IN THE COURT OF APPEALS OF MARYLAND

KOBINA EBO ABRUQUAH,	:	
Appellant,	:	
V.	:	**COA-REG-0010-2022**
STATE OF MARYLAND,	:	
Appellee.	:	

AMICUS CURIAE BRIEF IN SUPPORT OF APPELLANT KOBINA EBO ABRUQUAH

Dr. Michael Rosenblum
Professor of Biostatistics
Johns Hopkins University
2408 Everton Road
Baltimore, MD 21202
mrosen@jhu.edu

Dr. Susan VanderPlas
Assistant Professor of Statistics
University of Nebraska-Lincoln
340 Hardin Hall North Wing
Lincoln, NE 68583
svanderpls2@unl.edu

Dr. Kori Khan
Assistant Professor of Statistics
Iowa State University
2438 Osborn Dr.
Ames, IA 50011
kkhan@iastate.edu

Dr. Yawen Guan
Assistant Professor of Statistics
University of Nebraska-Lincoln
340 Hardin Hall North Wing
Lincoln, NE 68583
yguan12@unl.edu

Dr. Arturo Casadevall
Chair, Molecular Microbiology & Immunology
Bloomberg Distinguished Professor
Alfred & Jill Sommer Professor and
Chair Professor
Johns Hopkins University
615 N. Wolfe Street, Room E5132
Baltimore, MD 21205
acasade1@jhu.edu

Dr. Thomas D. Albright
Professor and Conrad T. Prebys Chair
The Salk Institute for Biological Studies
10010 North Torrey Pines Road
La Jolla, California 92037
tom@salk.edu

Dr. Emily Robinson
Assistant Professor in Statistics
California Polytechnic State University
Faculty Offices East, Building 25, Room 107D
San Luis Obispo, CA 93407
e.95.robinson@gmail.com

Dr. Heather Akin
Assistant Professor of Agricultural Leadership, Education & Communication
University of Nebraska-Lincoln
143 Filey Hall
P.O. Box 830947
heather.akin@unl.edu

Dr. Bonner Denton
Galileo Professor of Chemistry
University of Arizona
1306 E. University Blvd.
Tucson, AZ 85721
mbdenton@email.arizona.edu

Dr. Jennifer Clarke
Professor, Dep't of Food Science and Technology, Dep't of Statistics
University of Nebraska-Lincoln
340 Hardin Hall North Wing
Lincoln, NE 68583
jclarke3@unl.edu

Dr. Heike Hofmann
Professor in Charge of Data Science
Iowa State University
2438 Osborn Drive
Ames, IA 50011
hofmann@iastate.edu

Dr. Bertrand Clarke
Professor and Chair, Dep't of Statistics
University of Nebraska-Lincoln
340 Hardin Hall North Wing
Lincoln, NE 68583
bclarke3@unl.edu

Dr. Maria Cuellar
Assistant Professor of Criminology and Statistics
University of Pennsylvania
558 McNeil Building
3718 Locust Walk
Philadelphia, PA 19104
maria.cuellar@gmail.com

TABLE OF CONTENTS

STATEMENT OF INTEREST OF AMICUS CURIAE[1]

Amicus curiae are professors and researchers in scientific disciplines who are concerned with the use of scientific studies to support the reliability of forensic evidence in the legal system. Most of us are not involved in the study of forensic disciplines directly, but we are scientists, statisticians, and researchers who are qualified to assess research design, execution, and the claims which are made as a result of research studies in firearms and toolmark analysis. We speak for ourselves, as private parties, and not for our institutions.

---

[1] Both parties have consented to this amicus brief. No person, other than Amici, made any monetary or other contribution to the preparation or submission of the brief (*See* Md. Rule 8-511).

INTRODUCTION

This amicus brief outlines the fundamental research principles used to evaluate the scientific validity of a method. What is discussed in this brief is not new; it describes the research requirements adhered to in science-based fields. The brief then discusses the application of these principles to the method used by firearms and toolmark examiners.

Adhering to the principles of sound research design and statistical analysis is fundamental to any applied science. There is no exception for forensic science. While the firearms and toolmark field has made strides, current research does not yet support the claims made by the discipline. Specifically, existing research studies that evaluated accuracy, reliability, and reproducibility of firearms examination have substantial flaws, described below. Our conclusion is that firearms examination has not been demonstrated to be accurate, reliable, or reproducible. Error rates for firearms examination (e.g., false positive identifications) are currently unknown, since existing studies are inadequate to establish these.

