesis and conclude that the items must have been originated from the same source. If sufficient evidence to refute H_0 does not accumulate, we cannot say anything about H_0 or the alternative, H_A. That is, we do not ever "accept" that H_0 is true (that is, an examiner would never need to conclude that the sources of the items were different); we simply do not

[69] Hofmann, Vanderplas, and Carriquiry, *Id.*

have enough similarities to reject the hypothesis that the two items are from different sources. It would, of course, be possible to start from the opposite conclusion: we could start with a null hypothesis that the two items are from the same source, and look for differences. This is not, however, how examiners seem to arrive at their conclusions. Rather, it seems that by training and in describing how they arrive at their decisions, examiners overwhelmingly focus on similarities.

This statistical hypothesis testing logic is very similar to the framework of the criminal justice system. If the jury is convinced "beyond a reasonable doubt", then the defendant is declared to be guilty (the presumption of not guilty, H_0, is rejected in favor of the H_A of guilt). Otherwise, the defendant is declared to be "not guilty". There is no way for the defendant to be declared innocent, because the system is set up to refute the starting premise that the defendant is not guilty, with evidence presented that accumulates against that hypothesis until a certain threshold is met.

What the FBI and BBC are advocating for, that is, the utility of inconclusives, is akin to having a legal system in which individuals are judged guilty, not guilty, or unknown. While that is something that would reduce the probability that the innocent are convicted or the guilty go free, it also allows for a large grey area in what is set up to be a decisive, binary system. The judicial system would not function well if a large proportion of cases were inconclusive and did not reach some sort of decisive resolution, but forensic disciplines tolerate this situation because it decreases nominal error rates.

8 Conclusion

As we have demonstrated in this document and our previous affidavit, there are substantial threats to both the internal and external validity of currently available firearms studies. Statistically, these concerns are primarily the result of the design and analysis of firearms and toolmark error rate studies, rather than as a result of the work that examiners do on a day-to-day basis. The external validity of FTE error rate studies is threatened by participants' self-selection into the sample population, limited assessment of the impact of different combinations of ammunition and firearms, and poor assessment of the impact of nonresponse bias on the error rates reported in each study. In addition to these threats to the generalizability of results, there are also threats to the internal validity of the studies: past use of closed-set and multiple-known comparison set study designs, poor statistical practice, and the treatment of inconclusives.

We remain firm in our conclusion that the estimates established from fundamentally flawed studies with threats to both internal and external validity are not sufficiently sound to be used in high-stakes situations, including medicine, law, and engineering applications where individuals' lives, health, or freedom are at stake.

We declare under the penalty of perjury and pursuant to the laws of the state of Illinois that the statements above are true and accurate to the best of our knowledge.

Alicia Carriquiry	Heike Hofmann	Kori Khan	Susan Vanderplas

EXHIBIT 2

1-3-2022

Firearms and Toolmark Error Rates

Susan VanderPlas
University of Nebraska-Lincoln, svanderplas2@unl.edu

Kori Khan
Iowa State University, kkhan@iastate.edu

Heike Hofmann
Iowa State University, hofmann@iastate.edu

Alicia L. Carriquiry
Iowa State University, alicia@iastate.edu

Firearms and Toolmark Error Rates

Susan Vanderplas, Kori Khan, Heike Hofmann, Alicia Carriquiry

Statement

We declare under penalty of perjury and pursuant to the laws of the state of Illinois that the following is true and accurate to the best of our knowledge.

_____

Alicia Carriquiry

_____

Heike Hofmann

_____

Kori Khan

_____

Susan Vanderplas

1 Qualifications

1.1 Alicia Carriquiry

I, Alicia Carriquiry, hold a Bachelor of Science in Agricultural Engineering from the Universidad de la República del Uruguay, a Master of Science in Animal Genetics from the University of Illinois at Urbana-Champaign, a Master of Science in Statistics from Iowa State University, and a PhD in Statistics and Animal Science from Iowa State University.

I am a Distinguished Professor of Liberal Arts and Sciences at Iowa State University, which is the highest rank that a professor can achieve. I am also the inaugural President's Chair in Statistics. Between 2000 and 2004, I was Associate Provost at Iowa State University, and was Director of Graduate Studies in the Department of Statistics at Iowa State University between 2004 and 2014.

Since 2015, I have been the Director of the Center for Statistics and Applications in Forensic Evidence (CSAFE), a National Institute of Standards and Technology (NIST) Center of Excellence. The Center is a consortium of universities that includes the University of California Irvine, Carnegie Mellon University, the University of Virginia, Duke University, the West Virginia University, Iowa State University, the University of Nebraska Lincoln, the University of Pennsylvania, and Swarthmore College.

I am the only statistician in the history of Iowa State University to be elected to the National Academies. I became a member of the National Academy of Medicine in 2016 thanks to my applied work in statistics, that includes the design and analysis of studies and surveys. I am a member of the Advisory Board for the Division of Behavioral and Social Sciences and Education (DBASSE), of the Report Review Committee, and of the Committee on Applied and Theoretical Statistics at the National Academies of Science, Engineering and Medicine. I am also a member of the Intelligence Science and Technology Experts Group (ISTEG), a group of members of the National Academies that is available to the US intelligence agencies when questions about science and technology arise.

I was elected Fellow of the American Association for the Advancement of Science (AAAS) in 2016, and an Associate Member of the American Academy of Forensic Science (AAFS) in 2019. Since 2018, I have been a Technical Advisor for the Association of Firearms and Toolmark Examiners (AFTE). I am an elected Fellow of almost all major professional statistical organizations: the American Statistical Association, the Institute of Mathematical Statistics, the International Statistical Institute, and the International Society for Bayesian Analysis.

I have published over 140 peer-reviewed manuscripts in the scientific literature and have co-edited several books. I have directed the doctoral research work of 20 students in statistics and related areas, and have been invited to speak at national and international conferences, and at other venues hundreds of times. I have received competitive research funds from the National Science Foundation, the National Institutes of Health, the National Institute of Standards and Technology, the Office of Naval Research, the United States Department of Agriculture, the Federal Bureau of Investigations, and several other federal, state, and private organizations, totaling well over $50M during my professional career.

1.2 Heike Hofmann

I, Heike Hofmann, am a tenured full professor in Statistics at Iowa State University and the professor in charge of the data science program. I have been a faculty member of the Center for Statistics and Applications in Forensic Evidence (CSAFE) since 2015. CSAFE is supported by a cooperative agreement with the National Institute of Standards and Technology (NIST). In my role as CSAFE faculty I have been a key contributor for the work that received the ASA SPAIG (Statistical Partnerships Among Academe, Industry & Government) award in 2018.

I am an elected member of the International Statistical Institute and an elected fellow of the American Statistical Association (ASA). I have been named a Technical Advisor to the Association of Firearms and Toolmark Examiners. I serve as a member of the ASA committee on forensic statistics.

I have published extensively in peer reviewed, scientific journals. My publications cover several areas including

computationally intensive methodology for automatic matching of pattern and image evidence, error rate analyses and development of open-source software for reproducible, accessible comparison of firearm evidence. I maintain the `bulletxtrctr` and `x3ptools` R packages, which have been released to the community to facilitate analysis of bullets and 3D scans respectively.

1.3 Kori Khan

I, Kori Khan, hold a Bachelor of Science in Public Health in Biostatistics from the Gillings School of Global Public Health at the University of North Carolina at Chapel Hill, a Juris Doctorate from the Moritz College of Law School at the Ohio State University, a Masters of Science in Statistics from the Ohio State University, and a Ph.D in Statistics from the Ohio State University.

I am an Assistant Professor in Statistics at Iowa State University. My areas of expertise include the analysis of dependent data, which can include repeated observations of individuals in experiments. I frequently serve as a referee for the top journals of statistics and applied statistics, including the American Statistics Associations' flagship journal.

I have in the past and continue to provide consultations regarding the design and analysis of complex data collection efforts and experimental design for studies focused on human subjects for researchers at organizations including the Moritz College of Law at the Ohio State University, Weill Cornell Medicine, and the Institute for Population Research at the Ohio State University.

I currently serve as sub-committee member on the Organization of Scientific Area Committees (OSAC) for Forensic Science Forensic Nursing and the Statistics Task Group. I have, in the past, served as a member of a Scientific & Technical Review Panel for the same organization.

I have personally reviewed the experimental designs, reporting of the obtained data, and statistical methodologies of the 11 studies referenced in Appendix B.

1.4 Susan Vanderplas

I, Susan Vanderplas, hold a Bachelor of Science in Applied Mathematical Sciences and Psychology from Texas A&M University, a Masters of Science in Statistics from Iowa State University, and a Ph.D. in Statistics from Iowa State University.

I am an Assistant Professor in Statistics at University of Nebraska Lincoln. My areas of expertise include data visualization, perception, and communication of statistical information, as well as machine learning algorithms for statistical analysis of forensic pattern evidence, including firearms and footwear.

Between 2018 and 2020 I served as a research assistant professor at the Center for Statistics and Applications in Forensics (CSAFE) at Iowa State University — CSAFE is a collaboration among several university scholars conducting research on how to collect more accurate forensic evidence and more reliably convey that forensic evidence in court. CSAFE is supported by a cooperative agreement with the National Institute of Standards and Technology (NIST)—a federal agency whose mission is to advance measurement science, standards, and technology.

I am the author or co-author of more than a dozen peer-reviewed scientific papers. Publications most relevant here are published in the Handbook of Forensic Science, Law, Probability and Risk, and Forensic Science International. These papers cover machine learning methods for automatic matching of forensic pattern evidence, error rates in forensic studies, and the treatment of inconclusive results in firearm studies. In addition to my own publications, I have served or been asked to serve as a reviewer for Forensic Science International, the Harvard Data Science Review, Forensic Sciences Research, The American Statistician, Journal of the American Statistical Association, Science & Justice, and the Journal of Computational and Graphical Statistics. I also serve on the editorial board of the R Journal and the Journal of Computational and Graphical Statistics.

I am the primary investigator (PI) of a research and development grant on the acquisition and identification of footwear class characteristics funded by the National Institute of Justice. I am also PI on the CSAFE

sub-award at UNL, which covers research in firearms, footwear, and human factors. I am also a co-PI on grants from the National Science Foundation and the United States Department of Agriculture.

2 Introduction - Firearm and Toolmark Error Rates

There are an impressive array of existing studies of the "validity" of firearms and toolmark comparisons [Keisler, M. A., Hartman, S., and Kil, A. (2018); Lightstone (2010); Riva et al. (2017); Bunch and Murphy (2003); Brian Mayland and Tucker (2012); Duez et al. (2018); Pauw-Vugts et al. (2013); Neel and Wells (2007); Hamby et al. (2019); Gouwe, Hamby, and Norris (2008); Giroux (2009); Stroman (2014); Lyons (2009); Mattijssen et al. (2020); T. P. Smith, Smith, and Snipes (2016)] across different firearms, ammunition types, types of marks made, and even across different countries with different protocols for training and firearms and toolmark examination. As statisticians, however, we have significant qualms with the state of error rate studies in firearms and toolmark examination. Many studies are poorly designed, with problems ranging from a complete inability to characterize the full error rate (Hamby et al. 2019; Gouwe, Hamby, and Norris 2008; Giroux 2009; Thomas G. Fadul, Jr et al. 2013; Lyons 2009) to the acknowledged inability of examiners to follow the instructions set out by the researchers (Baldwin et al. 2014). Furthermore, all of the studies we are aware of which are applicable to the state of firearms and toolmark examination as practiced in the United States at this time suffer from sampling and non-response bias that renders them unreliable for the purposes of establishing the science of firearms and toolmark examination as a reliable discipline.

Here, we will lay out some of the fundamental problems with the state of firearms and toolmark examination error rate studies. We approach these problems both as statisticians who have experience in the design of scientific experiments and as researchers in statistical applications to forensic evidence. We have worked extensively with forensic examiners, metrologists, and other subject-matter experts, and we have an understanding of both the process of firearms and toolmark examination and the statistical underpinnings of estimation of error rates. Before we begin assessing the current state of error rate studies, however, it is useful to establish the characteristics necessary to have a reliable study.

3 Study Design

A statistical estimate is only as reliable as the data used to generate that estimate. In the case of black-box studies, this data originates from the examiners who participate in the study. In order for statistical estimates to be unbiased, the participants selected from the population to participate in the study (the sample) must be representative. The best way to ensure that a sample is representative is to randomly select this set of participants from the population.

Practically speaking, it is relatively rare that all participants selected for a study complete the study. Scientific ethics requires that research using human participants be undertaken with the participants' consent, and that consent can be withdrawn at any time. As a result, it is common for studies to have some level of drop-out rate. In these situations, the risk to the statistical validity of the study is that participants who drop out of the study vary systematically relative to those who remain in the study. One general guideline is that if less than 5% of the participants drop out, there is little threat to the statistical validity of the study, but if more than 20% of the participants drop out, the study's validity is severely compromised (Schulz and Grimes 2002). It is extremely important that when reporting study results, authors clearly state the level of participant drop out and assess whether there is any evidence of bias in the participants who remain relative to the participants who dropped out.

Another important factor in study design is that studies should be designed in such a way as to directly evaluate the desired conclusions and/or produce the desired numerical estimates. If a study is intended to assess the false positive error rate of firearms and toolmark evaluation, then examiners need to have the opportunity to make a false positive error. While this seems straightforward, many common firearms and toolmark black-box study designs do not allow for estimation of the number of different-source comparisons, which ensures that it is not possible to calculate the overall error rate, the correct decision rate, or the true negative rate (the specificity). This problem is detailed at length in Hofmann, Vanderplas, and Carriquiry

(2021) as well as in President's Council of Advisors on Science and Technology (2016) and National Research Council (U.S.) (2009). A well designed black-box study should have a defined number of pairwise comparisons, where each comparison is completed by the examiner with no possibility of eliminating comparisons based on the structure of the test set. In practice, this means that black-box studies should be open set studies (no guarantee that an unknown item matches any provided knowns in the set) and should involve the comparison of one standard (known sample) and one or more unknown samples at a time.

