Shining a Light on Forensic Black-Box Studies

Kori Khan* and Alicia L. Carriquiry**

*Department of Statistics, Iowa State University, Ames, IA; **Center for Statistics and Applications in Forensic Evidence (CSAFE), Iowa State University, Ames, IA

ABSTRACT

Forensic science plays a critical role in the United States criminal legal system. For decades, many feature-based fields of forensic science, such as firearm and toolmark identification, developed outside the scientific community’s purview. The results of these studies are widely relied on by judges nationwide. However, this reliance is misplaced. Black-box studies to date suffer from inappropriate sampling methods and high rates of missingness. Current black-box studies ignore both problems in arriving at the error rate estimates presented to courts. We explore the impact of each type of limitation using available data from black-box studies and court materials. We show that black-box studies rely on unrepresentative samples of examiners. Using a case study of a popular ballistics study, we find evidence that these nonrepresentative samples may commit fewer errors than the wider population from which they came. We also find evidence that the missingness in black-box studies is non-ignorable. Using data from a recent latent print study, we show that ignoring this missingness likely results in systematic underestimates of error rates. Finally, we offer concrete steps to overcome these limitations. Supplementary materials for this article are available online.

Forensic science is a central part of the United States (U.S.) criminal justice system. However, in the last two decades, it has become apparent that faulty forensic science has caused gross miscarriages of justice (Garrett and Neufeld 2009; Bonventre 2021; Fabricant 2022). This occurred partly because many forensic disciplines were developed by and for entities outside of the scientific community. A 2009 report commissioned by the National Academy of Sciences highlighted that, with the exception of DNA, no forensic science field had been empirically shown to be consistent and reliable at connecting a piece of evidence to a particular source or individual (National Research Council (U.S.) 2009). This problem was particularly concerning for feature-based comparison methods, such as latent print analysis, firearm and toolmark identification, and footwear impression examinations because these methods are not rooted in science but rather in subjective, visual comparisons.

In 2016, a technical report by the President’s Council on Advisors on Science and Technology (PCAST) highlighted that some feature-based comparison methods, like bitemarks, were known to be invalid and were still being used in U.S. courts (President’s Council of Advisors on Science and Technology 2016). It also pointed out that there was still no empirical evidence other methods in use were valid. The report stated that empirical testing through “black-box” studies is the only scientific way to establish the validity of feature-based comparison methods. In the years that followed, the PCAST report spurred a number of black-box studies in a variety of fields.

These studies immediately found their way into the U.S. criminal justice system, at times before peer review. Judges frequently use them when deciding whether or not results from feature-based comparison methods should be admissible. Currently, all federal courts and the majority of state courts evaluate expert scientific testimony by the Daubert standard (or a modified form of it) (Daubert v. Merrell Dow Pharmaceuticals, Inc. 1993; Kumho Tire Co. v. Carmichael 1999; Fed. R. Evid. 702). Under the Daubert standard, a trial judge must assess whether an expert’s scientific testimony is based on a scientifically valid methodology that has been properly applied to the facts at issue in the trial. The judge is asked to consider a number of factors. One is whether the theory or technique in question has a “known or potential” error rate. A “high” known or potential error rate would weigh in favor of excluding the testimony at a criminal trial.

Across disciplines, the vast majority of times the “known or potential” error rate is called into question, judges find error rates are low enough to favor admission (e.g., United States v. Cloud 2021; there are exceptions, see, for example, United States v. Shipp 2019). The results of black-box studies are frequently used to support such findings. In relying on the results from black-box studies to evaluate a forensic technique’s admissibility, U.S. judges are explicitly told that the conclusions from the black-box studies can be generalized to results obtained by examiners in the discipline in question. For example, a recent ballistics black-box study asserted that...
“[This] study was designed to provide a representative estimate of the performance of F/T [firearm and toolmark] examiners who testify to their conclusions in court” (Monson, Smith, and Bajic 2022). The Federal Bureau of Investigation’s Lab used this study to support a claim to the Court that “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 rate” (Federal Bureau of Investigation 2022).

Unfortunately, these statements are false. Our review of existing black-box studies found that current studies rely on nonrepresentative, self-selected samples of examiners. They also all ignore high rates of missingness, or nonresponse. These flaws, individually and jointly, preclude any statement about discipline-wide error rates. But perhaps more problematically, we also found evidence that, in some cases, these problems work to systematically underestimate the error rates presented to judges.

The rest of the article proceeds as follows. In Section 1, we introduce the concept of black-box studies and the accuracy measures they consider. In Section 2, we review the methods used to select examiners for participation in black-box studies. Using a popular ballistics black-box study as an illustrative example, we show that these methods lead to unrepresentative samples of participants. For this case study, we also explore how the methods employed could contribute to lower error rate estimates than may be present in the wider discipline. In Section 3, we explore the extent of the missing data problem in black-box studies. To put these problems in context, we give a brief overview of the analysis of missing data. Then, in Section 4, we use the experimental design and data (to the extent they are available) from two black-box studies. We find evidence that examiners who commit a disproportionate number of errors also have disproportionately high nonresponse rates in black-box studies. Using simulation studies, we show that ignoring this kind of missingness could result in gross underestimates of error rates. We also highlight how misleading the current standards for reporting results in black-box studies are. Finally, in Section 5, we offer concrete steps to address these limitations in future studies.

### 1. Black-Box Studies

Forensic feature comparison disciplines are difficult to evaluate empirically. These disciplines, which include latent print analysis, firearm and toolmark examination, and footwear impression examination, rely on inherently subjective methods. In response to this complication, President’s Council of Advisors on Science and Technology (2016) stated that the scientific validity of these disciplines could only be evaluated by multiple, independent “black-box” studies.

In a black-box study, researchers accommodate the subjectivity of the method by treating the examiner as a “black-box.” No data are collected on how an examiner arrives at his/her conclusions. Instead, the examiner is presented with an item from an unknown origin and an item(s) from a known source and is asked to decide whether the items from known and unknown sources came from the same source. Although the details of arriving at such a decision can vary by discipline, the general steps are the same. First, the examiner assesses the quality of the sample of unknown origin to determine whether it is suitable for comparison. Many disciplines have a categorical outcome for this stage. For example, latent print examiners often use three categories: value for individualization, value for exclusion only, or no value (Ulery et al. 2011; Eldridge, De Donno, and Champod 2021). If the item is deemed suitable for comparison, the examiner then arrives at another categorical conclusion about the origin of the unknown. The categories of this conclusion typically include Identification (the samples are from the same source), Exclusion (the samples are from different sources), or Inconclusive (with different disciplines having various special cases of inconclusive).