Issues with experimental design, participant selection, statistical analysis, and the interpretation of estimates pervade the current validation studies. As just one example, studies count inconclusive responses–those in which the examiner cannot make a definitive conclusion–as effectively correct (i.e., not as errors), which results in misleadingly low reported error rates.[2] Treating inconclusive responses as effectively correct results in reported error rates as low as zero percent. If inconclusives are instead

[2] Inconclusive responses are included in the total number of comparisons performed, the denominator, but not included as errors in the numerator.

treated as errors, error rates can be as high as 93%. The true error rate is likely between these two extremes, but until more well-designed research is performed, it remains *unknown*.

While there are encouraging developments in research design, data from a recent study shows an alarming lack of consistency in decisions when the same examiner was presented with the same evidence twice, and when different examiners were presented with the same evidence.[3] These new data further undermine the claim of a well-developed, scientifically valid method and cannot go unaddressed.

[3] Stanley Bajic et al., *Validation Study of the Accuracy, Repeatability, and Reproducibility of Firearm Comparisons*, 127 (2020), https://www.scribd.com/document/586448513/Ames-FBI-Validation-Study.

ARGUMENT

I. THE RELEVANT EXPERTISE

If the Court wishes to understand how weapons leave marks on fired ammunition, or how an examiner performs a comparison between two bullets or cartridge cases, practitioners are the best group to consult. If, however, the determination the Court must make is not about *how* the forensic process is performed, but rather *how well* it is performed –or, in the ideal, *how well it can be performed*–the person to consult is the research scientist.

An analogy to the field of medicine, a field with similarly high stakes consequences, is helpful. The epidemiologist who researches disease has a very different role and skill set than the doctor who treats patients.[4] If one wants to know about the effects of a disease on a population or how to slow the spread of an emerging virus, it is far more effective to consult the epidemiologist; if one wants to know how to treat a patient afflicted, then the doctor is the appropriate expert to consult.

[4] Another relevant and important distinction is between interested and disinterested parties. Those with a financial or personal stake in the outcome are generally not the only people who should be tasked with researching a particular issue.

II. SCIENTIFIC VALIDITY

The basic requirements of any valid scientific method are that it must be repeatable, meaning the examiner reaches the same conclusion when presented with the same evidence, reproducible, meaning different examiners reach the same conclusions when analyzing the same evidence, and accurate, meaning the conclusion reached is the correct one.

A *reliable* method or instrument gives consistent results. A scale, for example, can be perfectly reliable and report the same weight for the same object each time it's weighed. That does not mean the scale is *accurate*; it may consistently report the wrong weight. Reliability, or consistency, is a necessary component of a scientifically valid method, but does not, on its own, establish scientific validity.

A *valid* method produces accurate results. It is not possible to assess the accuracy of a method without testing it on samples where ground truth is known, meaning testing using samples of known origin. Because ground truth is unknown in case work–the examiner does not know if the two bullets were fired in the same gun or different guns–case work, even case work in which a second examiner agrees with the first examiner, cannot serve to support the validity of the method.

The goal of any validation study is 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. A high quality study design is needed in order to achieve these goals.

Evaluating the validity of an entire discipline requires many studies, over a range of conditions, with some replication; in addition, studies used to support the validity of a discipline must be well-designed, using appropriate test problems, instructions, sampling procedures, and statistical practices when analyzing the results. Scientifically, supporting studies should meet several conditions: they should be designed in consultation with statisticians, published in scholarly journals that require peer review by statisticians and subject matter experts,[5] the results must hold up over time and replication,[6] and the studies must be conducted over a wide range of conditions that are representative of those seen in applied settings. As an analogy, consider what is required for regulators such as the U.S. Food and Drug Administration (FDA) to approve a new drug. Multiple, high quality randomized trials are required, each of which needs to demonstrate efficacy of the

---

[5] Trade journals, including the AFTE Journal, are sometimes peer reviewed, but the peers are practitioners rather than research scientists; these reviews focus on the forensic procedures but neglect to consider the design of the study and the statistical validity of any reported results. As a result, studies from these journals often have serious methodological flaws. Research journals are not immune from this problem, but it is at least more likely that reviewers who are active research scientists and have training in statistical analysis and experimental design.