Executing a well-designed study is not an easy task, but it is important that during the study experimenters not provide participants with additional information. This means that experimenters must take care to provide different examiners with test samples which are not shared between examiners in the same lab, and which may have different numbers of same-source and different-source samples. This prevents information from "leaking" from examiner to examiner within a relatively small community. In addition, experimenters should observe strict protocols when communicating with participants to avoid sharing information about the task beyond what is specified in the instructions.

Finally, it is important that experimenters provide the community with all relevant information when reporting study results. It is difficult (and sometimes, impossible) to reconstruct the full set of aggregated answers when the only data reported in a study is the error rate. Experimenters should provide data at the individual level if it is possible to do so without identifying participants, and if this is not possible, data should be provided at the lowest level of aggregation which maintains participant anonymity.

4 Participant Sampling Problems

One of the primary concerns with error rates provided by "well-designed" studies is that even well designed, well-executed studies cannot compensate for sampling bias in the participant pool. That is, no matter how well the experiment is laid out, if the participants are not a representative sample from the population (in this case, all qualified firearms examiners in the United States), the results of the study do not generalize to that population. This principle is taught in even basic undergraduate statistics courses; it is fundamental to our discipline. One of the easiest ways to ensure that a sample is representative is to randomly select participants from the population; a more labor-intensive option is to conduct a full census of the population at a certain time.

Fundamentally, because almost all current black-box studies use volunteers, we can conclusively state that the participants in these studies are likely to meaningfully differ from those who did not volunteer to participate. Some of these differences are likely not related to the error rates estimated by the studies, but there are many potential lurking covariates that would meaningfully affect the error rates estimated by the studies. For instance, it is possible that experienced examiners are more likely to volunteer to participate in these studies out of a sense of duty to the discipline: these examiners might have lower error rates due to their experience, which would lead to an estimated error rate that is lower than the error rate of the general population of all firearms examiners (including those who are inexperienced). In fact, in studies which differentiate between trainee and qualified examiners, we find a higher error rate among trainees(Duez et al. 2018).

Not all potential biases are this direct: it is possible that examiners who have time to volunteer to participate in studies would tend to have lower case loads. Thus, these examiners would be over-represented in the study-wide estimate of the error rate, in that they account for fewer cases than the examiners who do not have time to participate in a voluntary study. Thus, the estimated error rate from the study would not be representative of the error rate of all examiners. There are many variables which might be expected to increase likelihood of volunteering for a study and also change the expected error rate: education, experience, confidence, amount of time available for study participation.

An additional source of participant recruitment bias is that many studies recruit via the AFTE membership forums (Thomas G. Fadul, Jr et al. 2013; Keisler, M. A., Hartman, S., and Kil, A. 2018; Chumbley et al. 2021), which lead to an over-representation of AFTE members and certified examiners (and in particular, those who spend time on the membership boards). There is not a list of examiners who are allowed to testify in court which could be used for sampling for such a study, but we might expect that people who are active in AFTE are more invested in the discipline, may have more training, and may thus have a lower error rate.

Without a random sample of all qualified examiners, there is no way to generalize the results of a biased sample to the whole population; any study only speaks for the error rate of the participants of that study. Random selection of participants mitigates these potential biases by ensuring that any differences in the sample of selected participants are the product of random assortment - while any single experiment might have a random sample of participants who are not fully representative of the population, each experiment's different samples will produce overall unbiased results. This is why it is not only important that study participants be randomly sampled from the population, but also that there are multiple studies.

Unfortunately, sampling bias is one of the hardest biases to work around. Because we cannot determine how the volunteer examiners might differ from the whole population of examiners, we cannot say that it is likely that the error rate is higher or lower than what is reported from the flawed studies. While in some areas it is necessary to work with volunteer populations (for instance, clinical trials take place on volunteers), this requires that statisticians can reasonably expect that there are no differences between the volunteer and non-volunteer populations, which is a claim that is more easily made in medicine than in forensics. [1]

This bias affects all studies conducted in the United States to date. Currently, there is no way to randomly sample all qualified examiners and compel them to participate. This problem is less pervasive in Europe, where lab certification and validation studies are conducted to assess this type of error rate and certification is conditioned upon participation.

5 Material Sampling Problems

As scientists, we want to derive knowledge that is generalizable to the population. The population includes all decisions made from a combination of one or more firearms of interest, and the ammunition that interacts with those firearms. This combination of characteristics is then evaluated by an examiner. We have previously addressed the fact that we need a random sample of examiners in order to generalize to the whole population of firearms examiners. However, we also need a representative set of ammunition/firearm combinations used in black-box studies in order to validate the discipline as a whole.

Many black-box studies are performed on a single type of firearm and a single type of ammunition. In some cases, these firearms are consecutively manufactured, which allows firearms examiners to prove that they can distinguish individual firearms on the basis of fine details, even when these firearms are manufactured closely in time. However, this does reduce the generalizability of the conclusions from these small studies: by examining only firearms of the same make and model manufactured closely in time, we lose the ability to make broad, sweeping claims about the discipline as a whole. In some cases, studies explicitly document that small manufacturing companies with limited operations (and thus generalizability to the wider population) were selected because it was more convenient to obtain consecutively manufactured components (Lyons 2009).

We need black box studies that examine large numbers of firearms as well as those that examine minute differences. We are not currently aware of any study of a large number of firearms of even one specific model, though we are aware of some data which has been collected but not published examining 600 Berettas confiscated by the LAPD. We cannot generalize error rates from small consecutively manufactured firearm studies to the entire population of firearms examinations, and as a result, we do not know how to assess the error rate of the discipline as a whole on the basis of these studies.

Researchers are well aware of these limitations and typically characterize their findings in a much more limited fashion than some professional expert witnesses. Riva et al. (2017) suggest their study is limited in scope and conclusions: "Finally, even if the results obtained in this study illustrate the impact of subclass characteristics for a given make and model of firearm, they cannot be easily transposed to all firearms at this stage. We remain conscious of the limitation of the sample used here. It is known that the quality and the quantity of these features will vary as a function of the type of firearms and the manufacturing process." This makes it

[1] In medicine, it is reasonable to think that someone's willingness to participate in research is not related to their biological response to cancer treatment, as one is psychological and the other is physiological. Forensic examiners, on the other hand, are trained as scientists - the effectiveness of their training is related to their willingness to participate in the scientific process, and they have a stake in the outcome of the experiment in that if the experiment shows that toolmark examination is not reliable, they are out of a job.

extremely difficult for researchers to provide a general estimate of the error rate of firearms and toolmark comparisons, as the discipline is so broad and the data under examination are affected by so many different factors: type of tool or firearm, ammunition, manufacturing process, material interactions, and so on.

We know that not all firearms and ammunition mark equally well - Glocks are renown for being difficult to compare bullets, but for having easy cartridge comparisons. In addition, some ammunition marks better than other ammunition. While it is possible to characterize the error rates of certain studies (subject to the sampling and non-response biases noted earlier), studies are generally conducted with well-marking tools or firearms. Baldwin et al. (2014) did not examine the samples for issues before sending them to examiners for evaluation; due to the design of the study the researchers estimated that 2-3% of known samples were judged to be inappropriate for comparison due to not marking well. If studies are done on ammunition that is generally known to mark well, then we should expect that this non-marking issue is more likely to show up in casework than in studies - that is, currently existing studies are not representative, and error rates from these studies may not generalize well to firearms and ammunition with different properties.

6 Missing Data and Non-response Bias

In addition to the fundamental problems with volunteer-only participation in black-box studies, there are additional statistical issues which plague most black-box studies: even participants who volunteer for the study may not return the full set of answers (or any answers). That is, most studies have a drop-out rate where participants who initially volunteered for the study did not complete the full study. In statistics, this sort of missing data is more problematic if it is "missing not-at-random" - that is, if the participants who do not return full sets of answers are systematically different from those who do return full sets of answers. Some studies deal with this issue in ways that would not be considered statistically valid: Lyons (2009) contacted someone who returned an incomplete answer sheet to prompt them to complete the questions and provide more satisfactory answers. In other studies, such as Chumbley et al. (2021), the issue is acknowledged but not mitigated or even assessed for its' impact on the error rate estimates reported[2].

An example of a situation in which non response or missing not-at-random bias is common is in telephone surveys. We know that people who are more likely to answer a phone call from an unknown number are different from those who do not answer phone calls from unknown numbers; we also know that those people who continue on the phone call to answer all of the poll questions are more likely to be engaged in their communities and have a higher sense of civic responsibility. These biases can sometimes be corrected for statistically when they are a well-studied quantity (as in political polling), but even the models which allow for pollsters to adjust their estimates to account for these biases are sometimes wrong, leading to systematically biased polls. The same set of problems arise when we ignore non-response bias or missing-not-at-random data in black-box studies, but we do not have sufficient data to adjust the estimates based on the issues, because black-box studies in forensics are not nearly as well studied as political polling. The scientific community also do not have the ability to start studying these issues because the authors of black box studies almost never make the data publicly available. This last point is particularly concerning given that the broader scientific community recognizes that publicly available data is necessary to ensure studies are reaching valid statistical conclusions (Wicherts, Bakker, and Molenaar (2011),Stodden (2015))

What is concerning is that most studies do not even indicate levels of non-response bias. This issue is seldom mentioned when describing the conduct of the study, so there is no way for us to assess the magnitude of the problem given the current conventions for reporting non-response bias. One option to bound the problem is to estimate the error rate with an assumed drop-out rate in order to determine just how bad the issue would be if the participants who dropped out had completely incorrect responses. Note that this requires two assumptions: one, that the drop out rate is a certain percentage - in this case, 20%, and two, that the participants who dropped out would have been completely wrong on every comparison. Keisler, M. A., Hartman, S., and Kil, A. (2018) has one of the lowest error rates and highest number of comparisons of any study reported (2520 comparisons, 0% error counting inconclusives as correct). With a 20% non-response

[2]"Slightly more cartridge case (10,110) than bullet (10,020) comparison are reported because some examiners returned partially-completed test packets that had results for all the cartridge case sets but not all the bullet sets." (Chumbley et al. (2021), page 8)

rate (which is conservative relative to rates reported in the literature), the overall error rate in the population (counting inconclusives as correct) could be as high as 16.56% if all non-responses made completely incorrect decisions. While this is unlikely, it does provide an upper bound of the order of magnitude of the possible bias that participant drop-out might have on estimated error rates.

This missing data may occur at the participant level (drop out bias) or at the question level (item non-response bias), but the likely reasons for non response in either form include insufficient time to commit to completing the study or finding the study more difficult than expected and not wanting to return the wrong answers. In either case, this is a potential source of bias for the estimated error rates - if the individual is not sure of the answers or does not have enough time to dedicate to the study, it is likely that estimated error rates with complete data would be higher than estimated error rates with incomplete data. That is, given what we know about why people drop out of black-box studies, we would expect that studies with non response bias under-estimate the error rate.

7 Types of Marks and Study Difficulty

It is common to talk of error rates for firearms and toolmark examination as a whole. While it is certainly true that there are some similarities across different comparison types (toolmarks and land-engraved areas both involve comparisons of striae), there are many different types of marks used for firearms and toolmark examination, and it would not be reasonable to assume they all have the same error rates. For example, striations on land engraved areas usually involve comparison of a number of different lands, most of which must match for the overall comparison to be deemed an identification - there is some redundancy in this comparison, whereas the comparison of a single scrape from a screwdriver involves only one set of striae. Even though the two comparisons are objectively similar in that the examiner is visually comparing striae, we would not expect them to have similar error rates because there is an internal check on incidental matches when bullets are compared that does not exist with screwdriver engraving comparisons.

Expert witnesses[3] have previously testified about high error rates in a closed-set study [Brian Mayland and Tucker (2012)] and low error rates in an open-set study (Keisler, M. A., Hartman, S., and Kil, A. 2018). What is particularly remarkable about this portion of the testimony is that Brian Mayland and Tucker (2012) examined obturation marks, while Keisler, M. A., Hartman, S., and Kil, A. (2018) examined breech face impressions. These two types of marks are entirely different in character, and thus the error rates should never be directly compared. Of the studies we have cited on firearms and toolmark examination, some examine bullet striae (Hamby et al. 2019; Pauw-Vugts et al. 2013), extractors (Lyons 2009), cartridge cases (Baldwin et al. 2014; Chapnick et al. 2021; Pauw-Vugts et al. 2013; Bunch and Murphy 2003; Duez et al. 2018; Keisler, M. A., Hartman, S., and Kil, A. 2018), aperture marks (Mattijssen et al. 2020), obturation marks (Brian Mayland and Tucker 2012), and tool marks (Giroux 2009). It is not reasonable to assume that the error rates in these different disciplines (with different amounts of information available to the examiners) would be similar; after all, studies which provide examiners with the full cartridge case instead of a sub-region of the case (such as the aperture shear or obturation marks) give examiners more information on which to make a decision, which should lower the error rate significantly.

Before we make direct comparisons between error rates in studies, we should also consider other factors beyond the type of comparison made which might make a difference in the reported error rate. Some open set studies include comparisons from marks with similar class characteristics as knowns in the set (Mattijssen et al. 2020; Pauw-Vugts et al. 2013), while others include comparisons with items that do not share class characteristics (this allows for examiners to eliminate based only on class characteristics, as some labs do not allow for elimination based only on individual characteristics) (Bunch and Murphy 2003). Thus, the level of difficulty of comparisons varies considerably between studies. In addition, the proportion of same source comparisons and different source comparisons varies considerably between studies: Baldwin et al. (2014) has about 1/3 same-source comparisons and 2/3 different-source comparisons, while Duez et al. (2018) has 3/4 same-source comparisons and 1/4 different-source comparisons. Obviously, the proportion of same-source and different-source comparisons may influence the precision with which any study can estimate certain error

[3]Todd Weller, California v Auimatagi, February 2021

rates.

The difficulty level across different studies is not necessarily comparable: in studies of European examiners, the goal is typically to differentiate different labs and procedures in skill and effectiveness, so the evaluations may be more difficult in order to separate good examiners and labs from those who are uniquely skilled; in these studies (Pauw-Vugts et al. 2013; Mattijssen et al. 2020), comparisons are typically harder and error rates much higher than in studies conducted in the US, where the study goal is typically to validate the process of firearms and toolmark evaluation.