The PCAST report stated that empirical studies must show that a forensic method is accurate, repeatable, and reproducible for the method to have foundational validity (President’s Council of Advisors on Science and Technology 2016). Accuracy is measured by the rate at which examiners obtain correct results for samples from the same source and different sources. Repeatability is measured by the rate at which the same examiner arrives at a consistent conclusion when re-examining the same samples. Reproducibility is the rate at which two different examiners reach the same conclusion when evaluating the same sample. All current black-box studies assess accuracy, and accuracy measures are the most frequently used measures in courts (United States v. Shipp 2019; United States v. Cloud 2021; Federal Bureau of Investigation 2022). The studies that assess repeatability and reproducibility do so after assessing accuracy. Typically, researchers recruit participants and give a set of items for comparison to assess accuracy. Then, they select (or request volunteers from) a subset of the original participants and redistribute new (potentially repeating) items for comparison in the repeatability or reproducibility stage. Because of this setup, the issues we discuss in this article always apply to repeatability and reproducibility in addition to accuracy. For simplicity, we restrict our attention to accuracy measures.

Accuracy is typically quantified with four measures: (a) the false positive error rate, (b) the false negative error rate, (c) sensitivity, and (d) specificity. However, the error rates tend to be considered the most important accuracy measures, so we focus on these (Smith, Andrew Smith, and Snipes 2016). The false positive error rate focuses on different source comparisons. Researchers typically divide the number of items for which examiners incorrectly concluded “Identification” by the number of total different source comparisons. The false negative error rate focuses instead on the same source comparisons and is defined similarly.

Importantly, items marked inconclusive are almost always considered “correct” comparison decisions. This practice apparently originated in the late 1990s among the feature comparison community. The Collaborative Testing Service (CTS) treated inconclusives as errors until approximately 1998 (Klees Expert Trial Transcript 2002). The decision to change the treatment of inconclusives was seemingly influenced, in part, by high error rates in ballistics (Klees Expert Trial Transcript 2002). Treating inconclusives as “correct” is problematic (for more details, see Hofmann, Vanderplas, and Carriquiry 2021; Dorfman and Valiant 2022a). However, there is no agreement on how to handle
inconclusives in analyses (see, e.g., Weller and Morris 2020). In this article, we treat inconclusives as “correct” decisions. We do so to view the impacts of sampling and nonresponse bias in the light most favorable to the error rates currently being reported to courts across the United States.

The PCAST report, and others, have pointed out several desirable features of black-box studies; for example, open set design, blind testing, and independent researchers overseeing the study (President’s Council of Advisors on Science and Technology 2016). In the wake of the PCAST report, authors of many black-box studies assume that if they have addressed these elements, they can use their studies’ results to make statements about the discipline-wide error rate. However, in this article, we assume that all the desirable features described in the PCAST report have been met. We show that even for a study where this is the case, the current methods of sampling examiners and handling nonresponse rates preclude any such conclusions.

2. Unrepresentative Samples of Examiners

The inferential goal of current black-box studies is to make a statement about either the discipline-wide error rate or the average examiner's error rate in a specific discipline (see, e.g., Chumbley et al. 2021; Smith 2021; Eldridge, De Donno, and Champod 2021; Monson, Smith, and Bajic 2022). In other words, these studies wish to take observations made on a sample of examiners and arrive at a conclusion about the broader population of examiners to which they belong.

The power of well-done statistics is the ability to do precisely that: take observations made on a sample and generalize these observations to a wider population of interest. However, this ability comes at a price: valid sampling methods must be used to ensure that the sample selected to participate in the study is representative of the larger population. The gold standard for achieving a representative sample is random sampling (Fisher 1992; Levy and Lemeshow 2013). In random sampling, members of the population are selected, with known probability, to be included in the study by the researcher.

Random sampling is desirable for many reasons. For example, it is necessary for many standard statistical techniques. However, most random sampling methods require, at least theoretically, that it be possible to enumerate the population of interest. There are many cases where this is not feasible or practical. The inability to use random sampling does not always preclude researchers from generalizing to the population of interest (e.g., Smith 1983; Elliott and Valliant 2017). In order to make such generalizations with nonrandom sampling, however, something must generally be known about the population of interest, and care must be taken to avoid known sources of bias.

One such source of bias in nonrandom sampling methods is selection bias. Selection bias occurs when the sampling method results in samples that systematically over-represent some members of the underlying population. This can result in biased estimates for the parameter of interest, such as the error rate. One of the most well-studied types of selection bias comes in the form of self-selection, or volunteer, bias. This occurs when the researcher does not choose the population members to be in the study but instead allows members of the population to “volunteer” to participate. Research in various fields has indicated that this typically results in unrepresentative samples because those who volunteer to participate tend to be different than those who do not volunteer (e.g., Strassberg and Lowe 1995; Ganguli et al. 1998; Taylor, Cahn-Weiner, and Garcia 2009; Jordan et al. 2013; Dodge et al. 2014).

2.1. Sample Selection in Black-Box Studies

For most feature-comparison disciplines, little is known about the population of examiners analyzing forensic evidence. The standards for serving as an expert witness in a U.S. trial are not high. For example, two days (or less) of formal training in a discipline can be sufficient to qualify as an expert in a forensic science discipline (see, e.g., State v. Moore 1991). Few states have established regulatory boards to establish minimum requirements to be an examiner. This, unfortunately, means that it is very difficult to determine what a representative sample of examiners might look like.

No black-box study attempts to address this information gap. Instead, to our knowledge, every black-box study to date has used self-selected samples of examiners (see, e.g., Ulery et al. 2011; Baldwin et al. 2014; Smith, Andrew Smith, and Snipes 2016; Richetelli, Hammer, and Speir 2020; Chumbley et al. 2021; Eldridge, De Donno, and Champod 2021; Hicklin et al. 2021; Smith 2021; Guttmann et al. 2022; Hicklin et al. 2022). These self-selected participants are typically solicited through E-mail listservs for one or more professional organizations. The black-box studies for some disciplines, like latent prints, often accept every volunteer examiner. While this does not alleviate the probable self-selection bias, it ensures that the researchers are not excluding self-selected examiners based on qualities that may be related to error rates.

On the other hand, black-box studies in other disciplines, such as firearm and toolmark identification, often impose inclusion criteria that reasonably could be expected to be related to error rates. For example, Monson, Smith, and Bajic (2022) reports that the ballistics study we refer to as the FBI/Ames study used self-selected examiners. However, the researchers also restricted participation in the study to “fully qualified examiners who were currently conducting firearm examinations, were members of [the Association of Firearm and Tool Mark Examiners] AFTE, and were employed in the firearms section of an accredited public crime laboratory within the U.S. or a U.S. territory.” They also excluded FBI employees to avoid a conflict of interest.

Other than perhaps for the word “qualified,” none of these criteria are directly related to a characteristic required of examiners who examine firearm and toolmark evidence or testify about it in U.S. courts. For example, AFTE is a professional organization. To our knowledge, no court has ever deemed membership in AFTE necessary to qualify an examiner as an expert witness. Additionally, many privately employed (or self-employed) examiners are actively testifying.