[6] Note that even prestigious scientific journals and respected institutions have published scientific results which do not hold up to the test of time. Ellis R. Lippincott et al., *Polywater: Vibrational spectra indicate unique stable polymeric structure.*, 164 Science 1482–1487 (1969). and other follow-up papers demonstrated unique properties of a form of water called polywater, first discovered in the USSR and then replicated in US labs and at the National Bureau of Standards (now known as NIST). These properties were later shown to be identical to those of sweat, D. L. Rousseau, *"Polywater" and Sweat: Similarities between the Infrared Spectra*, 171 Science 170–172 (1971)., suggesting that the original documented and peer-reviewed phenomenon was a result of replication of conditions producing laboratory contamination of samples. More details can be found in Joseph Stromberg, *The Curious Case of Polywater*, Slate, 2013, https://slate.com/technology/2013/11/polywater-history-and-science-mistakes-the-u-s-and-ussr-raced-to-create-a-new-form-of-water.html (last visited Aug 8, 2022).

drug for the target population. For firearms examination, though there have been randomized experiments, even the best ones have significant flaws.

III. FIREARMS AND TOOLMARK ANALYSIS

Firearms examiners compare two bullets or two cartridge cases under a comparison microscope. Historically, the Association of Firearms and Tool Mark Examiners (AFTE) method permitted an examiner to render three subjective judgments – identification, meaning they were fired in the same gun, exclusion, meaning they were fired from different guns, and inconclusive. The AFTE Theory of Identification has now adopted a five point scale, including three inconclusive options - "A. Agreement of all discernible class characteristics and some agreement of individual characteristics, but insufficient for an identification. B. Agreement of all discernible class characteristics without agreement or disagreement of individual characteristics due to an absence, insufficiency, or lack of reproducibility. C. Agreement of all discernable class characteristics and disagreement of individual characteristics, but insufficient for an elimination."[7]

There are many ways to attempt to quantify how often judgments are wrong, and it is important to fully understand the strengths and weaknesses of each potential approach. In the field of medicine, for example, the National Institutes of Health (NIH) has very strict requirements that ensure that the design of validation studies meet the highest

[7] AFTE Glossary, 6th Edition, *available at* https://afte.org/uploads/documents/AFTE_Glossary_Version_6.110619_DRAFT_.PDF.

standards, and the Food and Drug Administration regulates which tests can be used on patients.[8] There is currently no similar oversight mandating appropriately designed studies in the field of firearms and toolmark analysis. As such, the fact that a study was performed, or even published, does not mean that the results are reliable. One has to evaluate the design of the studies to determine whether they meaningfully contribute to the overall scientific validity of the discipline.

IV. STAGES OF SCIENTIFIC SUPPORT FOR FIREARMS AND TOOLMARK ANALYSIS

Examiners did not always claim to be able to identify a specific fired bullet to a specific gun. At the inception of firearms examination as a discipline, examiners made claims supported by their individual experience, borne of an understanding of the mechanics of firearms and the (relatively new) ability to accurately measure minute details of the firearm and ammunition.[9] These claims were supported by *descriptive* data, in that there were measurements being made in a laboratory setting, but examiners did not make source identification decisions nor establish any systematic data collection that would allow for inference that two bullets or cartridge cases were fired by the same gun.

---

[8] 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).

[9] A. L. Hall, *The Missile and the Weapon*, 39 BUFFALO MED. J. 727–736 (1900).

By the 1930s Valentine's Day Massacre, however, examiners began to make claims about the individualizing nature of the firearms manufacturing process.[10] These claims were still unsupported by any systematic data collection, but the claims were more expansive than previous written records, which highlight descriptive characteristics and do not attempt to draw a direct connection between fired ammunition and a specific weapon. Examiners had moved on to *inferential* claims, where the accumulated "data" of their past experiences were used to support more general claims about the methodology used in firearms and toolmark identification.

Over the next 60 years, the field focused on research into the investigative method and procedures, with some forays into initial attempts at quantitative evaluation methods. The next development of interest to the Court addressed the question of examiners' ability to apply a procedure to evaluate a set of samples of known provenance and come up with the correct answer. Such "black-box" studies are so called because they treat the examiner and evaluation procedure as an unobservable entity and evaluate only the resulting answer (rather than assessing the reasoning behind it). The subjective, visual comparisons performed during examiner evaluations cannot be tested step-by-step, a marked difference from disciplines like DNA where each step of a lab test can be audited separately.

---

[10] Calvin Goddard, *The Valentine Day Massacre: A Study in Ammunition-Tracing*, 1 Am. J. Police Sci. 60–78 (1930).