8 Inconclusives

In fact, the firearms examination community is very aware of the fact that sometimes samples are difficult to match to each other. This issue is so pervasive that the AFTE theory of identification allows for a middle category between identification and elimination: inconclusive. Actually, more accurately, the AFTE theory of identification allows for 3 different levels of inconclusive on the continuum between identification and elimination. Under the AFTE theory of identification, all of the hard decisions are automatically correct because firearms and toolmark examiners don't have to make a decision between elimination and identification. The treatment of inconclusives under the AFTE theory of identification is controversial and has recently seen significant scrutiny from the scientific community [Dror and Scurich (2020); Dror and Langenburg (2019); Biedermann et al. (2019)][4].

Scientifically, an inconclusive result has to be automatically incorrect: a comparison is either from a same-source or a different-source. AFTE rules allow inconclusives to be counted as both identifications and eliminations, and therefore artificially decrease error rates. If we focus on a correct source decisions only, the percentage of correct decisions can be as low as 49%, leaving at least 51% of the decisions as errors (correct source identification rate taken from bullet comparisons in Pauw-Vugts et al. (2013)). This is statistically worse than random chance - that is, examiners would perform about as well if they were flipping a coin to make the decision!

Furthermore, when we examine which comparisons are more likely to result in an inconclusive decision, we find that the overwhelming majority of inconclusive decisions in black-box studies were made on different-source comparisons. This suggests that examiners are more willing to make an identification, and less willing to make an elimination when there is some doubt about the strength of the match. Black box studies conducted outside of the US legal system and forensic training system suggest that this bias is not as strong: Mattijssen et al. (2020), which was conducted on primarily European and UK examiners, did not reveal nearly as strong of a bias towards inconclusive decisions for different-source comparisons (though the scales used differ by jurisdiction). This bias is only obvious when we look at the positive and negative predictive values, as discussed in Hofmann, Vanderplas, and Carriquiry (2021).

The bias towards identification at the expense of elimination is in some cases encoded within the lab evaluation rules. In some laboratories, including at the FBI, examiners are not allowed to make an elimination based on individual characteristics. This is one reason why studies involving the FBI firearms lab often do not report sensitivity, specificity, or the correct source decision rate directly: when inconclusives are considered, the correct source decision rate looks abysmal: 52.22% in Bunch and Murphy (2003) and 37.5% in Giroux (2009).

9 Summary

We have outlined several problems with the state of error rate studies on firearm and toolmark examination. Fundamentally, we do not know what the error rate is for these types of comparisons. This is a failure of the scientific study of toolmarks, rather than the examiners themselves, but until this is corrected with multiple studies that meet the criteria described in Section 3, we cannot support the use of this evidence in criminal proceedings.

[4]CTS used to treat inconclusives as errors, but in 1998 changed to treating inconclusives as correct decisions. The error rates dropped from 12% (firearms) and 26% (toolmarks) to approximately 1.4% and 4% respectively. UNITED STATES OF AMERICA, v. Joseph MINERD, Defendant., 2002 WL 32995663 (W.D.Pa.)

If inconclusives are errors, even if these errors are not necessarily attributable to the examiners' skill level, then examiners make the correct decision in black-box tests of screwdriver and bullet evaluations at rates that are worse than if they used coin flips to determine their answers (37.5% and 49.68%, respectively). Correct source decision rates for many studies mentioned in this assessment are provided in Appendix A. Even if inconclusives are considered correct decisions, as recommended by the Association of Firearms and Toolmark Examiners, however, there are still other problems with this evidence.

The error rates on different types of toolmarks are substantially different, and there are not multiple studies that even allow the estimation of the necessary error and accuracy rates (false positive rate, false negative rate, and correct source decision rate) for most types of evidence. Multiple studies are necessary because any set of participants and required comparisons may be non-representative in some way. Science thrives on replication of studies in slightly different conditions over a period of time; firearms and toolmark examination is no exception. In only one discipline (cartridge case evaluation) are there more than two studies which are designed in such a way that the full set of error rates are estimable. In these studies, overall error rates (counting inconclusives as correct decisions) are between 0% and 8.21%. While these error rates are relatively low, the studies these error rates are based on still have fundamental flaws that suggest that the true error rates in casework may be significantly higher.

One of the major issues with the studies that do exist is that they are plagued by non-response bias. In many disciplines, non-response biases of greater than 20% are sufficient to invalidate study results, and rates greater than 5% are sufficient cause for concern. Most studies in this discipline do not report non-response rates. Few studies report sufficient details to begin to estimate the non-response rates, as seen for the studies reviewed in Appendix B. For these studies, where sufficient details were reported, drop-out rates are up to 35% and item non-response are up to 17%. That most studies do not report drop-out rates is statistical malpractice; failure to report or address this information when reported is particularly egregious in studies that have practicing statisticians as authors. While it is theoretically possible that examiners who drop out of studies or leave missing items are similar to examiners who complete the studies (leading to an unchanged error rate), it is significantly more likely that if these examiners had completed the study, the error rate would have been higher. In the worst case scenario, the actual error rate could be as high as the sum of the dropout rate and the reported error rate. We stress again that there is no possibility of assessing the true impact of non-response bias in these studies when the authors do not make their data available to other researchers: as was the case in every study reviewed in Appendix A.

Even when non-response bias is ignored, there are further issues with the state of error rate studies in firearm and toolmark examination: studies do not cover a wide enough range of firearms and ammunition to generalize to the discipline as a whole. We know that not all firearm and ammunition combinations mark equally well, but studies are conducted using firearms and ammunition that are known to mark. As a result, there is a potential difference between the firearms and ammunition used in studies and those evaluated as part of casework that impacts our ability to generalize error rates from studies to a specific combination of ammunition and firearm in a particular case. Similarly, the firearms examiners who participate in studies are likely to be fundamentally different from the full population of firearms examiners. This, too, means that we cannot generalize error rates from studies conducted on volunteers the entire set of qualified firearms examiners. What we can say with confidence is that the sampling issues with firearms examination studies are problematic and cast doubt on the ability to generalize error rates from these studies to the wider population of all examiners.

As a result of these compounding issues, it is our opinion as statisticians and researchers in forensic science that error rates established from studies with sampling flaws, methodological flaws, non-response and attrition bias, and inconclusive results are not reliable estimates of the discipline-wide error rate.

10 References

Baldwin, David P., Stanley J. Bajic, Max Morris, and Daniel Zamzow. 2014. "A Study of False-Positive and False-Negative Error Rates in Cartridge Case Comparisons:" Fort Belvoir, VA: Defense Technical Information Center. https://doi.org/10.21236/ADA611807.

Biedermann, Alex, Silvia Bozza, Franco Taroni, and Joëlle Vuille. 2019. "Are Inconclusive Decisions

in Forensic Science as Deficient as They Are Said to Be?" *Frontiers in Psychology* 10 (March): 520. https://doi.org/10.3389/fpsyg.2019.00520.

Bunch, Stephen, and Douglas Murphy. 2003. "A Comprehensive Validity Study for the Forensic Examination of Cartridge Cases." *AFTE JOURNAL* 35 (2): 201–3.

Chapnick, Chad, Todd J. Weller, Pierre Duez, Eric Meschke, John Marshall, and Ryan Lilien. 2021. "Results of the 3D Virtual Comparison Microscopy Error Rate (VCMER) Study for Firearm Forensics." *Journal of Forensic Sciences* 66 (2): 557–70. https://doi.org/10.1111/1556-4029.14602.

Chumbley, L Scott, Max D Morris, Stanley J Bajic, Daniel Zamzow, Erich Smith, Keith Monson, and Gene Peters. 2021. "Accuracy, Repeatability, and Reproducibility of Firearm Comparisons Part 1: Accuracy." *arXiv Preprint arXiv:2108.04030*.

Dror, Itiel E., and Glenn Langenburg. 2019. "'Cannot Decide': The Fine Line Between Appropriate Inconclusive Determinations Versus Unjustifiably Deciding Not To Decide." *Journal of Forensic Sciences* 64 (1): 10–15. https://doi.org/10.1111/1556-4029.13854.

Dror, Itiel E., and Nicholas Scurich. 2020. "(Mis)use of Scientific Measurements in Forensic Science." *Forensic Science International: Synergy*, September. https://doi.org/10.1016/j.fsisyn.2020.08.006.

Duez, Pierre, Todd Weller, Marcus Brubaker, Richard E. Hockensmith, and Ryan Lilien. 2018. "Development and Validation of a Virtual Examination Tool for Firearm Forensics, ," *Journal of Forensic Sciences* 63 (4): 1069–84. https://doi.org/10.1111/1556-4029.13668.

Giroux, Brandon N. 2009. "Empirical and Validation Study: Consecutively Manufactured Screwdrivers." *AFTE Journal* 41 (2): 153–66.

Gouwe, J, JE Hamby, and SA Norris. 2008. "Comparison of 10,000 Consecutively Fired Cartridge Cases from a Model 22 Glock. 40 S&W Caliber Semiautomatic Pistol." *AFTE JOURNAL* 40 (1): 57.

Hamby, James E., David J. Brundage, Nicholas D. K. Petraco, and James W. Thorpe. 2019. "A Worldwide Study of Bullets Fired From 10 Consecutively Rifled 9MM RUGER Pistol Barrels—Analysis of Examiner Error Rate." *Journal of Forensic Sciences* 64 (2): 551–57. https://doi.org/10.1111/1556-4029.13916.

Hofmann, Heike, Susan Vanderplas, and Alicia Carriquiry. 2021. "Treatment of inconclusives in the AFTE range of conclusions." *Law, Probability and Risk* 19 (3–4): 317–64. https://doi.org/10.1093/lpr/mgab002.

Keisler, M. A., Hartman, S., and Kil, A. 2018. "Isolated Pairs Research Study." *AFTE Journal* 50 (1): 56–58.

Lightstone, L. 2010. "The Potential for and Persistence of Subclass Characteristics on the Breech Faces of SW40VE Smith and Wesson Sigma Pistols." *AFTE Journal* 42 (4): 308–22.

Lyons, Dennis. 2009. "The Identification of Consecutively Manufactured Extractors." *AFTE Journal* 41 (3): 246–56. https://unl.illiad.oclc.org/illiad/illiad.dll?Action=10&Form=75&Value=1315162.

Mattijssen, Erwin J. A. T., Cilia L. M. Witteman, Charles E. H. Berger, Nicolaas W. Brand, and Reinoud D. Stoel. 2020. "Validity and Reliability of Forensic Firearm Examiners." *Forensic Science International* 307 (February): 110112. https://doi.org/10.1016/j.forsciint.2019.110112.

Mayland, Brian, and Caryn Tucker. 2012. "Validation of Obturation Marks in Consecutively Reamed Chambers." *AFTE Journal* 44 (2): 167–69.

Mayland, B, and C Tucker. 2012. "Validation of Obturation Marks in Consecutively Reamed Chambers." *AFTE Journal* 44 (2): 167–69.

National Research Council (U.S.), ed. 2009. *Strengthening Forensic Science in the United States: A Path Forward*. Washington, D.C: National Academies Press.

Neel, M, and M Wells. 2007. "A Comprehensive Statistical Analysis of Striated Tool Mark Examinations Part 1: Comparing Known Matches and Known Non-Matches." *AFTE JOURNAL* 39 (3): 176.

Pauw-Vugts, P, A Walters, L Øren, and L Pfoser. 2013. "FAID 2009: Proficiency Test and Workshop." *AFTE Journal* 45 (2): 115–27.

President's Council of Advisors on Science and Technology. 2016. "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature Comparison Methods." https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.

Riva, Fabiano, Rob Hermsen, Erwin Mattijssen, Pascal Pieper, and Christophe Champod. 2017. "Objective Evaluation of Subclass Characteristics on Breech Face Marks." *Journal of Forensic Sciences* 62 (2): 417–22. https://doi.org/10.1111/1556-4029.13274.

Schulz, Kenneth F., and David A. Grimes. 2002. "Sample Size Slippages in Randomised Trials: Exclusions and the Lost and Wayward." *Lancet (London, England)* 359 (9308): 781–85. https://doi.org/10.1016/S0140-6736(02)07882-0.

Smith, Jaimie A. 2021. "Beretta Barrel Fired Bullet Validation Study." *Journal of Forensic Sciences* 66 (2): 547–56.

Smith, Tasha P., G. Andrew Smith, and Jeffrey B. Snipes. 2016. "A Validation Study of Bullet and Cartridge Case Comparisons Using Samples Representative of Actual Casework." *Journal of Forensic Sciences* 61 (4): 939–46. https://doi.org/10.1111/1556-4029.13093.

Stodden, Victoria. 2015. "Reproducing Statistical Results." *Annual Review of Statistics and Its Application* 2: 1–19.

Stroman, Angela. 2014. "Empirically Determined Frequency of Error in Cartridge Case Examinations Using a Declared Double-Blind Format." *AFTE Journal* 46 (2): 157–74.

Thomas G. Fadul, Jr, Gabriel A. Hernandez, Stephanie Stoiloff, and Sneh Gulati. 2013. "An Empirical Study to Improve the Scientific Foundation of Forensic Firearm and Tool Mark Identification Using 10 Consecutively Manufactured Slides." *AFTE Journal* 45 (4): 376–89. https://unl.illiad.oclc.org/illiad/illiad.dll?Action=10&Form=75&Value=1315164.

Wicherts, Jelte M, Marjan Bakker, and Dylan Molenaar. 2011. "Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results." *PloS One* 6 (11): e26828.

A Table of Reported Correct Source Decision Rates

Some Abbreviations:

- SS: Same Source comparisons
- DS: Different Source comparisons
- LEA: Land Engraved Area

Table 1: Correct source decision rates for many commonly-cited studies, along with number of evaluations and type of mark evaluated in the study.