There has never been an attempt to assess whether inclusion criteria used in black-box studies are representative of examiners. Here, we explore whether the FBI/Ames study criteria would have excluded examiners currently conducting firearm
and toolmark examinations for trials. We used Westlaw Edge’s collection of expert witness materials to identify 60 unique expert witnesses whose curriculum vitae (CV) indicated the witness was an expert with respect to firearm and toolmark identification (see SI section 1 for more details). These CVs cannot be viewed as a representative sample of expert witnesses. Among other problems, Westlaw Edge’s materials tend to be disproportionately from federal jurisdictions. However, they are still useful to explore whether the inclusion criteria used by the FBI/Ames study match the characteristics of expert witnesses who have interacted with courts.

For each of the 60 expert witnesses, we assessed whether each examiner was a current AFTE member and whether he/she worked for a private or public employer (see SI section 1 for more details). In Table 1, we found that just over 60% of the expert witnesses met each criterion separately, but only 38.3% met both. In other words, with fewer than two of the inclusion criteria used in the FBI/Ames study, the majority of these expert witnesses would have been excluded from participation.

More problematically, some of the inclusion criteria used in black-box studies are either known to be related to or could reasonably be related to error rates. Using the FBI/Ames study as an example again, this study excludes foreign examiners. Yet, foreign examiners can provide testimony in court, and the results of studies on foreign examiners are used to support the admissibility of firearm and toolmark examinations (see Federal Bureau of Investigation (2022) citing Kerkhoff et al. 2018). However, foreign examiners have also been linked to higher error rates than U.S. examiners in other disciplines (e.g., in latent palmar prints, see, Eldridge, De Donno, and Champod 2021). The exclusion of foreign examiners thus should not be done lightly. Additionally, forensic laboratory accreditation requires proficiency testing of examiners and for the lab to assess and certify that the examiners are competent (Doyle 2020). The expectation would be that this process reduces errors. However, examiners are not required to work for an accredited lab to present testimony at trial. Thus, these criteria make it likely that sampling bias is present. In the case of the FBI/Ames study, the particular inclusion criteria used also suggest that this bias works to result in underestimations of error rates.

### Table 1. % of 60 expert witnesses satisfying FBI/Ames criterion.

| Criterion | % Experts |
|-----------|-----------|
| Current AFTE member | 63.3% |
| Public Employer | 65.0% |
| Current AFTE member & Public Employer | 38.3% |

3. **Unit and Item Nonresponse in Black-Box Studies**

In this section, we borrow from survey methodology to distinguish between unit and item nonresponse (see, e.g., Little and Rubin 2019, pp. 5–6; Elliott et al. 2005, p. 2097). Unit nonresponse occurs when a participant who agreed to participate in a study does not respond to a single assigned comparison. Item nonresponse occurs when a participant responds to at least one but not all assigned comparisons. The presence of either unit or item nonresponse can lead to bias and a loss of power in statistical analyses which do not account for this missingness (Rubin 1976; Groves 2006). These problems can be exacerbated in studies, like many black-box studies, where both types of nonresponse are present. Yet, to our knowledge, no one has ever attempted to formally analyze the patterns of missingness in black-box studies or to adjust error rate estimates to account for it.

#### 3.1. The Nomenclature of Missing Data

To adjust for nonresponse, researchers must first explore the patterns of missingness in their data. Statistical methods to address missing data depend heavily on the mechanisms that caused the missingness to arise. Rubin (1976) was the first to formalize missing-data mechanisms, but the nomenclature has since changed (Kim and Shao 2014; Little and Rubin 2019). Currently, missing data mechanisms are described as falling into one of three categories: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) (Kim and Shao 2014; Little and Rubin 2019).

To understand the differences in the context of black-box studies, we define \( Y \) to be the \( n \times K \) matrix of complete data. This matrix would include, at a minimum, the response for every assigned item. It could also include auxiliary information about the item or responding examiner. We let \( M \) be the \( n \times K \) matrix where \( M_{ij} \) is 1 when \( Y_{ij} \) is missing and 0 otherwise. Finally, we let \( \phi \) be a set of unknown parameters on which the missingness mechanism depends. In the best-case scenario, whether data are missing should not depend on any element of \( Y \). More precisely, the following would hold:

\[
    f(M|\psi) \equiv f(M|\phi) \forall Y, \phi.
\]

In this case, the missingness would be MCAR. If, instead, the missingness depended only on observed values of the data matrix \( Y \), that is,

\[
    f(M|\psi) \equiv f(M|Y_{\text{obs}}, \phi) \forall Y, \phi,
\]

then we would call it MAR. The most problematic missingness mechanism is NMAR. This occurs when \( f(M|Y, \phi) \) cannot be simplified to the right-hand sides of either (1) or (2).

For many analytical goals, MCAR and MAR do not typically result in biased estimates. Instead, they primarily affect the uncertainty associated with such estimates. In the context of black-box studies, point estimates obtained for error rates may not be systematically skewed from the underlying population error rate. Instead, the confidence intervals associated with these point estimates may need to be adjusted. For sufficiently low nonresponse rates, some researchers consider ignoring MCAR (and sometimes MAR) missingness acceptable. This is part of the reason why MCAR and MAR are often referred to as ignorable missingness. There is no consensus on what “sufficiently low” means—rules of thumb range from 5% to 10% (Schafer 1999; Bennett 2001). Even with ignorable missingness, however, a sufficiently high nonresponse rate can become problematic (Madley-Dowd et al. 2019). Some researchers have proposed that nonresponse rates above 40% should always preclude generalizations to a wider population (Dong and Peng 2013; Jakobsen et al. 2017).

Unlike MAR and MCAR, NMAR can lead to bias in the point estimates and distortions in uncertainty estimates even with low
rates of nonresponse. It is inappropriate to ignore this type of missingness in statistical analyses. As a result, NMAR is often referred to as non-ignorable missingness.

### 3.2. Adjusting for Nonresponse

There are numerous methods to adjust for both unit and item nonresponse. The appropriate method will depend on the type of missingness mechanisms present. Assessing this will typically require a thorough analysis of all collected data. When in doubt about the missingness mechanisms, many researchers recommend conducting sensitivity analyses to assess the potential impacts of different types of mechanisms (Pedersen et al. 2017).

Most simple methods for handling missing data only result in unbiased estimates if the missing mechanism is MCAR. These methods include complete case analysis and available case analysis. In a complete case analysis, an analysis is only carried out on cases where the full set of analysis variables is observed. An available case analysis, on the other hand, uses all data available about the analysis variables. We emphasize these approaches are only appropriate when the missingness is MCAR (Schafer and Graham 2002). Even in this case, standard errors for the estimators can be adversely impacted.

Unfortunately, the missingness in real-world data of human subjects is rarely MCAR (see e.g., Mislevy and Wu 1996; Jakobson et al. 2017). When the missingness is not MCAR, the methods of adjusting for missingness almost always require auxiliary information. For MAR approaches, this auxiliary information can be limited to assumptions about the relationship between missingness in one analysis variable and other observed analysis variables. However, methods involving NMAR approaches almost always require auxiliary information beyond the analysis variables (Groves 2006; Riddles, Kim, and Im 2016; Franks, Airoldi, and Rubin 2020).