One of the first studies to attempt to test examiners' ability to reach the correct conclusion was Brundage (1994),[11] which served as a model for error rate studies in firearms and toolmark analysis for the next 15-20 years, with updated data published as recently as 2019.[12] Unfortunately, the design of the Brundage-Hamby studies is deeply flawed. As a result, the re-use of this study design has resulted in a collection of studies which cannot be relied upon for calculation of an error rate. These studies have two separate but related design flaws which, on their own, render the results unhelpful in understanding the performance of the method: they use multiple unknown and known samples in the same kit, and they are "closed-set" studies, meaning examiners know that all unknown samples have a matching known source.[13]

When multiple unknown and known samples are included in the same kit, examiners do not list out all comparisons which were performed. Instead, they fill in only the matching known sample for each unknown. This does not allow us to calculate the error rate for a comparison, because we do not know how many comparisons were performed.[14] As a result, it is impossible to estimate the probability of a missed

---

[11] David J. Brundage, *The Identification Of Consecutively Rifled Gun Barrels*, 1994, https://vufind.carli.illinois.edu/vf-sic/Record/sic_1201372/Description.

[12] 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 J. Forensic Sci. 551–557 (2019).

[13] The studies suffer from other design flaws as well, as discussed more fully below.

[14] To illustrate why a closed set test prevents the researcher from knowing the number of comparisons conducted, consider the case when there are two unknowns (A and B) and two knowns (C and D). One examiner might compare each of the unknowns to each of the knowns (A-C, A-D, B-C, B-D) for a total of four comparisons. Another examiner, however, might first compare A to C and determine them a match, and therefore refrain

elimination (where an examiner fails to eliminate samples from different sources). In addition, due to the knowledge that all unknown samples match a provided known, examiners can select the closest known sample instead of making a positive identification based on the visible evidence. All told, this leads to a misidentification rate that we can expect to be lower than in case work.[15] While these studies also have other issues (e.g. sampling bias), the structural flaws of the study are severe enough on their own to render the results unusable for evaluating examiners' ability to reach the correct conclusion.

Many studies which followed Brundage (1994) emulated the multiple-known to multiple-unknown study design, precluding a determination of the number of comparisons, essential data for an error rate study, though not all of these studies were also closed-set studies. In 2014, the Ames Laboratory undertook a study in conjunction with the Department of Defense. Recognizing the confounding problem of the previous studies,[16] the researchers modified the test problem design so that the number of

---

from comparing A to D. Accounting for all the possibilities, there could be anywhere from 2 to 4 comparisons. As the number of unknowns and knowns grow, the range of possibilities also increases. For example, if there were four knowns and four unknowns, the possible number of comparisons completed can range from 4 to 16.

[15] The issue of inconclusive responses, which figures so prominently in better designed open studies, does not typically arise in "closed set studies," in part because the additional information that all unknown items share a source with known items presented as part of the set is used by examiners when making their comparisons.

[16] "[T]he design of these previous studies, whether intended to measure error rates or not, did not include truly independent sample sets that would allow the unbiased determination of false-positive or false-negative error rates from the data in those studies." David P. Baldwin et al., *A Study of False-Positive and False-Negative Error Rates in Cartridge Case Comparisons*, 5 (2014), https://www.ojp.gov/pdffiles1/nij/249874.pdf (last visited Jan 29, 2020).

comparisons could be calculated. Similar designs were also adopted by Keisler (2018),[17] and Chapnick (2021).[18] While these studies have better test problem design (e.g. open v. closed), they still have some major flaws common to almost all studies in firearms and toolmark examination: there are significant levels of participant drop-out which are not accounted for in the analysis of results,[19] participants self-select instead of being randomly selected as part of a representative sample,[20] and there is no objective assessment of the difficulty of comparisons in each study (which makes it difficult to

---

[17] Keisler, M. A., Hartman, S., & Kil, A., *Isolated Pairs Research Study*, 50 AFTE J. 56–58 (2018).

[18] Chad Chapnick et al., *Results of the 3D Virtual Comparison Microscopy Error Rate (VCMER) Study for firearm forensics*, 66 J. Forensic Sci. 557–570 (2021).

[19] Participant drop-out is of particular concern because in many cases it occurs after participants have seen the study materials. If the materials are difficult comparisons, then less-skilled or less-confident examiners may drop out because they do not want to increase the published error rates for the discipline. Of course, there are other reasons participants may drop out, such as casework overloads, but the fact that there are explanations for the drop-out rate that would be related to the calculated error rate make the estimates generated from these studies statistically questionable. That the researchers do not account for these issues when calculating possible error rates (as is common in other disciplines with participant drop-out, such as medicine) is much more problematic.