Type of Mark	Study	# Evaluations	Correct Source Decision Rate
Obturation	Brian Mayland and Tucker (2012)	192 SS, ? DS	Not calculable
Extractor	Lyons (2009)	178 SS, ? DS	Not calculable
Aperture Shear	Mattijssen et al. (2020)	2947 SS, 1673 DS	0.6816
Screwdriver	Giroux (2009)	29 SS, 51 DS	0.3750
Bullet LEA	Hamby et al. (2019)	10455 SS, ? DS	Not calculable
Bullet LEA	Pauw-Vugts et al. (2013)	188 SS, 124 DS	0.4968
Bullet LEA	T. P. Smith, Smith, and Snipes (2016)	219 SS, 621 DS	0.8024
Cartridge Cases	Stroman (2014)	75 SS, ? DS	Not calculable
Cartridge Cases	Bunch and Murphy (2003)	70 SS, 290 DS	0.5222
Cartridge Cases	Pauw-Vugts et al. (2013)	127 SS, 189 DS	0.6867
Cartridge Cases	Chapnick et al. (2021)	491 SS, 693 DS	0.7508
Cartridge Cases	Baldwin et al. (2014)	1090 SS, 2180 DS	0.7633
Cartridge Cases	Keisler, M. A., Hartman, S., and Kil, A. (2018)	1512 SS, 1008 DS	0.9179
Cartridge Cases	T. P. Smith, Smith, and Snipes (2016)	199 SS, 437 DS	0.9324
Cartridge Cases	Duez et al. (2018)	336 SS, 112 DS	0.9375

Studies that have an unknown number of different source comparisons cannot be analyzed to produce a correct source decision rate because the total number of comparisons performed (and the number of correct eliminations) cannot be computed. These studies are usually closed-set designs, which have been cited as problematic in President's Council of Advisors on Science and Technology (2016). Totals from T. P. Smith, Smith, and Snipes (2016) are recreated using reported marginal totals and internal counts; the study results are internally inconsistent, so the approach with the least obfuscation and adjustment was taken for simplicity. The discrepancy in T. P. Smith, Smith, and Snipes (2016) also stems from the study design's inability to determine how many comparisons were conducted, but it is at least designed in a way that allows for the calculation of the minimal number of necessary comparisons.

B Table of Participant Sampling, Data Availability, Drop-out Rates, and Item Non-Response Rates

Drop-out Rate: Proportion of examiners who agreed to participate in the study and were not included in the final analysis. "Unreported"" indicates the authors did not provide sufficient information to calculate this number.
Item Non-Response Rate: This is a conservative measure. We are calculating the proportion of items missing only for participants who are included in the final analysis. Note this is a lower bound for the item non-response as it does not account for the items not responded to by participants who dropped out. "Unreported" indicates the authors did not provide sufficient information to calculate this number.

Table 2: Participant sampling types, data availability, drop-out rates, and item non-response rates. Only studies for which one of drop-out rates and item non-response rates are specified are included. This table focuses on data used to calculate accuracy of examiners conclusions.

Study	Volunteer Participants	Data Publicly Available	Drop-out Rate	Item Non-Response Rate
Lyons (2009)	Yes	No	Unreported	0%
B. Mayland and Tucker (2012)	Yes	Partially [5]	Unreported	0%
Thomas G. Fadul, Jr et al. (2013)	Yes	No	23% [6]	0%
Stroman (2014)	Yes	No	17%	Unreported
Baldwin et al. (2014)	Yes	No	23%	0.06%
T. P. Smith, Smith, and Snipes (2016)	Yes	No	34%	Unreported
Keisler, M. A., Hartman, S., and Kil, A. (2018)	Yes	No	Unreported	0%
Hamby et al. (2019)	Yes	No	Unreported	Unreported
Chapnick et al. (2021)	Yes	No	$\geq$ 29% [7]	3%
Chumbley et al. (2021)	Yes	No	Unreported	17%
J. A. Smith (2021)	Yes	No	35%	Unreported

[5]The authors did report the participant answers by test.
[6]This reflects the proportion of individuals who completed the test sets but were excluded from analysis because they had not had two years of training.
[7]The authors did not report the number of participants who agreed to participate. 107 participants completed some the test sets. The authors excluded all but 76 of them from analysis for the results reported in the abstract due to the participant being "unqualified" or working outside of the United States or Canada. Note, other studies have explicitly included examiners outside of the United States and Canada (e.g., Hamby et al. (2019) and Keisler, M. A., Hartman, S., and Kil, A. (2018)). The authors indicated the excluded participants had committed more errors than those included.

EXHIBIT 3

Methodological Problems in Every Black-Box Study of Forensic Firearm Comparisons

Maria Cuellar[a,b,f], Susan Vanderplas[d,f], Amanda Luby[c,f], and Michael Rosenblum[e]

[a]Department of Criminology, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA, 19104, United States

[b]Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Walnut Street, Philadelphia, PA 19104, United States

[c]Department of Mathematics and Statistics, Swarthmore College, 500 College Avenue, Swarthmore, PA 19081

[d]Department of Statistics, University of Nebraska-Lincoln, 340 Hardin Hall North Wing Lincoln, NE 68583-0963

[e]Department of Biostatistics, Johns Hopkins University, Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, MD 21205

[f]Center for Statistics and Applications in Forensics Evidence (CSAFE), Iowa State University, 613 Morrill Road, Ames, IA, 50011, United States

March 27, 2024

Abstract

Reviews conducted by the National Academy of Sciences (2009) and the President's Council of Advisors on Science and Technology (2016) concluded that the field of forensic firearm comparisons has not been demonstrated to be scientifically valid. Scientific validity requires adequately designed studies of firearm examiner performance in terms of accuracy, repeatability, and reproducibility. Researchers have performed "black-box" studies with the goal of estimating these performance measures. As statisticians with expertise in experimental design, we conducted a literature search of such studies to date and then evaluated the design and statistical analysis methods used in each study. Our conclusion is that all studies in our literature search have methodological flaws that are so grave that they render the studies invalid, that is, incapable of establishing scientific validity of the field of firearms examination. Notably, error rates among firearms examiners, both collectively and individually, remain unknown. Therefore, statements about the common origin of bullets or cartridge cases that are based on examination of "individual" characteristics do not have a scientific basis. We provide some recommendations for the design and analysis of future studies.

1 Introduction

The National Academy of Sciences Report "Strengthening Forensic Science in the United States: A Path Forward" (NAS, 2009) expressed a critical need for research aimed at establishing the scientific foundations of forensic science. Among other conclusions about firearms examination, it stated that "sufficient studies have not been

done to understand the reliability and repeatability of the methods." (NAS, 2009, p.154). The President's Council of Advisors on Science and Technology (PCAST, 2016) evaluated the scientific validity of some forensic methods, including firearms examination, that are used in criminal courts. They similarly concluded that the scientific validity of firearms examination has not been established.

PCAST (2016) explained why empirical studies to evaluate examiner performance, called validation studies, are of critical importance: "Without appropriate estimates of accuracy, an examiner's statement that two samples are similar–or even indistinguishable–is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact. Nothing–not training, personal experience nor professional practices–can substitute for adequate empirical demonstration of accuracy." (PCAST, 2016, p.46). We agree, since a fundamental principle of sound science is the need for demonstration by repeated experiments (NAS, 1992, p.38). The importance of empirical studies is also highlighted by the U.S. National Institute of Standards and Technology (NIST) in their report on evaluating scientific foundations of forensic methods (Butler et al., 2020). This motivates our focus on validation studies.

In addition to the aforementioned NAS and PCAST reports, there is a substantial literature describing problems with the study design and statistical analysis methods in validation studies of firearms examination, e.g., Spiegelman and Tobin (2012); Dror and Scurich (2020); Hofmann et al. (2020); Dorfman and Valliant (2022a,b); Scurich (2022); Khan and Carriquiry (2023a,b); Scurich et al. (2023); Rosenblum et al. (2024), among others. Our contribution builds on this prior work. We conducted a literature

search of validation studies (focusing on black-box studies) of firearms examination to date, and then evaluated the design and statistical analysis methods used in each. We identified serious methodological flaws that are common to all such studies, with minor exceptions (see Table 1 of Section 4). Most of these flaws were presented in the aforementioned prior work for a subset of studies; we show that these flaws are not the exception, but instead are generally the rule across such studies. Our search also revealed new flaws. We present this critique constructively with the goal of raising awareness of how to improve future studies.

In the next section, we explain why sound experimental design and statistical analysis methodology used in validation studies of medical diagnostic tests are applicable to validation studies of firearms examination. In Section 3 we describe our literature search of firearms examination validation studies. Study design and analysis flaws are summarized in Section 4 and explained in detail in Section 5. We end with a discussion in Section 6.

2 Relevance of research and evaluation methods used in medical diagnostic testing

The 2009 National Academies of Sciences report (NAS, 2009) summarized the need for scientifically sound evaluations of forensic analysis methods, including firearms examination. It highlighted the relevance of the sound research and evaluation methods used in medical diagnostic testing, in the following passage:

> A body of research is required to establish the limits and measures of

> performance and to address the impact of sources of variability and potential bias. Such research is sorely needed, but it seems to be lacking in most of the forensic disciplines that rely on subjective assessments of matching characteristics. These disciplines need to develop rigorous protocols to guide these subjective interpretations and pursue equally rigorous research and evaluation programs. **The development of such research programs can benefit significantly from other areas, notably from the large body of research on the evaluation of observer performance in diagnostic medicine...** (NAS, 2009, p.8)

We bolded the last sentence to highlight the relevance of medical diagnostic tests. There are many similarities between diagnostic testing involving human observers, e.g., breast cancer diagnostic testing using mammography, and firearms examination. In both mammography and firearms examination, a human observer examines images and based on their experience and training makes a subjective decision about it. In mammography an examiner decides whether the images (taken from different angles and possibly magnified) indicate cancer or not, while in firearms examination an examiner decides whether bullets/cartridge cases (using magnified images) were fired from the same or different guns. Both involve subjective analysis followed by a yes/no decision (with a possibility for inconclusive results).

Firearms examiners have drawn parallels between diagnostic testing and their field, e.g., Bunch and Murphy (2003, p.203), state that "For data analysis purposes, a forensic comparison examination can be considered analogous to a clinical test such as a blood test." We therefore refer to published guidance documents from the U.S.

Food and Drug Administration and the Clinical and Laboratory Standards Institute on how to design and analyze validation studies. Their recommendations are applicable to any diagnostic test procedure, including ones that involve human observation and subjective decisions, and therefore can be applied to firearms examination.

3 Selection of black-box studies

Firearms examination is a subjective method, i.e., key parts of the comparison rely on human judgment. Because of this, validation studies need to be black-box, i.e., they ask firearm examiners to make decisions about a series of test cases where the experimenter (but not the firearm examiners) knows the ground truth, and performance is measured by comparing firearm examiners' decisions to the ground truth across test cases. These are called "black-box" because they measure whether decisions are correct or not, while treating the examiner's internal decision process (which is difficult to quantify due to its subjective nature) as a "black-box". We use the term "black-box studies" to refer to black-box validation studies.

We compiled our list of black-box studies of firearms examination (see list in Appendix A) from multiple sources. Our criteria for including a black-box study in our list in Appendix A are the following: (i) it includes human examiners solving test cases involving fired cartridge cases and/or bullets with known ground truth; (ii) it is not a pilot/preliminary/exploratory study.

We first searched for studies by using specific terms (i.e., "validation", "firearms", "examination", and "comparison") in the Association of Firearms and Toolmarks

Examiners (AFTE) Journal Article Index. We then considered studies from the Interpol review papers covering the period 2012-2022; that publication compiles validation studies of firearms examination (among other forensic methods) from the following journals: Australian Journal of Forensic Sciences, Forensic Science International, Forensic Science International: Synergy, Forensic Sciences Research, Journal of Forensic Identification, Journal of Forensic Sciences, Science and Justice.

We also considered the studies listed in the PCAST (2016) report and in a response[1] to this report by the Firearms and Toolmarks Subcommittee of the Organization of Scientific Area Committees (OSAC) at the U.S. National Institute of Standards and Technology (NIST) (2016). We conferred with a practicing firearms examiner to ask for additional studies. Each author of this manuscript included any studies that they are familiar with through previous work on the scientific validity of firearms examination.

A total of 28 different studies meeting our inclusion criteria were obtained. These are listed in Appendix A. (There are 33 citations there, but some refer to the same study.)

4 Evaluation of study designs and statistical analyses

A validation study needs to have sound experimental design and statistical analysis principles to provide reliable results. Some standard principles include the following:

[1]This response was not endorsed by OSAC nor NIST, but in an effort to be as inclusive of potential black-box studies as possible, we considered the studies that it listed.

selecting participants and materials that are representative of the full spectrum of real casework; including sufficient numbers of such participants and materials to answer the scientific question with the desired level of precision; estimating error rates correctly along with valid confidence intervals; and addressing missing data with appropriate statistical methods. With minor exceptions described below, none of the 28 studies adhered to any of these key principles. Although it can be challenging for validation studies of firearms examination to adhere to all of these principles, it is possible and is necessary in order to produce reliable results. The challenge of adhering to all of these principles is faced in other domains, e.g., in medical-imaging validation studies, where these challenges have been successfully met.

Table 1 lists study design and analysis flaws that we found in every study in Appendix A, with some exceptions for flaws C and E indicated by footnotes. It is not an exhaustive list. Each of the methodological flaws B-E in Table 1 is so consequential that having even one such flaw renders a validation study scientifically unsound. Also, flaw A is important because it exacerbates the other flaws, as described below. Most of the flaws in Table 1 have been identified for at least some studies in prior work, as described in Section 1. Our contribution is to summarize the flaws that are both common and consequential across the black-box studies in Appendix A. Because these flaws are shared across studies, it is not possible to combine results from the studies to overcome the flaws. The only way to demonstrate scientific validity is to conduct future, adequately designed and analyzed studies.

[2]One study had no inconclusive responses and some studies did not report a study-wide error rate. Some studies computed error rates in multiple ways, but all are flawed.

[3]Some of the smaller studies had no missing data.