### 3.3. Implications for Black-Box Studies

Black-box studies are plagued by both unit and item nonresponse. We emphasize that we treat inconclusive decisions as observed in this discussion. However, the authors of black-box studies pay little attention to the nonresponse. In fact, they often fail to release enough information to even calculate the relevant nonresponse rates.

Ideally, it would be possible to calculate the nonresponse rates for each analysis conducted. For example, to calculate the unit and nonresponse rates for a false positive error rate, we need to know the number of different source items assigned to participants. At the time of writing, only two black-box studies have released sufficient information to calculate item nonresponse rates for both false positive and false negative error rate estimations (see, Eldridge, De Donno, and Champod 2021; Hicklin et al. 2022).

In Table 2, we provide unit and item nonresponse rates for black-box studies aggregated over all potential analyses. This table is limited to black-box studies that released sufficient information to calculate both unit and item nonresponse rates (see SI section 2). These unit nonresponse rates reflect those seen in other black-box studies that do not release sufficient information to calculate item nonresponse (see SI section 2 for more details about these and other studies). We note that unit nonresponse rates are often over 30%, far above any definition of “low” nonresponse rates (see e.g., Ulery et al. 2011; Richetelli, Hammer, and Speir 2020; Eldridge, De Donno, and Champod 2021; Smith 2021).

In Table 2, we have calculated nonresponse rates assuming that any recorded response is observed, even if no comparison decision was rendered. However, practically speaking, only a recorded comparison decision can be used to estimate error rates. If only comparisons with a comparison decision are considered observed, the item nonresponse could be much higher than those given in Table 2.

To illustrate this, we consider one of the two black-box studies that have released sufficient information to calculate item nonresponse rates for the false positive rate. In Eldridge, De Donno, and Champod (2021), 328 participants enrolled in the study, and 226 actively participated. Each participant was assigned 22 different source comparisons. Twenty-five of the 226 active participants failed to start all 22 different source comparisons assigned to them. Another 3 participants stated the image was of no value for every different source comparison they responded to. Finally, another 13 failed to render a single comparison decision for their assigned different source comparisons, despite finding at least one such comparison to be suitable for comparison. Out of the original 328 enrolled participants, only 185 rendered a comparison decision on at least one different source comparison. For the purpose of calculating the false positive error rate, there was a unit nonresponse rate of $(1 - \frac{185}{328}) \times 100 = 44\%$. The 185 participating examiners rendered 2560 comparison decisions for their assigned $185 \times 22 = 4070$ different source comparisons. The item nonresponse rate was thus $(1 - \frac{2560}{4070}) \times 100 = 37\%$ for the false positive error rate calculations in this study (where we treat inconclusives as observed). As we have discussed, such high rates of missingness preclude generalization to a broader population of examiners.

Despite the high nonresponse rates, authors of black-box studies have yet to report examining the patterns of missingness in their data. Black-box studies also do not attempt to adjust for the missingness in their analyses. Instead, there are two ways that black-box studies deal with missing data. Some black-box studies drop a participant from the analysis if the participant did not answer all items assigned to him/her (Ulery et al. 2011; Hicklin et al. 2021). This is an example of complete case analysis discussed in Section 3.2, and it is only appropriate for MCAR missingness. The second way black-box studies handle missingness is to analyze on the observed responses, ignoring the nonresponse. This approach is marginally better than dropping participants with missing items and is most similar to an

| Study                          | Unit      | Item       |
|-------------------------------|-----------|------------|
| Ulery et al. (2011) (Prints)   | 33.5%     | 0.2%       |
| Baldwin et al. (2014) (Cartridges) | 23.2% | 0.06%      |
| FBI/Ames Study (Bullet)       | ≥ 32.4%   | 35.6%      |
| FBI/Ames Study (Cartridges)   | ≥ 32.4%   | 35.1%      |
| Eldridge, De Donno, and Champod (2021) (Palmar Prints) | 31.1% | 27.5%      |
| Richetelli, Hammer, and Speir (2020) (Footwear) | 33.0% | 9.6%       |
available cases approach. However, again this is only potentially defensible if the missingness is MCAR.

It is highly unlikely that the missingness in black-box settings is MCAR. Technically, MCAR missingness would require a random sample (as opposed to self-selected samples) of examiners. Outside of forensic science, missingness is typically assumed to be potentially non-ignorable in testing settings (Mislevy and Wu 1996; Pohl, Gräfe, and Rose 2014; Dai 2021). In such settings, Mislevy and Wu (1996) suggests "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." If an examinee is proficient enough to know when he/she are likely to be incorrect, then this type of behavior will lead to an underestimate of error rates if missingness is ignored.

To appropriately adjust for missingness in black-box studies, researchers will likely need auxiliary information. This information could come in the form of examiner characteristics or item characteristics. Many black-box studies collect this type of information, but only one has released any portion of such data to the public in a meaningful way. Indeed, most black-box studies fail to release any de-identified data at all. Instead, as we will explore in Section 4.2, they give aggregated summaries that could be misleading in the context of high nonresponse rates.

The authors of black-box studies typically reject the possibility of non-ignorable missingness. Many state that missingness occurs because examiners are too busy to participate in the study or, alternatively, to complete all items if they do participate (Private communication with Heidi Eldridge 2021). However, it is possible for missingness to be non-ignorable and for non-response to be due to examiners being too busy. For example, busy examiners may choose to respond to the easier item comparisons as they take less time. Rather than speculating, however, the appropriate step would be to assess the patterns of missingness in the data. As no one has done this, we offer the first such attempt to do so now.

4. Case Studies

In this section, we focus on the potential impact of item nonresponse on error rate estimates in black-box studies. As previously referenced, only two black-box studies have released data in a form that an independent researcher can analyze (i.e., the data are released with enough detail about the study design that, at a minimum, the item nonresponse rates can be calculated for individual analyses). In Section 4.1, we use one of these datasets to explore some of the patterns of item nonresponse. We show there is evidence of non-ignorable missingness. Using that insight, we use simulation studies to replicate the FBI/Ames study's exploration of false positive error rates in bullet comparisons in Section 4.2 and highlight how misleading the current trends of reporting responses in black-box studies can be.

4.1. EDC Study: Palmar Prints

This sub-section focuses on the item nonresponse in a study described in Eldridge, De Donno, and Champod (2021) (hereafter the EDC study). This study assessed the accuracy of latent print examiners' analysis of palmar prints. Each participant was asked questions about his/her demographic information, training, and employer. All participants then received 75 comparisons to complete. The study design assumed participants in this study followed a multi-stage approach to analyzing items. First, examiners were asked to assess the images of the prints for suitability for comparisons. If examiners found an item suitable for comparison, they could enter a conclusion of Inconclusive, Identification, or Exclusion. As part of their analysis, examiners were also asked to rate each item's comparison difficulty. In this section, we treat any response to an item as non-missing. Thus, we consider items marked as not suitable for comparisons as observed.