[20] Most scientific studies involving humans take place on volunteer samples. What is problematic in FTE studies is that researchers make no effort to ensure that the participants in the study accurately reflect characteristics of the active examiner population, such as experience, lab type, training, and education level. Again, medical studies are a good comparison group: participants in pharmaceutical trials are also volunteers, but substantial effort is devoted to try to enroll participants who are representative of the general population, in accordance with guidelines from the NIH. Without a representative sample, it is difficult to justify generalizing the results of the study to the wider population – a critical step for utilizing these studies in a legal setting. 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).

compare studies or assess the relevance of a study to a specific case). The treatment of inconclusive responses is also a significant issue discussed below.

As the discipline of firearms and toolmark analysis has matured, and as pressure to validate the conclusions made by examiners using scientific studies of the examination process has increased, more sophisticated study designs have been developed which provide more nuanced ways to assess the discipline than raw error rates. The most recent set of studies to be released, colloquially known as Ames II,[21] discussed further below, examined not only error rate, but also the repeatability and reproducibility of examiner conclusions when assessing both bullets and cartridge cases. While Ames II still has many of the same flaws identified in other modern studies, it demonstrates that studies in this discipline are maturing and that it is possible in the future to design studies which directly answer questions of interest to the Court: is firearms analysis repeatable and reproducible? Do the method's error rates support its conclusions?

---

[21] Bajic et al., *Validation Study of the Accuracy, Repeatability, and Reproducibility of Firearm Comparisons*, 127 (2020), https://www.scribd.com/document/586448513/Ames-FBI-Validation-Study. We reference the full 127 page report of the Ames Laboratory to the FBI, comprehensively detailing data and analysis estimating accuracy, repeatability, and reproducibility inter alia of forensic firearms examinations. It was released to the public in early 2021 and then withdrawn. Before being withdrawn, it circulated widely enough to have been put into evidence in several court cases, including in this case. Portions of the report have been published online as preprints (L. Scott Chumbley et al., *Accuracy, Repeatability, and Reproducibility of Firearm Comparisons Part 1: Accuracy*, (2021), http://arxiv.org/abs/2108.04030 (last visited Jun 20, 2022)).

V. THE CURRENT DOMAIN-WIDE ERROR RATE IS UNKNOWN

While the state of research has matured to some degree, there remain significant and unaddressed problems with the design of the recent studies beyond the design of the test problem. Addressing these issues is not an impossible task. Medicine, for example, employs strict standards for the design and execution of clinical trials before adopting any new test or method. Unless and until the field employs the rigor seen in other scientifically mature disciplines, it is not possible to assess the utility of the current reported error rates.

There are two quantities of interest when evaluating a particular diagnostic test. Returning to medicine as an example, the sensitivity, or true positive rate, estimates how often the test identifies cancer when cancer is present. The specificity, or true negative rate, estimates how often the test identifies no cancer is present when there is no cancer. The sensitivity and specificity combined determine the overall accuracy rate and are useful for an agency such as the FDA in determining whether the test works as claimed.

A patient taking the test may be interested in different statistics describing the test performance. If the patient's test was positive, they would be interested in the positive predictive value: the probability that the patient has cancer given a positive test. If the patient's test was negative, they would instead be interested in the negative predictive value: the probability that the patient does not have cancer given a negative test.[22]

[22] To provide another example relevant to the current COVID pandemic, BinaxNow rapid antigen tests have a sensitivity of about 43% relative to PCR (55/127) but have a specificity of 100% relative to PCR (642/642). From an individual perspective, however,

Error rate studies with independent pairwise comparisons do allow calculation of the sensitivity, specificity, and false positive and false negative rates because they explicitly measure how many comparisons were performed along with the outcome of the comparisons. As alluded to before, however, this basic design characteristic is only present in a few modern firearms studies. While these few studies involving a known number of single pair comparisons allow for the calculation of the full set of error rates, they have other significant flaws which make their error rate estimates misleading and unreliable for the Court's purposes.

In order to rely on these studies and generalize their error rates to casework, validation studies not only need to be well designed, but must also include test samples that are representative of comparisons found in casework. In addition, the calculated error rates must account for any study flaws so that if error rates cannot be precisely estimated, they can at least be bounded by a reasonable interval.

The sections below discuss the issues in research design, both acknowledged by the firearms and toolmarks community, and yet to be acknowledged. Even looking only at factors that have been acknowledged, it is clear that the reported error rates are incorrect and misleading. Without further research, however, it is impossible to know

---

a positive BinaxNow test suggests a 100% chance of a positive PCR test (55/55), where a negative BinaxNow test suggests a 90% chance of a negative PCR test (642/714). That is, the BinaxNow test misses some COVID cases (because the PCR test is much more sensitive), but it is a very good screening tool because a positive antigen test is a very good indicator of an active COVID infection. Numbers from Krishna Surasi et al., *Effectiveness of Abbott BinaxNOW Rapid Antigen Test for Detection of SARS-CoV-2 Infections in Outbreak among Horse Racetrack Workers, California, USA*, 27 Emerg. Infect. Dis. 2761–2767 (2021).

how significant an effect the unacknowledged factors have on the true error rate for the discipline.