Flaw	Impact
A. Inadequate sample size: No sample size calculation was done to determine how many firearms, examiners, and bullets/cartridge cases are needed to achieve the study goals. The sample size calculation is "one of the most important parts of any experimental design problem." (Montgomery, 2017, p.44).	Insufficient number of firearms are used, leading to low precision.
B. Non–representative sample: Study conditions, materials (firearms and ammunition), and participants (examiners) are not representative of the full spectrum of real casework.	Results not applicable to real casework.
C. Error rates incorrectly computed because inconclusive responses are essentially treated as correct or ignored.[2]	Error rates may be underestimated.
D. Invalid confidence intervals for error rates. Confidence intervals do not account for the structure of study designs (multiple examiners and multiple firearms, which leads to statistical dependence), or no confidence interval is provided.	Reported range of plausible error rates in studies is too small or is unknown
E. Missing data.[3] Missing data, such as item and unit non–response, are not handled with appropriate statistical methods.	Error rates may be biased.

Table 1: Flaws with the study design and statistical analysis that occur in every black-box study that we analyzed, along with the impact of each flaw.

Prior work has identified some of the flaws. For example, Spiegelman and Tobin (2012) found flaws A, B, D, and E, as well as other flaws not on our list, in several firearms examination studies. Hofmann et al. (2020) study the treatment of inconclusive responses in error rate calculations (flaw C). Khan and Carriquiry (2023a,b) consider some large validation studies and show that they have high rates of "missingness", i.e., unit and item nonresponse (flaw E).

5 Flaws with study design and statistical analysis

5.1 Flaw A: Sample size

Contrary to standard practice in the experimental design of validation studies, none of the black-box studies in Appendix A carried out the important step of computing the sample size, i.e., the number of examiners, the number of firearms of each type, the number of ammunition types, and the number of bullets/cartridge cases required to achieve the study goals. According to the textbook *Design and Analysis of Experiments* by Montgomery (2017, p.44), "Selection of an appropriate sample size is one of the most important parts of any experimental design problem." This requirement is stated in FDA guidelines for validating diagnostic tests (FDA, 2007, p.42), which states that "The study should be of sufficient size to evaluate the candidate test based on the desired level of uncertainty." The sample size needs to be determined prior to starting the study and is computed in order to provide a high likelihood of achieving the study's goals. Without a sample size calculation, there is no reason to expect that a validation study will be able to do this.

The Human Factors Committee of the Organization of Scientific Area Committees (OSAC) for Forensic Science (2020, p.13) also acknowledges the need for a sample size calculation when it states that, "To properly assess the accuracy of a method at any difficulty level, it is important to include adequate numbers of test specimens at that level." It also recommends that, "Statisticians or other experts familiar with statistical power and sample size requirements for experimental research can assist in determining the appropriate numbers of test specimens and examiners for validation

and reliability studies." As statisticians, we agree with these statements. Skipping this important step can lead to studies that fail to estimate accuracy precisely. This occurred in all of the studies in Appendix A. Not doing a sample size calculation may be acceptable in pilot studies or exploratory studies, but not in validation studies.

Kerkhoff et al. (2018, p.262) state that, for their firearms examination study, "The sample size was relatively small (53 uncorrected, 39 corrected conclusions) and a much larger sample would be needed to get a good estimate of the rate of misleading evidence in practice." We applaud Kerkhoff et al. (2018) for clearly stating the limitations of their study and for stressing the need for adequate sample size and proper study design more generally. The quote above explains why we did not include the studies of Kerkhoff et al. (2015, 2018); Neuman et al. (2022) in our list of black-box studies, i.e., they are exploratory studies. Such studies are valuable for hypothesis generation and for informing the planning of future studies; however, they are not appropriate for validation.

The sample size calculation depends on what type of experimental design is used. Almost every black-box study in Appendix A uses a design where one or more test cases from each firearm (or firearm pair) is separately evaluated by multiple examiners, and similarly each examiner evaluates multiple test cases. In such designs, a correct sample size calculation needs to simultaneously account for (i) dependence among test cases evaluated by the same examiner and (ii) dependence among test cases involving the same firearm(s) (FDA, 2007; Montgomery, 2017). This general setup is called a multi-reader, multi-case (MRMC) design, and there is a rich statistical literature on how to appropriately analyze it, e.g., Zhou et al. (2009); Gallas

et al. (2009).

Zhou et al. (2009, pp.219–228) provides detailed guidance about how to calculate sample sizes in MRMC studies, while appropriately handling dependencies that arise from the experimental design. The software package iMRMC in R by Gallas (2023) does sample size calculations, is free to download, and is provided by the FDA. This tool can be used to determine how many firearms, ammunition types, examiners, and test cases are needed.[4]

5.2 Flaw B: Representative samples

For a set of studies to assess the accuracy, repeatability, and replicability of a pattern-comparison method such as firearms comparison, it is critical to include representative samples, i.e., a set of participants, materials, and test conditions that are representative of the full spectrum encountered in real casework. As we explain below, none of the studies that we reviewed gives any evidence that its conditions, participants, and materials (firearms and ammunition) are representative of the full spectrum of real casework. But first we consider statements by several scientific groups about the importance of having a representative sample.

The American Statistical Association (ASA) is the largest community of statisticians in the world and includes members from over 90 countries working in government, academia, and industry; their goal is "promoting sound statistical practice" (ASA, 2024). According to the *ASA Position on Statistical Statements for Forensic*

[4]If the estimated false positive rate is close to 0 or 1, one may need to use exact methods to construct confidence intervals, e.g., generalizing the Clopper-Pearson method described in Appendix B.

Evidence, "To be applicable to casework, rigorous empirical studies of the reliability and accuracy of forensic science practitioners' judgments must involve materials and comparisons that are representative of evidence from casework" (ASA, 2024). Thus, the ASA also agrees that representation is important.

We next consider the report *NIST Scientific Foundation Reviews* by the the National Institute of Standards and Technology (Butler et al., 2020, p.2). The report describes key principles and criteria for evaluating studies of the scientific validity of forensic methods. One of the key criteria is the following: "Are test results fit for purpose (i.e., are they expected to reflect performance under casework conditions)?" In other words, a representative sample is needed. The National Institute of Forensic Science of Australia/New Zealand (2019, p.6) also provides a checklist that includes this point.

Monson et al. (2022, pp.7-8) state that "To detect a small or rare effect (examiner errors in the present case) one must make a large number of observations within a representative population." Though we disagree that examiner error rates have been demonstrated to be small or rare, we do agree that in order to have any opportunity to demonstrate this, one needs a large number of observations from a representative population. This is especially important since some firearm types produce harder/easier to classify toolmarks, and so assessing a relatively small number of firearm types may lead to conclusions that don't generalize to what's commonly encountered in real casework.

The FDA warns about the dangers of not including a representative sample. The "FDA recommends the set of subjects and specimens to be tested include:

subjects/specimens across the entire range of disease states. . . " and that "If the set of subjects and specimens to be evaluated in the study is not sufficiently representative of the intended use population, the estimates of diagnostic accuracy can be biased." (FDA, 2007, p.18).

A prerequisite to conducting experiments to evaluate reliability of firearms examination is to first characterize the full spectrum of firearms (and firearms manufacturing techniques) and ammunition known to be used in crimes, as best possible given available data sources, or new data sets should be created for this purpose. Only after this is done can studies be designed that include representative samples of firearm types. An analogous issue arises for ammunition and for examiners, e.g., with respect to level of training and to protocols used in different crime labs.

Khan and Carriquiry (2023b) consider how representative the sample of firearm examiners who participated in the study of Monson et al. (2023) are with respect to characteristics that may affect error rates, such as years of expertise, laboratory protocols, and amount of training. They show that this sample of examiners is not representative of the population of examiners, which may lead to underestimating error rates.

It is not necessary to include every possible firearm, ammunition type, or examiner. What is needed is a representative sample, which can be obtained by first characterizing the full spectrum of firearms, ammunition, and examiners and then taking a random or systematic sample from these. This is not easy, but it is a necessary step since otherwise it is not possible to determine whether a sample is representative of the full spectrum of real casework.

Study conditions should also be representative of real casework, so that study results generalize to real casework conditions. A discrepancy between study and casework conditions is that examiners in a study know that their answers will be compared to known ground truth. This creates a potential motivation to modify behavior, e.g., they may spend more time analyzing the samples and/or may be less likely to make an identification decision. Changes in behavior could lead to error rates in studies that are not the same as the error rates in real casework (Dror and Scurich, 2020). One way to partially address this is to conduct studies that insert test cases with known ground truth into the workflow of real casework in a blinded manner (i.e., without informing examiners which samples are test cases or not). This has been pilot tested in quality control programs (Kerkhoff et al., 2018; Neuman et al., 2022). Mejia et al. (2020) consider the logistical challenges in implementing such blinded programs and provide recommendations on how to overcome them.

Another discrepancy between conditions in studies and real casework is the presence of contextual information in the latter that may bias decision making (Kassin et al., 2013; Mattijssen et al., 2016). Contextual information could include, for example, other evidence in a case. According to the aforementioned related work, such contextual bias may be reduced in real casework if examiners are blinded to all information about a case other than what is needed to compare bullets/cartridge cases.

Additional factors have potential to impact the difficulty of classifying toolmarks; see e.g., the list of factors in (Spiegelman and Tobin, 2012, Section 1.2). For example, the condition of firearms (e.g., newly manufactured versus damaged/corroded) may

impact the difficulty of classifying toolmarks, yet this has not been systematically varied and analyzed in any black-box study, to the best of our knowledge. None of the black-box studies in the Appendix reports the use of damaged/corroded firearms for generating test cases. Therefore, it's unknown whether the error rates from these studies would apply to such firearms.

5.3 Flaw C: Treatment of inconclusives

In firearms comparisons, examiners assess whether a pair of samples (bullets/cartridge cases) were fired from the same firearm. Roughly speaking, they select one of three conclusions for each pair: identification, elimination, or inconclusive.[5] A black-box study of sample comparisons should report the false positive rate, the false negative rate, the true positive rate, the true negative rate – alternatively, the sensitivity and specificity. It is not sufficient to only report the false positive rate, for example. In the extreme case, all the participants in a study could mark all pairs as eliminations – this would result in a zero percent false-positive rate, but it would be trivially uninformative. In addition, the study should report the treatment of inconclusives.

In most black-box studies that we reviewed, the "inconclusive" responses are effectively treated as correct or ignored when computing error rates[6]. Both approaches

[5]Different subcategories are also possible, as well as being "unsuitable" for evaluation; also, some laboratories report likelihood ratios instead of categorical responses. We focus on the three aforementioned conclusions for clarity of explanation.

[6]Some studies included additional ways to treat inconclusives. Law and Morris (2021) present results treating inconclusives as correct, incorrect, and using a consensus-based scoring approach; they do not report a study-wide error rate. Some studies either did not report error rates but rather number of decisions (Hamby et al., 2009), while other studies resulted in no inconclusives (Fadul, 2011; Cazes and Goudeau, 2013). Mattijssen et al. (2020, 2021) used a 'forced choice' design and reported error rates both including and excluding inconclusives. None of the aforementioned methods, however, adequately handles how to account for inconclusive responses when computing

(treating "inconclusives" as correct or ignoring them) are invalid and can lead to underestimation of error rates. The first approach artificially reduces the error rates because inconclusive responses get counted in the denominator but not the numerator when computing each error rate. For both approaches, there may be an incentive for examiners to opt out of the most challenging test cases by responding "inconclusive", knowing that this will either decrease or have no impact on the study's error rate.

Hofmann et al. (2020) provide an in-depth exploration about inconclusives and their different treatments in black-box studies. They find that the rates of inconclusives may vary depending on the norms for training and reporting in different regions. They also find evidence that the inconclusive rate is higher for different-source than same-source samples.

In the context of evaluating diagnostic tests, a guidance from the FDA (2007, p.20) lists four statistical practices that it calls "inappropriate". The second "inappropriate" practice on this list is to "discard equivocal new test results when calculating measures of diagnostic accuracy or agreement". In the context of evaluating the accuracy of firearms examination, this means that "inconclusives" should not be ignored nor omitted when computing error rates, as is done in some of the studies in Appendix A.

The aforementioned FDA guidance also has an entire section on how to handle "inconclusive" results (also called "equivocal" or "indeterminate" results) titled: "Avoid elimination of equivocal results", that recommends not to ignore inconclusives

study-wide error rates.

(FDA, 2007, p.18). There, the FDA describes a procedure to handle "equivocal" results, which is to report error rates in the following two ways: first, set all "equivocal" responses to "positive" and compute error rates; second, set them all to "negative" and recompute error rates.

The corresponding procedure in the context of firearms examination would be to set all "inconclusive" responses to "identification" and compute error rates; second, set them all to "elimination" and recompute error rates. We recommend this procedure since it gives estimated lower and upper bounds on the error rates. Applying this procedure to the cartridge case data from Monson et al. (2023), the corresponding false positive error rates are 51.4% and 0.92%, respectively; the corresponding false negative error rates are 1.76% and 25.6%, respectively. This is for illustration only, temporarily setting aside the other flaws in Table 1. All such flaws need to be addressed by future studies in order to produce valid results.

Another measure of accuracy is the likelihood ratio, which is discussed and computed by Guyll et al. (2023). Computing a study-wide likelihood ratio could be a useful approach, as discussed by Taroni and Biedermann (2005); Thompson et al. (2013); Robertson et al. (2016); Champod et al. (2016); Lund and Iyer (2017); Ommen and Saunders (2018). However, because the study of Guyll et al. (2023) has flaws A, B, D, and E, the reported likelihood ratios are invalid. For additional methodological flaws in that study, please see (Rosenblum et al., 2024).

Another related issue is conformance, i.e., the need for examiners in black-box studies to adhere to a clearly defined protocol. A lack of conformance has been described by Baldwin et al. (2023), who show that the meaning of inconclusive re-

sponses varies by examiner. They explain that forensic laboratories vary in whether they allow elimination decisions to be based on "individual characteristics". Labs that do not allow this are in conflict with how the AFTE Range of Conclusions (Association of Firearm and Tool Mark Examiners, 2024) defines inconclusive and elimination responses. We put "individual characteristics" in quotes because it is unknown whether these characteristics (surface contour patterns) constitute a signature that is unique to each individual firearm, up to the level that can be observed using current technology such as comparison microscopes, 3D imaging, etc. (NAS, 2008). The AFTE (2016) Theory of Identification assumes the truth of this unproven uniqueness.