The authors released information about all items that examiners responded to and demographic information for most participants. They also released fairly rich information about the quality of images for compared items. When asked, they declined to release information about the comparisons examiners did not respond to.

We begin the analysis of item nonresponse with variables that have been previously linked to non-ignorable missingness. As alluded to in Section 3.3, one common pattern of non-ignorable missingness on assessment tests is when examinees fail to respond to items they believe they would answer incorrectly. In this dataset, examiners were asked to rate each item's difficulty on a Likert-type scale, including possible responses of: Very easy/obvious, Easy, Moderate, Difficult, and Very difficult. For the items examiners responded to, 910 were deemed Very easy/obvious, and only 545 were deemed Very difficult. In seven cases, examiners ranked the item difficulty level and then failed to give a comparison decision. These items were all marked as either Moderate or Very Difficult. These patterns suggest that examiners were more likely to respond to items they deemed Very easy and less likely to respond to items they deemed Very difficult. As Eldridge, De Donno, and Champod (2021) (and others outside of black-box studies) have observed, more errors are committed on items ranked as difficult. Thus, there is evidence of non-ignorable missingness here. Ideally, we would have a list of every item assigned to each examiner. While it is not possible to know how a particular examiner would have ranked a particular item, it would be possible to use auxiliary information (e.g., other examiner's rankings or information about the quality of items) to more formally assess whether the items with no response were more likely to be viewed as difficult. However, this data has not been released.

There are other ways to use the released data to formally assess whether there is evidence of non-ignorable missingness. Because we know that each examiner was assigned 75 items, we can calculate each examiner's item nonresponse rate. The study's authors also identified various examiner characteristics they claimed were associated with higher error rates. If the item nonresponse was ignorable, we should not see any relationship between high rates of item nonresponse and characteristics associated with high error rates. We can use permutation tests to formally assess whether a statistically significant relationship exists between high degrees of item nonresponse and examiner characteristics associated with high error rates.

To do this, we restrict our attention to the 197 examiners who had released both demographic information and at least
one response. We define an examiner to have a high degree of item nonresponse if he/she failed to respond to over half of the 75 assigned items. The EDC study identified several characteristics that were associated with high error rates. For example, participants employed outside of the United States made half of the false positive errors in the study despite only accounting for 18.1% of the 226 active study participants. Similarly, the EDC study noted that non-active latent print examiners (LPEs) disproportionately made false positive error rates. Using machine learning approaches, the EDC study also noted that working for an unaccredited lab and not completing a formal training program were weakly associated with higher error rates (lower accuracy) among examiners. We note that these last two observations relied on analyses that, themselves, could have been impacted by missing data. However, we take these findings at face value here.

For the permutation tests, we focus on the four characteristics associated with higher error rates: working for a non-US employer, being a non-active LPE, working for an unaccredited lab, and not completing a formal training program. Because each of the four characteristics under consideration is binary, we can use the same general approach. We explain the methodology by focusing on whether examiners work for a non-US employer.

We consider the following hypothesis test:

- $H_0$: Foreign-employed examiners and U.S.-employed examiners are equally likely to leave over 50% of their items blank.
- $H_A$: Foreign-employed examiners are more likely than U.S.-employed examiners to leave over 50% of their items blank.

Thirty-eight (19.2%) of the 197 examiners worked for non-U.S. employers. Under the null hypothesis, we would expect that approximately 19.2% of the examiners with high rates of missingness worked for non-U.S. employers. Instead, we observe that 28.6% (or 14 examiners) of the 49 examiners with a high degree of item nonresponse worked for non-U.S. agencies.

If the null hypothesis were true, the probability that 28.6% or more of the examiners with a high degree of item nonresponse worked for foreign agencies would be about 4.9%. We refer to this probability as the $p$-value of the hypothesis test specified above. Because it is low, there is weak evidence that examiners working for non-U.S. agencies are not only more error-prone, but they are also more likely to fail to respond to over half of their assigned items. Note, the terminology “weak evidence” comes from historical hypothesis testing, where the decision to fail to reject or reject a null hypothesis was often made based on whether a $p$-value was less than 0.05. This type of decision-making can be problematic (Wasserstein and Lazar 2016). Here, we emphasize that we view this $p$-value as more of a data exploration tool: the closer it gets to zero, the less reasonable it is to assume that the item nonresponse is operating independently of the considered characteristic (in this case, type of employer). Here, we believe a $p$-value of 4.9% warrants further investigation before using analyses that ignore missingness.

We can repeat this procedure for the remaining three characteristics identified by Eldridge, De Donno, and Champod (2021) as associated with high error rates. When we do so, the $p$-values for the corresponding hypothesis tests involving being a non-active LPE, working for a non-accredited lab, and having not completed a formal training program are 46.6%, 3.1%, and 44.7%, respectively. Thus, there is evidence that examiners who work for an unaccredited lab are more likely to fail to respond to over 50% of the items assigned to them than examiners who work for an accredited lab. However, there is no evidence of such a relationship for either being a non-active LPE or having not completed a formal training program. In sum, two of the four characteristics identified by Eldridge, De Donno, and Champod (2021) as being associated with higher error rates also may be associated with higher rates of item nonresponse.

To summarize, there is evidence that examiners are more likely to respond to very easy items than very difficult items. Furthermore, some of the more error-prone examiners also are more likely to leave over half of their items blank than their less error-prone counterparts. These trends are evidence of non-ignorable missingness. The association between high nonresponse and higher error rates suggests that ignoring this missingness will result in underestimates of the associated error rates. However, because there is no information about items each examiner was assigned but chose not to respond to, it’s difficult to develop a formal attempt to adjust for missingness in error rate calculations. Thus, in the next section, we use simulation studies to demonstrate the potential impact of non-ignorable missingness.

### 4.2. FBI/Ames Study: Bullet Comparisons

In this section, we use simulation studies to demonstrate the impact that item nonresponse can have on estimates of error rates in black-box studies. We do not claim that any of the results are an indication of the truth of a particular existing black-box study. The EDC study suggests that missingness likely depends on the characteristics of both examiners and comparisons. There are no released data to account for these things in a formal statistical way. Instead, our primary purpose is to use a simple model to illustrate how misleading the current methods of reporting error and nonresponse rates can be. We base our simulations on the FBI/Ames black-box study previously considered in Section 2.

