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.

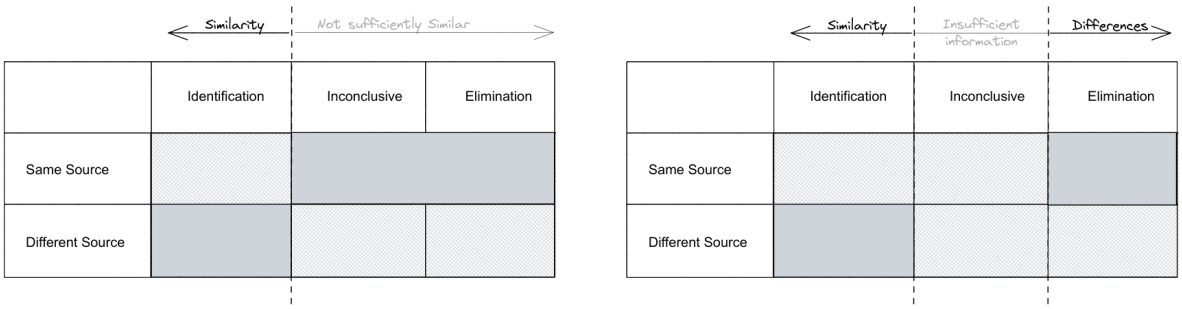


Figure 1: If examiners only spot similarities, then the classification scheme on the left is appropriate and examiners should confine themselves only to claiming to be able to make identifications, grouping inconclusives and eliminations together as having insufficient similarity to make an identification. 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 |
|---|---|---|---|