5.4 Flaw D: Confidence intervals for error rates

We first describe what confidence intervals are and why they are critical for interpreting the results of firearms examination studies. Next, we explain why the statistical analysis methods used to compute confidence intervals in all the black-box firearms examination studies in Appendix A are invalid.

Each of the black-box studies reports estimates (also called "point estimates") of error rates. For example, the false positive error rate is typically estimated by dividing the number of false identifications by the total number of different-source comparisons. For example, the point estimate of the false positive error rate for cartridge cases in the Monson et al. (2023) study is 0.933%. This point estimate alone is insufficient to draw any statistical conclusions (called statistical inferences) about the false positive error rate for firearms examiners (even when putting aside

all of the other study design and analysis flaws in Table 1). The reason is that we also need to have some measure of uncertainty such as a confidence interval to put the point estimate into perspective. A confidence interval for the false positive error rate represents the range of plausible values for it that are consistent with the study data. The importance of reporting measures of uncertainty in addition to point estimates is emphasized in FDA recognized clinical guidelines for evaluating diagnostic tests (CLSI, 2023, p.42) which states that "Point estimates alone are not enough to evaluate test performance because they do not reflect the variability in the estimates."

The National Academies of Science (NAS, 2009, p.116) also highlights the importance of providing confidence intervals that reflect the sources of variability. They state the following:

> A key task for the scientific investigator designing and conducting a scientific study, as well as for the analyst applying a scientific method to conduct a particular analysis, is to identify as many sources of error as possible, to control or to eliminate as many as possible, and to estimate the magnitude of remaining errors so that the conclusions drawn from the study are valid. Numerical data reported in a scientific paper include not just a single value (point estimate) but also a range of plausible values (e.g., a confidence interval, or interval of uncertainty).

Confidence intervals are closely related to the margin of error in an opinion poll. For instance, consider an opinion poll asking likely voters whether they will vote for candidate A or B. As a simple example, if the opinion poll reports an estimate

that 48% will vote for candidate A and 52% for candidate B, with a 1% margin of error, then candidate B is in good shape. However, the same point estimates with a 5% margin of error means that neither candidate is clearly in the lead since the uncertainty (as measured by the margin of error) is too high. Similarly, a confidence interval represents the uncertainty, i.e., the range of plausible values for the quantity of interest (e.g., the false positive error rate) that are consistent with the data. The margin of error is related to the sample size of a study in that a smaller sample size leads to a larger margin of error, ceteris paribus.

Though different methods are used to construct confidence intervals in each of the black-box studies, they all suffer from (at least) the same major flaw. The flaw is that none of them accounts for variation among firearms in their likelihood of producing easier/harder to match toolmarks. This variation can arise from different makes/models of firearms (such as a 9mm Glock vs, Ruger SR9) or from individual firearms of the same make/model such as the 27 Beretta M9A3-FDE semiautomatic pistols used by Bajic et al. (2020); Monson et al. (2023). The problem is that failing to account for such variation can (and did) lead to invalid confidence intervals, as described below.

Some of the largest studies, e.g., Guyll et al. (2023) and the study of Bajic et al. (2020) (which we refer to as "AMES II"), show differences between firearm makes/models (such as Jimenez vs. Beretta) in their likelihood of producing easier/harder to match toolmarks. There is also evidence for the existence of such variation among individual firearms of the same make/model from the AMES II study. For example, examiners were asked to rate the difficulty of test cases as being

"easy", "average", or "hard".[7] For each of the 27 Beretta pistols the ratings were averaged across all examiners and reported in Table F1 of the AMES II study (Bajic et al., 2020, p.113). For cartridge case comparisons, the proportion of "hard" test cases generated by a Beretta pistol ranged from 0% to 62% (for Beretta pistols "F" and "E", respectively) when the ground truth was same source. The range was 1% to 43% (for Beretta pistols "O" and "Z", respectively) when the ground truth was different source. This shows substantial variability among individual Beretta pistols (all of the same model) in terms of how difficult the resulting test cases were judged to be by examiners. As another example, the proportion of examiner evaluations that were "exclusions" varied widely, i.e., from 26% to 86%, depending on which of the 27 Beretta pistols was used to generate cartridge case comparisons when the ground truth was different source (Bajic et al., 2020, p.116, Table F3). The analogous result for "identifications" under same source ground truth is a range from 60% to 97%.

Next, consider the black-box study of (Mattijssen et al., 2020, p.12, Appendix A)[8] which used 38 9mm Luger Glock pistols to generate same-source test cases. Each pistol was fired twice and each of the resulting 38 pairs of cartridge cases was evaluated by 77 examiners. For each of the 38 cartridge case pairs, the true positive rate averaged over 77 examiners was estimated by dividing the number of identifications by the total number of conclusive judgements. The estimates ranged from 100% (i.e., all examiners correctly reported "identification") down to 45% (i.e., only 45% of the examiners reported "identification"). This demonstrates substantial

[7] A limitation is that the data here are examiners' subjective assessments of difficulty and do not necessarily reflect actual difficulty.

[8] A caveat is that test cases in this study are, according to the authors, intended to be challenging and examiners make decisions based only on images of firing pin aperture shear marks.

variability across individual firearms in terms of average examiner performance, at least in this study.

An analogous result holds for the different-source test cases from (Mattijssen et al., 2020), which were generated from 22 pairs of Glock pistols. Each pistol was fired once and each of the resulting 22 pairs of cartridge cases was evaluated by 77 examiners. For each pair of the 22 pairs, the true positive rate averaged over 77 examiners was estimated by dividing the number of exclusions by the total number of conclusive judgements. The estimates ranged from 100% (i.e., all examiners correctly reported "exclusion") down to 40% (i.e., only 40% of the examiners reported "exclusion"). This again demonstrates substantial variability across different pairs of the same firearm in average examiner performance.

Chapnick et al. (2021) use a study design that they reference from the Montgomery (2017) textbook on experimental design, but they did not use the corresponding statistical method specified on the pages that they cited in this textbook for analyzing data when using that study design. That statistical method is designed to account for multiple sources of variation, such as the ones discussed above, but it was not used and instead an invalid method was used.

To demonstrate the impact of taking variability across individual firearms (and pairs of individual firearms) into account when computing confidence intervals, consider (Monson et al., 2023), which used 37 individual firearms for cartridge case comparisons. To appropriately compute confidence intervals, we advocate using the statistical model from Gallas et al. (2009), which is also used in the software referenced by the FDA. We advocate to combine this with an exact confidence interval

method (since the estimated false positive probability is close to 0) that extends the commonly used Clopper-Pearson method to this setting. Setting aside all of the other flaws in Table 1, an exact 95% confidence interval for the false positive error rate using this approach includes values greater or equal to 18%. See Appendix B for a full description and proof. This is in contrast to the reported confidence interval (0.548%, 1.57%), which only has values going up to 1.57%. The computation and justification for the 18% depends only on the number of individual firearms used, the estimated false positive error rate from the study, and the study design. Since many of the other studies in Appendix A involve fewer individual firearms and similar estimated false positive rates, the approach in Appendix B can be used to show analogous results, i.e., that ignoring the impact of variability due to individual firearms can lead to much smaller confidence intervals than correctly taking this variability into account.

The remedy to the above problem, which occurs in all studies in Appendix A, is to apply one of the methods referenced above that is valid for analyzing data with multiple sources of variation. Until that is done, it is unknown how much uncertainty to attach to the point estimates of error rates reported from these studies. Because of this, it is difficult or impossible to interpret the results from these studies (akin to learning the estimated percentages from an opinion poll without knowing the margin of error).

5.5 Flaw E: Missing data

Study validity requires that missing data, such as item- and unit-non-response, be handled by appropriate statistical methods. This was not done in any of the studies in Appendix A, with a few exceptions. Some of the smallest studies had no missing data reported, so flaw E does not apply. One of the methods used in the AMES II analysis, which computes examiner-specific error rates which it then averages across examiners, does implicitly address one part of missing data. However, that method is not sufficient for addressing missing data bias since it ignores potential factors that may impact both dropout and also estimator performance.

The importance of addressing missing data has been recognized at least since 2010 when the National Academies of Sciences published a seminal report on this topic for clinical studies (National Research Council and others, 2010). In the peer-reviewed summary of their report, they state that "Substantial instances of missing data are a serious problem that undermines the scientific credibility of causal conclusions from clinical trials." (Little et al., 2012, p.1355). We agree, and missing data is a problem in all types of controlled experiments. Black-box experiments involving firearms examination are no exception, and unfortunately the amount of missing data is substantial (where it is reported at all). For example, as described by Khan and Carriquiry (2023b), the amount of missing data due to examiner dropout and non-response is greater than 30% in Monson et al. (2023). At a minimum, every study should report the amount of missing data and conduct statistical analyses to address the additional bias and uncertainty introduced by it appropriately. This is not done in any of the firearms examination studies.

According to the National Academies of Science report on missing data in clinical trials (National Research Council and others, 2010, p.2):

> Modern statistical analysis tools–such as maximum likelihood, multiple imputation, Bayesian methods, and methods based on generalized estimating equations–can reduce the potential bias arising from missing data by making principled use of auxiliary information available for nonrespondents. The panel encourages increased use of these methods.

These or other appropriate statistical methods can be used to reduce the potential bias caused by missing data in firearms examination studies. None of these methods were used in any of the 28 firearms examination studies in Appendix A.

According to joint FDA and European Medicines Agency (EMA) guidances, the statistical analysis plan for every clinical trial should be written before the study starts (FDA/EMA, 2023, p.27) and should provide the "Procedure for accounting for missing, unused, and spurious data." (FDA/EMA, 2018, p.42). This applies to any validation study as well, regardless of whether it is a medical diagnostic test (for example) or a firearms examination black-box study. Failing to do an adequate statistical analysis to deal with missing data, as in the firearms examination studies (with the exception of some of the small studies that have no missing data), is invalid.

6 Discussion

We provided a statistical evaluation of 28 black-box validation studies, i.e., all such studies found in our literature search. We used the methodology of experimental

design and guidance documents from the FDA about statistical standards in research studies. Our main finding is that methodological problems are pervasive and consequential, and thus the scientific validity of firearms examination has not been established.

Although current practice in firearms examination relies heavily on human judgements, researchers have been working on the development of automated, objective methods for forensic comparisons, as recommended by NAS (2009); PCAST (2016); Kafadar (2019). Some black-box studies have already included automated methods, e.g., Mattijssen et al. (2020, 2021); Law and Morris (2021). Including automated methods as a "participant" in a black-box study is an opportunity to transition from human comparisons to algorithmic comparisons, or a hybrid combination of the two. A potential advantage of automated methods is increased transparency and replicability. In order to have this advantage, however, automated algorithms need to be made open-source and all data in the corresponding validation studies (including that used to train the algorithms) need to be publicly available (Kafadar, 2019). Currently, this is not the case for any of the black-box studies involving automated methods in our literature search.

We evaluated the experimental design and statistical analysis methods used in studies that aim to assess the accuracy of firearms comparisons. We did not evaluate whether examiners in validation studies or real casework are conforming (i.e., adhering) to a clearly defined protocol or procedure. Conformance to such a protocol is needed in order to be able to generalize results from a study to a population of examiners (who use that protocol) in real casework as explained by Mejia et al. (2020);

Swofford et al. (2024). The authors of a large black-box study (Baldwin et al., 2023, p.4), to their credit, acknowledge the following about the set of possible conclusions that firearms examiners can report:

> The language used to define the AFTE Range of Conclusions is specific and concise, but our data suggest that interpretation and application of this system is not consistent across the profession. Rather, our analysis strongly suggests that, as currently applied, the categories of the current five-point AFTE Range of Conclusions scale cannot be said to have consistent, meaningful interpretations.

This lack of conformance to a well-defined, clearly interpretable, firearms examination protocol in practice is a major roadblock to any research program that aims to evaluate scientific validity of a forensic method. One needs to have both adequate conformance and adequate study design/analysis in order for a study to reliably assess scientific validity of a method (Swofford et al., 2024).

We endeavored to provide useful recommendations for issues to consider and methods that can be used in the design and statistical analysis of future firearms examination validation studies. It also may be beneficial when planning future black-box studies to have an independent set of researchers and practitioners evaluate the proposed study design and statistical analysis plan before the study is started. This can help to find and fix potential problems before it is too late, and such pre-reviews are standard in the conduct of medical research.

Disclosures and Acknowledgments

This work was partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College and University of Nebraska, Lincoln. Dr. Rosenblum was supported in this research by a Nexus Award from Johns Hopkins University. The opinions expressed herein are solely those of the authors and do not necessarily reflect the views of any institution or other person. Some material in this document is reproduced, sometimes in a modified format, from Dr. Rosenblum's report (declaration) in a criminal case where he is an expert witness for the defense (DC Superior Court, 2023). Dr. Rosenblum is currently an expert witness in another case. His work as expert witness is through the consulting company Evolution Trial Design, Inc. of which he is co-owner and president. Additional related work includes expert declarations co-authored by Drs. Susan Vanderplas, Kori Khan, Heike Hofmann, and Alicia Carriquiry (2022) that were filed in the aforementioned case, and an amicus curiae brief co-authored by Drs. Rosenblum, Vanderplas, and Cuellar, among others, for a case at the Maryland Court of Appeals (2022).

References

AFTE (2016, October). Response to pcast report on forensic science. https://afte.org/resources/afte-position-documents. Accessed: 2023-12-13.

ASA (2024). American Statistical Association. https://www.amstat.org/. Accessed: 2024-01-24.

ASA (2024). American Statistical Association position on statistical statements for forensic evidence: Presented under the guidance of the ASA forensic science advisory committee. https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf. Accessed: 2024-01-24.

Association of Firearm and Tool Mark Examiners (2024). AFTE Range of Conclusions, Firearms Examiner Training. https://nij.ojp.gov/nij-hosted-online-training-courses/firearms-examiner-training/module-11/afte-range-conclusions. Accessed: 2024-02-08.