A. Acknowledged Research Design Issues

The following section discusses those issues which have been acknowledged by the firearms and toolmark community, though they remain currently unresolved.

1. The Reported Error Rates

Current validation studies report error rates for the method between zero[23] and 11.3 percent.[24] A zero percent error rate for any method, much less a subjective method using human judgment, is not scientifically plausible.[25] Even though many of the studies in this list have previously-identified methodological issues, we will work with this range of estimates for the moment.

---

[23] Michelle Cazes & Jeff Goudeau, *Validation Study Results from Hi-Point Consecutively Manufactured Slides*, 45 AFTE J. 175–177 (2013); Hamby et al., *supra* note 12; James E Hamby, *The Identification of Bullets Fired from 10 Consecutively Rifled 9mm Ruger Pistol Barrels: A Research Project Involving 507 Participants from 20 Countries*, 41 12 (2009); David J. Brundage, *The Identification of Consecutively Rifled Gun Barrels*, 30 AFTE J. 438–444 (1998).

[24] Erwin J. A. T. Mattijssen et al., *Validity and reliability of forensic firearm examiners*, 307 Forensic Sci. Int. 110112 (2020).

[25] "Although there is limited information about the accuracy and reliability . . . . claims that these analyses have zero error rates are not scientifically plausible.", pg. 142, STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES: A PATH FORWARD, (National Research Council (U.S.) ed., 2009).

2. Inconclusive Responses

One complication in calculating the error rates for firearms and toolmark examination is that the AFTE Theory of Identification (ToI) does not directly correspond with the physical state of the evidence, which is either from the same gun or from different guns. Instead, the AFTE ToI allows for an examiner to make an identification (same gun), elimination (different gun), or to make an inconclusive decision, indicating that there is insufficient information to make either definitive conclusion.

Given this mismatch, there are many potential ways to deal with inconclusive responses when calculating the error rate. Inconclusive decisions can be (1) removed entirely, (2) included as correct responses, or (3) included as incorrect responses.[26] These variations generate wildly different error rates based on the same data.

A hypothetical example highlights the confounding nature of this factor when evaluating reported error rates. In a test with 10 questions, if an examiner answers five questions correctly and does not answer five questions, these three methods generate different results. Removing inconclusive responses entirely (1) means the examiner had a 0% error rate. This rate, however, reflects only the error rate for the questions the examiner chose to answer, which could possibly be the easier questions. We do not know how the examiner would have performed on the five (potentially more difficult) questions

---

[26] They can be treated as *potential* errors, or effective eliminations as well. *See* Dror, N. Scurich, *(Mis)use of scientific measurements in forensic science, Forensic Science International: Synergy* (2020), https://doi.org/10.1016/j.fsisyn.2020.08.006;
H. Hofmann, A. Carriquiry, S. Vanderplas, *Treatment of inconclusives in the AFTE range of conclusions*, Law, Probability and Risk 19 (3–4) (2020) 317–364. https://doi.org/10.1093/lpr/mgab002.

she chose not to answer. In this example, counting the five unanswered questions as correct (2) would also generate a 0% error rate. Counting the inconclusive responses as wrong (3), however, would lead to a 50% error rate.

If instead the examiner answered four questions correctly, one incorrectly, and did not answer five questions, each of the different methods would generate: (1) 20% error rate (2) 10% error rate, and (3) 60% error rate *for the same data*. Dorfman & Valiant (2022) expand on this conundrum, concluding that inconclusives are at least potential errors, and the use of inconclusives in studies may mask potential errors in casework.[27]

Even without taking a position as to which most accurately presents the results of the study, it is a significant factor that cannot be ignored. The existing validation studies have all used the second method, counting inconclusive responses as correct responses (i.e., not as errors and not as unanswered). The reported error rates, when treating inconclusives as correct, are between 0 and 11.3% as described above; however, if inconclusives are instead treated as errors, error rates are instead between 0%[28] and 93%[29] (the latter value resulting from 335 inconclusives out of 360 total cartridge case comparisons). In reality, the true error rate is likely somewhere between the two ranges; this is consistent with Dorfman (2022)'s treatment of inconclusives as *potential* errors.[30]

---

[27] Alan H. Dorfman & Richard Valliant, *Inconclusives, errors, and error rates in forensic firearms analysis: Three statistical perspectives*, 2022 Forensic Sci. Int. Synergy (2022).
[28] Michelle Cazes and Jeff Goudeau, *supra* note 22.
[29] Erich Smith, *Cartridge Case and Bullet Comparison Validation Study with Firearms Submitted in Casework*, 36 AFTE J. 130–135 (2005).
[30] Dorfman et al., *supra* note 27.