The FBI/Ames study was conducted by researchers at the Ames Laboratory and the Federal Bureau of Investigation (FBI). The study's purpose was to assess the performance of forensic examiners in firearm and toolmark identifications (Chumbley et al. 2021; Monson, Smith, and Bajic 2022). It aimed to assess accuracy, repeatability, and reproducibility for both bullet and cartridge comparisons. Although the statistical analyses of collected data remained unpublished until October of 2022, the results were (and continue to be) used to support the admissibility of ballistics expert testimony in criminal trials prior to peer-review (United States v. Shipp 2019). Importantly, the repeatability and reproducibility analyses remain unpublished, but continue to be used in courts nationwide (note, these analyses have been challenged by others see, e.g., Dorfman and Valliant 2022b). Initially, the FBI/Ames study released no data capable of being independently analyzed. When we requested the de-identified data to explore patterns of missingness, an Ames lab researcher stated the FBI had not given Ames lab researchers permission to share such data. Our requests for clarification about missingness in these studies remain unanswered. We note that the authors subsequently released some
data from the accuracy stage in Monson, Smith, and Peters (2023) while this article was under review; however, these data were limited to only the observed responses for the accuracy stage. The reporting methods of the FBI/Ames study prior to this partial release of data remain the predominant practice in the field. Because our simulation studies are meant to highlight how misleading such reporting methods can be and are not meant to be a reflection of the truth of the FBI/Ames study, we first focus on the methods used prior to the publication of (Monson, Smith, and Peters 2023). As only results from the accuracy stage are publicly available, we focus on accuracy measures. For simplicity, we further limit our review to the false positive error rates for bullet comparisons.

The FBI/Ames study reported an estimated false positive error rate of approximately 0.7%. This estimate, like all estimates in black-box studies, did not account for missingness. Instead, it was based on the 20 observed false positive errors made in 2842 observed comparison decisions. We note that 2891 decisions for the accuracy stage were recorded, but 49 of the responses were dropped from the analysis by the original authors (see Chumbley et al. 2021).

We estimate the item nonresponse rate for bullet comparisons across all stages of the study is approximately 35%. Our estimate assumes the FBI/Ames study authors gave the 173 self-selected examiners 6 packets (see SI subsection 3.1 for details on this estimate). The FBI/Ames study authors report that each packet contained 15 bullet comparisons, across all stages of the study. There is no published paper reporting the number of packets assigned in the accuracy stage, but the abstract of an unpublished paper (Bajic et al. 2020) reports each examiner was intended to receive 2 packets in the accuracy stage (see SI subsection 3.1 for more details; note that the partially released data in Monson, Smith, and Peters (2023) indicate at least one examiner was assigned more than two packets in the accuracy stage). Thus, published papers do not provide sufficient information to calculate the item nonresponse rate for the accuracy stage of the study generally. To our knowledge, no papers, published or unpublished, currently report the number of different source items assigned in the accuracy stage, which makes it impossible to calculate the item nonresponse for the false positive accuracy error rate specifically.

If missingness is non-ignorable, as the percentage of missing items increases, the bias of the estimate obtained from an analysis that does not account for missingness will increase. To view the impact of the missingness in the light most favorable to the current estimates, we attempt to use the study design to make a reasonable estimate of the item nonresponse for assigned different source comparisons. Specifically, we assume that two packets were assigned to each examiner in the accuracy stage. The FBI/Ames study reports that approximately 2/3rds of all items were different source comparisons. Based on the number of recorded responses for different source comparisons and these assumptions, the item nonresponse for different source bullet comparisons was at least 17.9% (see SI subsection 3.1 for more details on the steps of this estimation).

We design a set of simulation studies to mirror the described experiment. To simulate our data, we let \( Y_{ij} \) be an indicator of whether examiner \( i \) makes an error on item \( j \). We define \( M_{ij} \) to be an indicator of whether examiner \( i \)'s response to item \( j \) is missing. We assume that there are 173 participants who are each given 20 comparison items. We generate the data in the following way:

\[
P(Y_{ij} = 1) \sim \text{Bern}(p_i),
\]

\[
P(M_{ij} = 1|Y_{ij} = 1) \sim \text{Bern}(\pi_i),
\]

\[
P(M_{ij} = 1|Y_{ij} = 0) \sim \text{Bern}(\theta_i),
\]

for \( i = 1, \ldots, 173, j = 1, \ldots, 20 \). We note that one way missingness could be ignorable is if \( \pi_i = \theta_i \) for all \( i \). However, given our results in the EDC case study, we explore how a range of non-ignorable missingness mechanisms could impact inference. Throughout all possible scenarios, we ensure that there is always approximately 17.9% item nonresponse.

The parameter \( \pi_i \) represents the probability that an examiner fails to respond to an item on which he/she would have made an error. An examiner is more likely to have a missing response for an item they would have made an error on as \( \pi_i \) increases in value. We let \( \pi_i = \pi \) be constant across individuals. We note that this is likely not the case in real black-box studies: The EDC data suggest that examiners with different error rates respond at different rates (i.e., likely have different \( \pi_i \)s). However, this simulation study is meant to give a simple illustration of the wide range of possibilities in observed datasets when missingness is ignored, so we proceed with this simple model. We consider 101 potential values for \( \pi \), varying between 0 and 1. For each value of \( \pi \), we simulate responses for all 173 examiners on 20 items.

In Chumbley et al. (2021) the authors stressed there was evidence error rates were not constant across examiners. They also noted that 10 of the 173 examiners committed all 20 of the false positive errors. To mirror this, \( p_i \) is randomly generated from a uniform distribution on \([0,0.007]\) for the first 163 examiners. For the other 10 examiners, \( p_i \) is randomly generated from a uniform distribution on \([0.55,0.6]\). In this way, each examiner has his/her own error rate, and we also reflect the pattern observed in the FBI/Ames study of having 10 examiners more error-prone than the others. To ensure that approximately 17.9% of the items are missing, \( \theta_i \) is chosen as a function of \( \pi \) and \( p_i \) (see SI subsection 3.1 for more details).

For each value of \( \pi \), we really have two datasets: the "observed" dataset (restricted to cases where \( M_{ij} = 0 \)) and the "full" dataset. We calculate the false positive error rate estimate and the 95% confidence interval for both the observed data and the full data. The FBI/Ames study used a beta-binomial model to arrive at the error rate estimates and 95% confidence intervals reported in Chumbley et al. (2021). The authors state the estimates were produced with R packages including VGAM (Yee and Wild 1996; Yee 2010, 2015). The beta-binomial model in VGAM is numerically unstable, as the expected information matrices can often fail to be positive definite over some regions of the parameter space (Yee 2023). This makes it unsuitable for simulation studies. Instead, we use the Clopper-Pearson estimator (Clopper and Pearson 1934). In settings like this simulation study, the primary difference will typically be in the width of the confidence intervals- with the beta-binomial model expected to be wider. For example, the Clopper-Pearson estimate and confidence interval for the FBI/Ames false positive error rate are
The full error rate estimate is 3.8% when these simulations, the observed error rate estimate is 0%, and the corresponding estimates from the beta-binomial model are 0.7% and (0.3%, 1.4%).