Bajic, S., L. S. Chumbley, M. Morris, and D. Zamzow (2020). Validation study of the accuracy, repeatability, and reproducibility of firearm comparisons. Technical report, Ames Lab., Ames, IA (United States).

Baldwin, D. P., S. J. Bajic, M. D. Morris, and D. S. Zamzow (2023). A study of examiner accuracy in cartridge case comparisons. part 2: Examiner use of the afte range of conclusions. *Forensic Science International 349*, 111739.

Bunch, S. and D. Murphy (2003). A comprehensive validity study for the forensic examination of cartridge cases. *AFTE JOURNAL 35*(2), 210–210.

Butler, J., H. Iyer, R. Press, M. Taylor, P. Vallone, and S. Willis (2020, 2020-12-18 00:12:00). Nist scientific foundation reviews.

Cazes, M. and J. Goudeau (2013). Validation study of results from hi-point consecutively manufactured slides. *AFTE Journal 45*, 175–177.

Champod, C., A. Biedermann, J. Vuille, S. Willis, and J. De Kinder (2016). ENFSI guideline for evaluative reporting in forensic science: A primer for legal practitioners. *Criminal Law and Justice Weekly 180*(10), 189–193.

Chapnick, C., T. J. Weller, P. Duez, E. Meschke, J. Marshall, and R. Lilien (2021). Results of the 3d virtual comparison microscopy error rate (vcmer) study for firearm forensics. *Journal of forensic sciences 66*(2), 557–570.

CLSI (2023). Clinical and laboratory standards institute (clsi). evaluation of qualitative, binary output examination performance. 3rd ed. clsi guideline ep12.

DC Superior Court (2023). Criminal Action No. 2018-CF1-4356. Washington, D.C., July 4, 2023.

Dorfman, A. H. and R. Valliant (2022a). Inconclusives, errors, and error rates in forensic firearms analysis:three statistical perspectives. *Forensic Science International: Synergy 5*, 100273.

Dorfman, A. H. and R. Valliant (2022b). A re-analysis of repeatability and reproducibility in the ames-usdoe-fbi study. *Statistics and Public Policy 9*(1), 175–184.

Dror, I. E. and N. Scurich (2020). (mis)use of scientific measurements in forensic science. *Forensic Science International: Synergy 2*, 333–338.

Fadul, T. G. (2011). An empirical study to evaluate the repeatability and uniqueness of striations/impressions imparted on consecutively manufactured glock ebis gun barrels. *AFTE Journal 43*(1), 37–44.

FDA (2007, March). Guidance for Industry and FDA Staff: Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Devices and Radiological Health, Diagnostic Devices Branch, Division of Biostatistics, Office of Surveillance and Biometrics. 2007. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/statistical-guidance-reporting-results-studies-evaluating-diagnostic-tests-guidance-industry-and-fda. Accessed June 25, 2023. Technical report, FDA, Rockville, MD: US FDA.

FDA/EMA (2018). US Food and Drug Administration and European Medicines Agency. ICH E6 Good Clinical Practice Guidance For Industry. (GCP) https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e6r2-good-clinical-practice-integrated-addendum-ich-e6r1.

FDA/EMA (2023). US Food and Drug Administration and European Medicines Agency. ICH E9 Statistical Principles for Clinical Trials - Scientific Guideline https://www.ema.europa.eu/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf.

Firearms and Toolmarks Subcommittee of the Organization of Scientific Area Committees (OSAC) at the U.S. National Institute of Standards and Technology

(NIST) (2016, December). Response to the President's Council of Advisors on Science and Technology (PCAST) Call for Additional References Regarding its Report 'Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods'. https://www.nist.gov/organization-scientific-area-committees-forensic-science/firearms-toolmarks-subcommittee Accessed October 26, 2022. Technical report, OSAC.

Gallas, B. (2023). Software: Multi-reader, multi-case analysis methods (roc, agreement, and other metrics). `https://www.fda.gov/medical-devices/science-and-research-medical-devices/imrmc-software-do-multi-reader-multi-case-statistical-analysis-reader-studies`.

Gallas, B. D., A. Bandos, F. W. Samuelson, and R. F. Wagner (2009). A framework for random-effects roc analysis: biases with the bootstrap and other variance estimators. *Communications in Statistics - Theory and Methods 38*(15), 2586–2603.

Guyll, M., S. Madon, Y. Yang, K. A. Burd, and G. Wells (2023). Validity of forensic cartridge-case comparisons. *Proceedings of the National Academy of Sciences 120*(20), e2210428120.

Hamby, J., D. Brundage, and J. Thorpe (2009). The identification of bullets fired from 10 consecutively rifled 9mm ruger pistol barrels: A research project involving 507 participants from 20 countries. *AFTE Journal 41*(2), 99–110.

Hofmann, H., A. Carriquiry, and S. Vanderplas (2020). Treatment of inconclusives in the afte range of conclusions. *Law, Probability and Risk 19*(3-4), 317–364.

Human Factors Committee of the Organization of Scientific Area Committees (OSAC) for Forensic Science (2020, March). Human factors in validation and performance testing of forensic science. Technical report, OSAC.

Kafadar, K. (2019). The need for objective measures in forensic evidence. *Significance 16*(2), 16–20.

Kassin, S. M., I. E. Dror, and J. Kukucka (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition 2*(1), 42–52.

Kerkhoff, W., R. Stoel, C. E. H. Berger, E. Mattijssen, R. Hermsen, N. Smits, and H. J. J. Hardy (2015). Design and results of an exploratory double blind testing program in firearms examination. *Science & justice : journal of the Forensic Science Society 55 6*, 514–9.

Kerkhoff, W., R. Stoel, E. Mattijssen, C. Berger, F. Didden, and J. Kerstholt (2018). A part-declared blind testing program in firearms examination. *Science & justice 58*(4), 258–263.

Khan, K. and A. Carriquiry (2023a). Hierarchical bayesian non-response models for error rates in forensic black-box studies. *Philosophical Transactions of the Royal Society A 381*(2247), 20220157.

Khan, K. and A. L. Carriquiry (2023b). Shining a light on forensic black-box studies. *Statistics and Public Policy 10*, 1–21.

Law, E. F. and K. B. Morris (2021). Evaluating firearm examiner conclusion variability using cartridge case reproductions. *Journal of Forensic Sciences 66*(5), 1704–1720.

Little, R. J., R. D'Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, J. D. Neaton, A. Rotnitzky, D. Scharfstein, W. J. Shih, J. P. Siegel, and H. Stern (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine 367*(14), 1355–1360. PMID: 23034025.

Lund, S. P. and H. Iyer (2017). Likelihood ratio as weight of forensic evidence: A metrological perspective. *Journal of Research of the National Institute of Standards and Technology 122, Article 27*, `https://doi.org/10.6028/jres.122.027`.

Maryland Court of Appeals (2022). Case number: COA-REG-0010-2022. September 2, 2022.

Mattijssen, E., W. Kerkhoff, C. Berger, I. Dror, and R. Stoel (2016). Implementing context information management in forensic casework: Minimizing contextual bias in firearms examination. *Science & Justice 56*(2), 113–122.

Mattijssen, E. J., C. L. Witteman, C. E. Berger, N. W. Brand, and R. D. Stoel (2020). Validity and reliability of forensic firearm examiners. *Forensic science international 307*, 110112.

Mattijssen, E. J. A. T., C. L. M. Witteman, C. E. H. Berger, X. A. Zheng, J. A.

Soons, and R. D. Stoel (2021). Firearm examination: Examiner judgments and computer-based comparisons. *Journal of Forensic Sciences 66*(1), 96–111.

Mejia, R., M. Cuellar, and J. Salyards (2020). Implementing blind proficiency testing in forensic laboratories: Motivation, obstacles, and recommendations. *Forensic Science International: Synergy 2*, 293–298.

Monson, K. L., E. D. Smith, and S. J. Bajic (2022). Planning, design and logistics of a decision analysis study: The fbi/ames study involving forensic firearms examiners. *Forensic Science International: Synergy 4*, 100221.

Monson, K. L., E. D. Smith, and E. M. Peters (2023). Accuracy of comparison decisions by forensic firearms examiners. *Journal of Forensic Sciences 68*(1), 86–100.

Montgomery, D. C. (2017). *Design and analysis of experiments*. John wiley & sons.

NAS (1992). *Responsible Science: Ensuring the Integrity of the Research Process. Volume I. National Academy of Sciences (US), National Academy of Engineering (US) and Institute of Medicine (US) Panel on Scientific Responsibility and the Conduct of Research. https://www.ncbi.nlm.nih.gov/books/NBK234526/*. Washington (DC): National Academies Press (US).

NAS (2008). Ballistic imaging. Technical report, National Academy of Sciences, Washington, DC.

NAS (2009). *Strengthening Forensic Science in the United States: A Path Forward*. Washington, D.C.: National Academies Press.

National Institute of Forensic Science of Australia/New Zealand (2019). Empirical study design in forensic science: A guideline to forensic fundamentals.

National Research Council and others (2010). The prevention and treatment of missing data in clinical trials. Technical report, National Academy of Sciences.

Neuman, M., C. Hundl, A. Grimaldi, D. Eudaley, D. Stein, and P. Stout (2022). Blind testing in firearms: Preliminary results from a blind quality control program. *Journal of forensic sciences 67*(3), 964–974.

Ommen, D. M. and C. P. Saunders (2018). Building a unified statistical framework for the forensic identification of source problems. *Law, Probability and Risk 17*(2), 179–197.

PCAST (2016). *President's Council of Advisors on Science and Technology: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-comparison Methods*. Executive Office of the President of the United States, President's Council.

Robertson, B., G. Vignaux, and C. E. Berger (2016). *What Questions Can the Expert Deal With?*, Chapter 4, pp. 53. John Wiley & Sons, Ltd.

Rosenblum, M., E. T. Chin, E. L. Ogburn, A. Nishimura, D. Westreich, A. Datta, S. Vanderplas, M. Cuellar, and W. C. Thompson (2024, 01). Misuse of statistical method results in highly biased interpretation of forensic evidence in Guyll et al. (2023). *Law, Probability and Risk 23*(1), mgad010.

Scurich, N. (2022, 11). Inconclusives in firearm error rate studies are not 'a pass'. *Law, Probability and Risk 21*(2), 123–127.

Scurich, N., D. L. Faigman, and T. D. Albright (2023). Scientific guidelines for evaluating the validity of forensic feature-comparison methods. *Proceedings of the National Academy of Sciences 120*(41), e2301843120.

Spiegelman, C. and W. A. Tobin (2012, 10). Analysis of experiments in forensic firearms/toolmarks practice offered as support for low rates of practice error and claims of inferential certainty. *Law, Probability and Risk 12*(2), 115–133.

Swofford, H., S. Lund, H. lyer, J. Butler, J. Soons, R. Thompson, V. Desiderio, J. Jones, and R. Ramotowski (2024). Inconclusive Decisions and Error Rates in Forensic Science. Presentation at Center for Statistics and Applications in Forensic Evidence Webinar on February 27, 2024. Presenter: Henry Swofford. Presentation sponsored by National Institute of Standards and Technology (NIST) `https://www.youtube.com/live/w0Y0fEOaaWY?si=nJOQSI7ZtqTmw3t5`.

Taroni, F. and A. Biedermann (2005, 03). Inadequacies of posterior probabilities for the assessment of scientific evidence. *Law, Probability and Risk 4*(1-2), 89–114.

Thompson, W. C., J. Vuille, A. Biedermann, and F. Taroni (2013). The role of prior probability in forensic assessments. *Frontiers in Genetics 4*, 220.

Zhou, X.-H., D. K. McClish, and N. A. Obuchowski (2009). *Statistical methods in diagnostic medicine*. John Wiley & Sons.

Appendix A: List of Black-Box Studies That We Reviewed

Below we give references for black-box studies of firearms examination. Some studies are reported both in technical reports and in subsequently published papers, or are reported in multiple papers. In such cases, we only count each study once in determining how many studies the flaws in Table 1 apply to. That is why the total number of studies is 28 (and not 33, which is the length of the list).

The list includes two types of studies. The first type consists of black-box studies where examiners are presented with test cases of the following form: decide whether a single item (called the "unknown") was fired from the same gun as a set of reference items (called the "knowns"). The items are either cartridge cases or bullets, and the reference items are fired from the same individual gun.

Studies that use the other type of test cases, called "set-based" studies, not only have every flaw in Table 1 but also have more flaws. In brief, "set-based" studies (e.g., where a set of items are all compared to each other, or where every item in one set needs to be matched to an item in another set) involve multiple comparisons in the same test question. The process of elimination (or similar logic) can be used to reduce the number of possibilities for the remaining comparisons after some comparisons have been decided on; this logical dependence can make the overall set of comparisons easier than if the comparisons had all been separate.

We included AFTE Journal articles but give a note of caution that there are potential issues with inadequate peer-review process during some years.

1. Baldwin, D.P. Bajic, S.J., Morris, M., Zamzow, D. A study of false-positive and false-negative error rates in cartridge case comparisons. Technical report, AMES LAB IA, 2014.

2. Baldwin, D. P., and Stanley J. Bajic, Max D. Morris, Daniel S. Zamzow, A study of examiner accuracy in cartridge case comparisons. Part 1: Examiner error rates, Forensic Science International, Volume 349, 2023, https://doi.org/10.1016/j.forsciint.2023.111733.

3. Baldwin, D. P., and Stanley J. Bajic, Max D. Morris, Daniel S. Zamzow, A study of examiner accuracy in cartridge case comparisons. Part 2: Examiner use of the AFTE range of conclusions, Forensic Science International, Vol. 349, 2023,

4. Bajic, S.J., Chumbley, L.S., Morris, M., and Zamzow, D. Report: Validation study of the accuracy, repeatability, and reproducibility of firearm comparisons. Technical Report # ISTR- 5220, Ames Laboratory-USDOE, 2020.

5. Best, B.A. & Gardner, E.A. (2022). An assessment of the foundational validity of firearms identification using ten consecutively button-rifled barrels. AFTE Journal, 54(1), 28-37.