3. Repeatability and Reproducibility

Recent data on the consistency of examiner decisions further undermines the discipline's claim of a low and well-understood error rate. A study "conducted between 2016 and 2020, in collaboration between the Federal Bureau of Investigation (FBI) and Ames Laboratory-USDOE," (colloquially known as Ames II) was the first modern study to test the repeatability and reproducibility of firearms examiners.[31] Over three different rounds, the study compared examiner decisions when presented with the same samples at different times (repeatability), and compared the conclusions of two different examiners when presented with the same evidence (reproducibility).

When the bullets were fired from the same gun, examiners disagreed with themselves on 21 percent of the comparisons. When the bullets were fired from different guns, examiners disagreed with themselves on 35.3 percent of the comparisons. Comparing results among different examiners, a second examiner disagreed with the first on 32.2 percent of the comparisons when bullets had the same source. For different-source bullets, a second examiner disagreed on 69.1 percent of the comparisons. While there is disagreement over the significance of these numbers,[32] the relatively poor repeatability / reproducibility rates need to be presented alongside the study's (much smaller) estimated false positive rates of 0.656 % and 0.933% for bullets and cartridge

---

[31] Bajic et al., *supra* note 21.

[32] *See e.g.*, Alan H. Dorfman & Richard Valliant, *A Re-analysis of Repeatability and Reproducibility in the Ames-USDOE-FBI Study*, (2022), http://arxiv.org/abs/2204.08889 (last visited Aug 19, 2022). In this preprint, Dorfman and Valiant re-analyze the repeatability and reliability data from Ames II and conclude that the analysis used in Ames II is flawed; instead, the results show weak repeatability and reproducibility.

cases, and false negative rates of 2.87% and 1.87% for bullets and cartridge cases, respectively.[33] The study's results are a potentially significant problem for the discipline, but until further well-designed research is performed, the scientific validity of the discipline as a whole remains unknown.

B. Unacknowledged Issues with Research Design

The following section outlines issues which remain unacknowledged by the current validation studies.

1. Hawthorne Effect

When evaluating validation studies, one has to consider the extent to which the experiment measures the real-life thing of interest. As just one example, when participants in a study are aware they are being observed, this can affect their behavior. This phenomenon is sometimes known as the Hawthorne effect.

This variable can be studied or controlled. There is at least one study that used blind proficiency testing intermixed with casework.[34] Thus, while the examiners knew they would be tested they did not know which item was a test and which was actual

---

[33] Because the study's data has not been released, further analysis of the disagreement between the reported error rates and the reported reproducibility and reliability numbers are not possible. It is likely that the discrepancy stems from the treatment of inconclusives (and sub-categories of inconclusive), but we cannot confirm this hypothesis until the researchers release the data.

[34] Maddisen Neuman, et al., *Blind testing in firearms: Preliminary results from a blind quality control program*, J. For. Sci. (2022), https://doi.org/10.1111/1556-4029.15031.

casework. Much of the current research, however, does not acknowledge the potential impact of the Hawthorne effect.

One possible example of how this could impact study results lies in the high percentage of inconclusive responses seen in many of the error rate studies. It is possible that the examiners are modifying their behavior and reaching inconclusive decisions at a higher rate because they know the potential effects of a false positive in a validation study. Without further study, the effect of this factor on error rate estimates is unknown.

2. Attrition

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.[35]

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

[35] 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), Jae Kwang Kim & Jun Shao, *Statistical methods for handling incomplete data* (2014), and National Research Council, *The Prevention and Treatment of Missing Data in Clinical Trials* (2010) The National Academies Press. https://doi.org/10.17226/12955.

the impact of nonresponse and drop-out bias.[36] More troubling, these studies do not release any data to facilitate other researchers filling in these gaps. In other scientific disciplines, such as medicine, these oversights would likely render a study unpublishable. Analysis of the effect of participant attrition on the calculated statistics of interest would be required in most pharmaceutical studies. In addition, the convention in many disciplines is that data are made available upon request, or, more commonly, are published alongside the paper in a repository to ensure the data are preserved for future study. That the researchers in firearms and toolmarks do not publish their data or release it to interested researchers is a demonstration of the distance between the status quo in this discipline and the scientific method as it is practiced in most other disciplines; even some studies in other pattern matching disciplines, such as handwriting, have published participant responses in anonymized form.[37] This is of particular concern in cases such as the Ames II study, where the analysis methods used are questionable, and researchers are not willing to release the data upon request despite the study being funded by agencies within the federal government.