Table 4. Clopper-Pearson error rates and CI for observed and full datasets.

| Data                      | FP rate | 95% CI     |
|---------------------------|---------|------------|
| FBI/Ames (Obs. Data)      | 0.7%    | (0.4%, 1.1%) |
| Simulated (Obs. Data)     | 0.7%    | (0.4%, 1.1%) |
| Simulated (Full)          | 3.6%    | (3.0%, 4.3%) |
| FBI/Ames (Full)           | ?       | ?          |

With sufficient data, it may be possible to rule out some of the more extreme discrepancies between the estimates obtained for the observed data and the full data. However, black-box studies do not report such details. For example, before Monson, Smith, and Peters (2023), the FBI/Ames study only reported the number of examiners with 0 false positives, 1 false positive, and 2 or more false positives, similar to Table 3 (they also report the equivalent for false negatives) (Bajic et al. 2020). We calculated the same type of summary for each of our observed simulated datasets. As shown in Table 3, we obtained summary statistics equivalent to those reported in the FBI/Ames study when \( \pi = 0.87 \).

For this observed dataset, just like in the FBI/Ames study, there are 20 errors made by only 13 examiners. The Clopper-Pearson estimate and 95% confidence intervals were equivalent to those for the FBI/Ames study, as shown in Table 4. However, if the error rate had been calculated on the corresponding full dataset, it would have been 3.6% instead of 7%. In other words, the “true” false positive error rate would have been over 414% greater than the reported one.

We emphasize here that even if the item nonresponse rate was different than 17.9% or the missingness mechanism is not similar to the one explored in our simulation studies, the general principles from these simulation studies would hold.

We take a moment to compare these simulation studies to the partially released information for the accuracy stage in Monson, Smith, and Peters (2023). We applaud the authors of the FBI/Ames study for releasing some information, but we note the information released is still inadequate for meaningful exploration of the missingness. To explore missingness, the unobserved is as important as the observed. The data released with Monson, Smith, and Peters (2023) included only the observed responses for examiners (see SI subsection 3.1 for more information), and no further information was released about the assigned items that did not receive a response. We note the information released was sufficient to illustrate that the simple simulation studies are not representative of the missingness patterns observed in the data. In the observed data, all false positives were committed by examiners with a 0% nonresponse rate. Our simulation study included examiners with false positives with nonzero item nonresponse. However, the released data still do not allow an explicit calculation of the item nonresponse for false positive rates (or false negative rates). To allow others to assess the potential impact on nonresponse, nonresponse rates must be explicitly reported (or sufficient information about the data and study design must be reported so these values can be calculated). To adjust for nonresponse, researchers need, at minimum, the assigned items for each examiner, the examiner’s response (or lack of response) to each assigned item, a way to link responses to unique items across examiners, and a way to link examiner demographics to item responses (for more details on the information needed to adjust for missingness, see Khan and Carriquiry 2023). We note that the nonresponse rate in the released data was either 50% or 0% for each examiner. In such a situation, it is critical to examine the demographic differences between low responders and complete responders and the characteristics of items with no response.

5. Discussion

Two major issues currently affect all black-box studies: self-selected participants and large proportions of missingness that go unaccounted for in the statistical analyses of examiner responses. We are the first to explore either of these issues in black-box studies. Using real-world court materials, we have shown that black-box studies are likely relying on unrepresentative samples of examiners. Similarly, we have used actual black-box data to show the missingness in forensic black-box studies is likely non-ignorable. Current estimates of error rates could be significantly biased, and we show there is evidence this bias works to underestimate error rates.

There are ways to overcome both of these problems. The nonresponse rates are easier to address. There is a rich literature on methods to identify missing data mechanisms, adjust statistical analyses to minimize nonresponse bias, and properly account for the associated estimation uncertainty. It would be relatively simple to produce less biased, more reliable estimates of error rates in black-box studies, even without collecting additional data. To do this, however, authors of black-box studies must share enough de-identified information about the participants and the experimental design to enable independent researchers to conduct their own explorations. This information must include the items assigned to each examiner for each analysis and the associated responses (including a nonresponse). When demographic information is available, this information must be linkable to the response data. Similarly, when multiple examiners evaluate the same comparison, the data must allow their responses must also be linkable. These practices are standard in other scientific disciplines but rare in forensic science.

More difficult to address is the question of the representativeness of black-box study participants. As a threshold matter, participants should not be allowed to self-select. Most black-box
studies currently use members of professional organizations, and there is no reason that random samples cannot be taken from these lists to invite examiners to participate in black-box studies. While this would not ensure representative samples, it would at least give more insight into unit nonresponse from a broader population. Beyond this threshold issue, too little is known about individuals who examine forensic evidence. A huge challenge is that almost anyone can be admitted as an expert by a judge. Even identifying members of the relevant population is an elusive problem. The lack of accessible data collected by courts at all levels adds to the challenge. In the short term, this means that further research about the population of people examining forensic evidence is required. As courts continue to transition to electronic court filings, there will be more opportunities to explore court records. For example, one plausible study to identify examiners could involve a multi-stage sampling design: (a) First, draw a random sample of courts that is itself representative; (b) In each, enumerate criminal cases filed in the last year; (c) Identify experts who testified in these (or a random subset of these) cases and request curriculum vitae from the respective representation. We have completed a pilot study in a single-state jurisdiction similar to this setup and shown that it is possible (albeit difficult) to identify testifying experts in this way. At the moment, obtaining the curriculum vitae of experts who were not subject to an objection relies on the cooperation of state and defense attorneys.

In this article, we have focused on how black-box studies are currently being used in courts. In courts, judges are told that these studies can be considered representative of a broader population of studies, and this has shaped our emphasis on representative samples. However, we acknowledge that judges (and juries) may be interested in simply assessing how much weight to give an examiner’s testimony rather than determining its admissibility. In other words, they may want to know which type of training or experience can help to improve an examiner’s accuracy. In this case, estimating an error rate for all examiners in a given discipline is not the inferential goal. However, from a practical standpoint, research into the population of people examining forensic evidence will still be necessary to begin to understand the factors that may influence an examiner’s credibility.

Funding
This work was partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreements 70NANB15H176 and 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College and University of Nebraska, Lincoln.