6. Brundage, D.J. (1998). The identification of consecutively rifled gun barrels. AFTE Journal, 30(3), 438-444.

7. Bunch, S.G., and Murphy, D.P. A Comprehensive Validity Study for the Forensic Examination of Cartridge Cases. AFTE J. 2003, 35(2), 201-203.

8. Cazes, M., & Goudeau, J. (2013). Validation study of results from Hi-point consecutively manufactured slides. AFTE Journal, 45(2), 175-177.

9. Chapnick, C., Weller, T.J., Duez, P., Meschke, E., Marshall, J. and Lilien, R. (2021), Results of the 3D Virtual Comparison Microscopy Error Rate (VCMER) Study for firearm forensics. J Forensic Sci, 66: 557-570. https://doi.org/10.1111/1556-4029.14602

10. DeFrance and Van Arsdale, M.D. Validation Study of Electrochemical Rifling. AFTE J. 2003. 35(1).

11. Duez, P., Weller, T., Brubaker, M., Hockensmith, R.E., II and Lilien, R. (2018), Development and Validation of a Virtual Examination Tool for Firearm Forensics. J Forensic Sci, 63: 1069-1084. https://doi.org/10.1111/1556-4029.13668

12. Fadul, T.G. (2011). An empirical study to evaluate the repeatability and uniqueness of striations/impressions imparted on consecutively manufactured Glock EBIS gun barrels. AFTE Journal, 43(1), 37-44.

13. Fadul, T.G., Hernandez, G.A., Stoiloff, S., & Gulati, S. (2013). An empirical study to improve the scientific foundation of forensic firearm and tool mark identification utilizing 10 consecutively manufactured slides. AFTE Journal, 45(4), 376-391.

14. Fadul, T.G., Hernandez, G.A., Wilson, E., Stoiloff, S., & Gulati, S. (2013) 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. Technical Report 244232 NCJRS.

15. Guyll M, Madon S, Yang Y, Burd KA, Wells G. Validity of forensic cartridge-case comparisons. Proceedings of the National Academy of Sciences. 2023 May 16;120(20):e2210428120.

16. Hamby, J.E., Brundage, D.J., & Thorpe, J.W. (2009). The identification of bullets fired from 10 consecutively rifled 9mm Ruger pistol barrels: A research project involving 507 participants from 20 countries. AFTE Journal,41(2), 99-110.

17. Hamby, J.E., Brundage, D.J., Petraco, N.D.K. and Thorpe, J.W. (2019), A Worldwide Study of Bullets Fired From 10 Consecutively Rifled 9MM RUGER Pistol Barrels–Analysis of Examiner Error Rate. J Forensic Sci, 64: 551-557. https://doi.org/10.1111/1556-4029.13916

18. Keisler, M.A. & Hartman, S. & Kilmon, A. & Oberg, M. & Templeton, M. (2018). Isolated pairs research study. AFTE Journal. 50. 56-58.

19. Knapp, J. and Garvin, A. (2012), Consecutively manufactured .25 auto F.I.E. barrels–A validation study. Presentation at AFTE 43rd Annual Training Seminar, Buffalo, NY.

20. Law, E.F. and Morris, K.B. (2021), Evaluating firearm examiner conclusion variability using cartridge case reproductions. J Forensic Sci, 66: 1704-1720. https://doi.org/10.1111/1556- 4029.14758

21. Lyons, D.J. (2009). The identification of consecutively manufactured extractors. AFTE Journal, 41(3), 246-256.

22. Mattijssen, EJAT, Witteman, CLM, Berger, Berger, CEH, Brand, N.W., Stoel, R.D. Validity and reliability of forensic firearm examiners, Forensic Science International, Volume 307, 2020, 110-112, https://doi.org/10.1016/j.forsciint.2019.110112.

23. Mattijssen, EJAT, Witteman, CLM, Berger, CEH, Zheng, XA, Sons, JA, Stoel, R.D., Firearm examination: Examiner judgments and computer-based comparisons. J Forensic Sci. 2021; 66: 96- 111. https://doi.org/10.1111/1556-4029.14557

24. Mayland, B., & Tucker, C. (2012). Validation of obturation marks in consecutively reamed chambers. AFTE Journal, 44(2), 167-169.

25. Mikko, Don. An Empirical Study/Validation Test Pertaining to the Reproducibility of Toolmarks on 20,000 Bullets Fired Through M240 Machine Gun Barrels. AFTE Journal 45(3), 2013

26. Monson, K.L., Smith, E.D., Bajic, S.J. Planning, design and logistics of a decision analysis study: The FBI/AMES study involving forensic firearms examiners. Forensic Science International: Synergy, 4:100221, 2022.

27. Monson, K.L., Smith, E.D., and Peters, E.M. Accuracy of comparison decisions by forensic firearms examiners. Journal of forensic sciences, 68(1):86-100, 2023.

28. Monson K.L., Smith E.D., Peters E.M. Repeatability and reproducibility of comparison decisions by firearms examiners. Journal of Forensic Sciences 2023;

68: 1721–1740. https://doi.org/10.1111/1556-4029.15318

29. Pauw-Vugts, P., Walters, A., Oren, L., Pfoser, L. (2013). FAID2009: Proficiency test and workshop. 45. 115-127.

30. Smith E. Cartridge case and bullet comparison validation study with firearms submitted in casework. AFTE J. 2005; 37(4): 130-135.

31. Smith, T.P., Smith, A.G., & Snipes, J.B. (2016). A validation study of bullet and cartridge case comparisons using samples representative of actual casework. Journal of Forensic Sciences, 61(4), 939-946.

32. Smith, J.A. (2021), Beretta barrel fired bullet validation study. J Forensic Sci, 66: 547-556. https://doi.org/10.1111/1556-4029.14604

33. Stroman, A. (2014). Empirically determined frequency of error in cartridge case examinations using a declared double-blind format. AFTE Journal, 46(2), 157-175.

Appendix B: Confidence Interval Procedure for Black-Box Studies

As described in Section 5.4, the confidence intervals in all studies in our literature search are invalid since they fail to account for variation across firearms. We propose a direct extension of the Clopper-Pearson method that accounts for variation across both examiners and firearms. The method that we propose handles the resulting

dependence structure (MRMC structure–see Section 5.4) of the data, which applies to almost all of the black-box studies in Appendix A.

We use the nonparametric statistical model from Gallas et al. (2009), which is also used in the software referenced by the (FDA, 2007) guidance on validation studies for diagnostic testing. This model applies to the MRMC data structure. We then define an exact confidence interval method (since the estimated false positive probability may be close to 0) that extends the commonly used Clopper-Pearson method to this setting. Lastly, we describe how this exact confidence interval method can be applied to data from the large black-box study of Monson et al. (2023), if we were to temporarily set aside all of the other flaws in Table 1. We show that the exact 95% confidence interval for the false positive error rate using this approach includes values greater or equal to 18%, which implies that it is much wider than the 95% confidence interval reported in that study.

Statistical Model:

The statistical model defined by (Gallas et al., 2009, p. 2588) is nonparametric except for independence assumptions given next. For simplicity, we focus on different source comparisons and the false positive error rate defined by Monson et al. (2023) as the total number of identification responses divided by the total number of different source test cases[9] though the ideas can be applied to same source comparisons as well. When the Gallas et al. (2009) model is applied to our context, it assumes that each individual firearm and examiner is an independent draw from independent populations, and that conditioned on an individual firearm (or pair of firearms for

[9]They exclude "unsuitable" test cases, but we do not consider that here, for conciseness.

different source comparisons), the probability of a false positive response is independent across examiners, and that conditioned on each examiner the probability of a false positive is independent across individual firearms (or pairs of firearms for different sourse comparisons). The target of inference is the false positive rate averaged over the populations of examiners and individual firearms.

Confidence Interval Procedure: We next define an exact 95% confidence interval procedure for our problem (involving MRMC data) that is based on inverting hypothesis tests using a natural ordering of the sample space. This is precisely the approach used by, e.g., the Clopper-Pearson method for a sequence of n independent, identically distributed data; there, the sample space is ordered by the number of 1's out of the total number of responses n. We use the analogous ordering applied to our problem. The sample space in our problem is defined to be all possible data sets that could result when using a given study design, where for simplicity each response is coded as 1 ("identification") or 0 (not "identification"). The ordering is by total number (denoted m) of identification decisions across all examiners and test cases (denoted n).

The inputs to the confidence interval procedure are the following (using only different source comparisons, as explained above): the total number of "identification" responses m, the total number of test cases n, and the study design (i.e., the plan for which test cases from each firearm or pair of firearms is evaluated by each examiner). Included in the study design is the number of individual firearms which we denote by f. In the study of Monson et al. (2023), for example, there are $f = 37$ firearms involved for cartridge case comparisons (which we focus on below). The output of

the confidence interval procedure is a sub-interval of $[0, 1]$ that roughly speaking, represents the range of plausible values of the studywide false identification rate.[10]

Consider any data generating distribution P on the sample space defined above that is in the aforementioned statistical model, i.e., for which the conditional independence relations defined by Gallas et al. (2009) hold. Let $e(P)$ denote the parameter of interest, i.e., the population false identification rate, which is the probability of a false identification if a randomly selected pair of firearms (having the same class characteristics) is examined by a randomly selected examiner. Lastly, let P_0 denote the true, unknown data generating distribution.

The confidence interval procedure for the false positive error rate $e(P)$ is based on inverting the two-sided, level 0.05 test of null hypothesis $H_0(p) : e(P_0) = p$ (versus alternative hypothesis $e(P_0) \neq p$) for all values of p in the unit interval [0,1]. The null hypothesis is composite, since it is defined to represent the class of all possible distributions P on the sample space that are consistent with the aforementioned model and such that the false positive error rate $e(P)$ equals p.

The following hypothesis testing procedure generalizes the Clopper-Pearson method to our statistical model. For any candidate value p in [0,1], the null hypothesis $H_0(p)$ is rejected if for every distribution in $H_0(p)$, either the event A (defined below) has probability at most 0.025 or the event B (defined below) has probability at most 0.025. Event A is the subset of the sample space where the number of identification responses (1's) is less than or equal to the observed number of identifications

[10]We do not endorse this definition of the error rate, but since it was used by Monson et al. (2023), we use it as well in order to contrast our confidence interval procedure directly with the one from their paper.

m; event B is the subset of the sample space where the number of identification responses (1's) is greater than or equal to the observed number of identifications m. The confidence interval (technically, a confidence set) is defined to be the set of all p in $[0, 1]$ such that for at least one distribution in $H_0(p)$ the corresponding hypothesis test fails to reject. By construction, the hypothesis test has Type I error rate at most 0.05. There are other potential ways to compute a valid confidence interval than described above, but the above method is a direct generalization of the commonly used Clopper-Pearson method.

Application to Study of Monson et al. (2023):

We next prove that a valid, two-sided, 95% confidence interval for the false positive error rate, based on Monson et al. (2023) study design and data from its different source cartridge case comparisons and using the above confidence interval procedure, includes values greater than 18%. We prove this by constructing a data generating distribution P in the statistical model with $e(P) = 0.1845$ for which the corresponding probability under P of event A is at least 0.025 and the corresponding probability of event B is at least 0.025 and therefore $H_0(0.1845)$ cannot be rejected by the confidence interval procedure and so $p = 0.1845$ is contained in the confidence interval.

Let $v = 0.0942$, which will be explained below. Define δ to be the left endpoint of the Clopper-Pearson 95% confidence interval for the false positive rate, e.g., $\delta = 0.006$ for the different-source cartridge case comparisons from Monson et al. (2023). Let m denote the number of false positives that occurred in the study out of n total different-source comparisons. Therefore, by definition of the Clopper-Pearson confidence interval procedure, the probability that m or more 1's occur in n independent

Bernoulli draws each with probability δ is 0.025.[11]

Define the following distribution P on the sample space:

> Each individual firearm j is represented as an independent, identically distributed draw X_j from a Bernoulli(v) distribution, i.e., it takes value 1 with probability v and 0 otherwise. Consider any pair of firearms i, j, and any randomly drawn examiner k who compares a single unknown bullet/cartridge case from one of these guns to one or more known bullets/cartridge cases from the other. If $X_i = X_j = 0$, then the examiner makes a false positive error with probability δ based on an independent Bernoulli(δ) draw denoted Y_{ijk}; otherwise, if either $X_i = 1$ or $X_j = 1$, then the examiner makes a false positive error with probability 1.

Consider the corresponding probability that bullets/cartridge cases fired from two (different) randomly chosen firearms i, j and examined by a randomly chosen examiner k results in a false positive. This equals the probability that (X_i or X_j is 1) or ($X_i = X_j = 0$ and $Y_{ijk} = 1$), which by construction above is $p = 2v - v^2 + (1-v)^2\delta = 0.1845$. Therefore, the above distribution is in $H_0(0.1845)$.

Recall that m denotes the number of false positives that occurred in the study out of n total different-source comparisons. The probability of m or fewer false positives (across all n different source comparisons in the study) under the aforementioned distribution P is at least the product of $(1 - v)^f$ (for $f = 37$) and the probability that at most m 1's occur in n independent Bernoulli draws each with probability δ, i.e., the probability of m or fewer false positives is at least $(1 - v)^{37} \times 0.975 > 0.025$.

[11] This assumes that $m > 0$. The case of $m = 0$ is analogous.

We now explain how we selected the value of v defined above. We had chosen the value of v above so as to approximately maximize $(p = 2v - v^2 + (1 - v)^2)$ under the constraint that $P(A) \geq 0.025$. In fact, by our construction, the probability under P of event A above is at least 0.025 regardless of how many false positives m are observed, due to the ordering defined in the Confidence Interval Procedure above. Also, the probability under P of event B is at least $1 - (1 - v)^{37} = 0.974$ (which exceeds 0.025) since event B occurs whenever at least one firearm j has $X_j = 1$, due to the structure of the study design given in Table 4 on page 110 of the AMES II technical report Bajic et al. (2020). Since both events A and B have probability greater than 0.025 under the previously defined distribution, it follows that the hypothesis test in the Confidence Interval Procedure defined above fails to reject and therefore $p = 0.1845$ is contained in the 95% confidence interval for the false positive error rate.