---

[36] 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).

[37] R. Austin Hicklin, Linda Eisenhart, Nicole Richetellia, Meredith D. Miller, Peter Belcastro, Ted M. Burkes, Connie L. Parks, Michael A. Smith, JoAnn Buscaglia, Eugene M. Peters, Rebecca Schwartz Perlman , Jocelyn V. Abonamah, and Brian A. Eckenrode, *Accuracy and Reliability of Forensic Handwriting Comparisons*, 119 PNAS 132 (2022).

3. Sampling Bias

Even well designed and well executed studies cannot compensate for sampling bias in the participant pool. If the participants in the study do not make up a representative sample of the population (in this case of firearms, ammunition, and toolmark examiners in the United States), the results of a study cannot be generalized. If we want to understand the distribution of colors in a standard bag of M&Ms, we would not want to take a sample from a bag of Christmas M&Ms that only include red and green, because this would result in a biased sample; the results could not be generalized to all M&Ms because the sample was taken from a different population.

Current studies rely on participants to volunteer. This is, of course, consistent with the practices of many other disciplines, such as clinical trials in medicine. With a self-selected sample of participants, however, it becomes even more critical to take steps to ensure the participants are representative of the population of interest. Individuals, for example, who have the interest, time, and / or lower caseloads to participate in studies may not be representative of the wider population of firearms examiners. In addition, some studies[38] exclude examiners who are not actively working on casework, including expert witnesses who were firearms examiners by training. These different inclusion criteria result in differences in the appropriate populations the study results might generalize to. This issue is entirely unaddressed in current error rate studies.

---

[38] Baldwin et al., *supra* note 15.

In addition to worries about sampling bias among participants in the studies, the firearms studied must also be representative of the population seen in case work. The studies to date tend to focus on a single firearm[39] or a small sample of firearms. Neither scenario supports extrapolating the results to the entire population of possible firearms.

While many scientific journals and funding agencies rely on peer review to identify and correct these issues, the review which takes place in trade journals such as the AFTE journal does not necessarily catch and correct issues with the description and presentation of study results. Even in cases where statisticians are on the research team, such as Ames I and II,[40] the issue of sampling bias is not addressed, perhaps because the two reports referenced were not peer-reviewed. While the lack of concern with sampling bias in forensic science seems to be a cultural problem within the forensic community, it is a problem when the legal system requires general scientific acceptance of forensic methods. Until forensic science is held to the same standards as other scientific disciplines, it will not meet the bar of "general acceptance" within the scientific community.

[39] Baldwin et al., *supra* note 15.
[40] *Id.*; Bajic et al., *supra* note 21.

CONCLUSION

While firearms and toolmark analysis studies have improved, the current state of the discipline is still well below thresholds of scientific validity applicable to other disciplines. Studies attempting to establish firearm and toolmark examination error rates are plagued with problems in statistical design, participant selection, statistical analysis, and the interpretation of estimates. As scientists, we support continued research in this area, but caution against interpreting the currently reported error rates as reliable in light of the problems we have highlighted here.

CERTIFICATION OF WORD COUNT AND COMPLIANCE WITH RULE 8-112

1. This brief contains 6,338 words, excluding the parts of the brief exempted from the word count by Rule 8-503.
2. This brief complies with the font, spacing, and type size requirements stated in Rule 8-112.

Dr. Michael Rosenblum*

*Signed with permission

CERTIFICATE OF SERVICE

In accordance with Maryland Rules 20-201(g)(3) and 20-404(c), I certify that on this day, September 2, 2022, the brief is being delivered by courier or first-class mail to the State and counsel for Kobina Ebo Abruquah:

J. Bradford McCullough
Stanley J. Reed
Assigned Public Defenders
7600 Wisconsin Avenue, Suite 700
Bethesda, MD 20814

Andrew J. DiMiceli
Assistant Attorney General
Office of the Attorney General
Criminal Appeals Division
200 St. Paul Place
Baltimore, MD 21202

Michael Rosenblum (BLZ)

Dr. Michael Rosenblum*

*Signed with permission