References
Bajic, S., Scott Chumbley, L., Morris, M., and Zamzow, D. (2020), “Validation Study of the Accuracy, Repeatability, and Reproducibility of Forensic Comparisons,” Technical Report, Ames Lab., Ames, IA (United States). [8,9]
M., et al. (2022), “Accuracy and Reliability of Forensic Handwriting Comparisons,” Proceedings of the National Academy of Sciences, 119, e2119944119. [3,5]

Hofmann, H., Vanderplas, S., and Carriquiry, A. (2021), “Treatment of Inconclusives in the AFTE Range of Conclusions,” Law, Probability and Risk, 19, 317–364. DOI:10.1093/lpr/mgb002. [2]

Jakobsen, J. C., Glud, C., Weterslev, J., and Winkel, P. (2017), “When and How Should Multiple Imputation Be Used for Handling Missing Data in Randomised Clinical Trials—A Practical Guide with Flowcharts,” BMC Medical Research Methodology, 17, 1–10. [4,5]

Jordan, S., Watkins, A., Storey, M., Allen, S. J., Brooks, C. J., Garaiova, L., Heaven, M. L., Jones, R., Plummer, S. F., Russell, I. T., et al. (2013), “Volunteer Bias in Recruitment, Retention, and Blood Sample Donation in a Randomised Controlled Trial Involving Mothers and their Children at Six Months and Two Years: A Longitudinal Analysis,” PLoS One, 8, e67912. [3]

Kerkhoff, W., Stol, R. D., Mattijssen, E. J. A. T., Berger, C. E. H., Didden, F. W., and Kerkhof, J. H. (2018), “A Part-Declared Blind Testing Program in Firearms Examination,” Science & Justice, 58, 258–263. [4]

Khan, K., and Carriquiry, A. (2023), “Hierarchical Bayesian Non-response Models for Error Rates in Forensic Black Box Studies,” Philosophical Transactions of the Royal Society A, 381, 20220157. [9]

Kim, J. K., and Shao, J. (2014), Statistical Methods for Handling Incomplete Data, Boca Raton, FL: Chapman and Hall/CRC. [4]

Klees Expert Trial Transcript. (2002), Proceedings Transcript of Daubert Hearing Commencing on Wednesday, April 3, 2002 in United States v. Minerd, 2002 WL 32995663 (W.D.Pa.). [2]

Kumho Tire Co. v. Carmichael. (1999), 526 U.S. 137. [1]

Levy, P. S., and Lemeshow, S. (2013), Sampling of Populations: Methods and Applications, Hoboken, NJ: Wiley. [3]

Little, R. J. A., and Rubin, D. B. (2019), Statistical Analysis with Missing Data (Vol. 793), Hoboken, NJ: Wiley. [4]

Madley-Dowd, P., Hughes, R., Tilling, K., and Heron, J. (2019), “The Proportion of Missing Data Should not be Used to Guide Decisions on Multiple Imputation,” Journal of clinical epidemiology, 110, 63–73. [4]

Mislevy, R. J., and Wu, P.-K. (1996), “Missing Responses and IRT Ability Estimation: Omits, Choice, Time Limits, and Adaptive Testing,” ETS Research Report Series, 1996, i–36. [5,6]

Monson, K. L., Smith, E. D., and Bajic, S. J. (2022), “Planning, Design and Logistics of a Decision Analysis Study: The FBI/Ames Study Involving Forensic Firearms Examiners,” Forensic Science International: Synergy, 4, 100221. [2,3,7]

Monson, K. L., Smith, E. D., and Peters, E. M. (2023), “Accuracy of Comparison Decisions by Forensic Firearms Examiners,” Journal of Forensic Sciences, 68, 86–100. [8,9]

National Research Council (U.S.), editor. (2009), Strengthening Forensic Science in the United States: A Path Forward, Washington, DC: National Academies Press. [1]

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., and Petersen, I. (2017), “Missing Data and Multiple Imputation in Clinical Epidemiological Research,” Clinical Epidemiology, 9, 157–166. [5]

Pohl, S., Gräfe, L., and Rose, N. (2014), “Dealing with Omitted and Not-reached Items in Competence Tests: Evaluating Approaches Accounting for Missing Responses in Item Response Theory Models,” Educational and Psychological Measurement, 74, 423–452. [6]

President's Council of Advisors on Science and Technology. (2016), “Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature Comparison Methods,” Technical Report. Available at https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/peast_forensic_science_report_final.pdf. [1,2,3]

Richetelli, N., Hammer, L., and Spier, J. A. (2020), “Forensic Footwear Reliability: Part III—Positive Predictive Value, Error Rates, and Inter-Rater Reliability,” Journal of Forensic Sciences, 65, 1883–1893. [3,5]

Riddles, M. K., Kim, J. K., and Im, J. (2016), “A Propensity-Score-Adjustment Method for Nonignorable Nonresponse,” Journal of Survey Statistics and Methodology, 4, 215–245. [5]

Rubin, D. B. (1976), “Inference and Missing Data,” Biometrika, 63, 581–592. [4]

Schafer, J. L. (1999), “Multiple Imputation: A Primer,” Statistical Methods in Medical Research, 8, 3–15. [4]

Schafer, J. L., and Graham, J. W. (2002), “Missing Data: Our View of the State of the Art,” Psychological methods, 7, 147–177. [5]

Smith, J. A. (2021), “Beretta Barrel Fired Bullet Validation Study,” Journal of Forensic Sciences, 66, 547–556. [3,5]

Smith, T. P., Andrew Smith, G., and Snipes, J. P. (2016), “A Validation Study of Bullet and Cartridge Case Comparisons Using Samples Representative of Actual Casework,” Journal of Forensic Sciences, 61, 939–946. [2,3]

Smith, T. M. F. (1983), “On the Validity of Inferences from Non-random Samples,” Journal of the Royal Statistical Society, Series A, 146, 394–403. [3]

State v. Moore. (1991), 122 NJ. 420. [3]

Strassberg, D. S., and Lowe, K. (1995), “Volunteer Bias in Sexuality Research,” Archives of Sexual Behavior, 24, 369–382. [3]

Taylor, A. M., Cahn-Weiner, D. A., and Garcia, P. A. (2009), “Examination of Volunteer Bias in Research Involving Patients Diagnosed with Psychogenic Nonepileptic Seizures,” Epilepsy & Behavior, 15, 524–526. [3]

Ulery, B. T., Austin Hicklin, R., Buscaglia, J., and Roberts, M. A. (2011), “Accuracy and Reliability of Forensic Latent Fingerprint Decisions,” Proceedings of the National Academy of Sciences, 108, 7733–7738. [2,3,5]

United States v. Cloud. (2021), No. 1:19-CR-02032-SMJ-1, 2021 WL 7184484 (E.D. Wash. Dec. 17, 2021). [1,2]

United States v. Shipp. (2019), 422 E.Supp.3d 762. [1,2,7]

Wasserstein, R. L., and Lazar, N. A. (2016), “The ASA Statement on p-values: Context, Process, and Purpose,” The American Statistician, 70, 129–133. [7]

Weller, T. J., and Morris, M. D. (2020), “Commentary on: I. Dror, N. Scurich “(Mis) Use of Scientific Measurements in Forensic Science” Forensic Science International: Synergy 2020,” Forensic Science International: Synergy, 2, 701. DOI:10.1016/j.fsisy.2020.08.006. [3]

Yee, T. W. (2010), “The VGAM Package for Categorical Data Analysis,” Journal of Statistical Software, 32, 1–34. [8]

Yee, T. W., and Wild, C. J. (1996), “Vector Generalized Additive Models,” R package version 1.1-8. Available at https://CRAN.R-project.org/package=VGAM. [8]