

# Treatment of Inconclusives in the AFTE Range of Conclusions \*

Heike Hofmann<sup>1, 2</sup>, Susan Vanderplas<sup>3</sup>, and Alicia Carriquiry<sup>1, 2</sup>

<sup>1</sup>Statistics Department, Iowa State University, 2438 Osborne Dr, Ames, IA 50011

<sup>2</sup>Center for Statistics and Applications in Forensic Evidence (CSAFE), Iowa State University, 613 Morrill Rd, Ames, IA 50011

<sup>3</sup>Statistics Department, University of Nebraska Lincoln, 340 Hardin Hall North Wing, Lincoln, NE 68583-0963

June 21, 2022

## Abstract

In the past decade, and in response to the recommendations set forth by the [National Research Council Committee on Identifying the Needs of the Forensic Sciences Community \[2009\]](#), scientists have conducted several black-box studies that attempt to estimate the error rates of firearm examiners. Most of these studies have resulted in vanishingly small error rates, and at least one of them [\[Baldwin et al., 2014\]](#) was cited by the President’s Council of Advisors in Science and Technology (PCAST) during the Obama administration, as an example of a well-designed experiment. What has received little attention, however, is the actual calculation of error rates and in particular, the effect of inconclusive findings on those error estimates. The treatment of inconclusives in the assessment of errors has far-reaching implications in the legal system. Here, we revisit several black-box studies in the area of firearms examination, investigating their treatment of inconclusive results. It is clear that there are stark differences in the rate of inconclusive results in regions with different norms for training and reporting conclusions. More surprisingly, the rate of inconclusive decisions for materials from different sources is notably higher than the rate of inconclusive decisions for same-source materials in some regions. To mitigate the effects of this difference we propose a unifying approach to the calculation of error rates that is directly applicable in forensic laboratories and in legal settings.

## 1 Introduction and Background

It has now been more than a decade since the publication of the [National Research Council Committee on Identifying the Needs of the Forensic Sciences Community \[2009\]](#) report, which discussed the need for better-designed studies to estimate error rates in forensic disciplines. In this paper, we examine several new studies that have been published since the 2009 report and assess the state of error rate studies in firearms and toolmark analysis. For each study, we calculate the error rates from the published study results using standardized methods. We also assess the impact of study design and treatment of inconclusives on the calculated error rates. To lay a foundation for the common calculation methods used in this paper, we introduce some theory and vocabulary before comparing the studies and assessing the state of error rates in firearms and toolmark analysis.

Examiners visually classify the similarity of toolmark and firearm evidence according to the AFTE theory of identification using a three-point scale: identification, inconclusive, or elimination [\[AFTE Criteria for](#)

---

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

Identification Committee, 1992, see also Table 1]. Exact guidelines for this classification vary from lab to lab; some labs will exclude only based on non-matching class characteristics, such as the direction of rifling, the number of lands and their width, or the type of rifling [Bunch and Murphy, 2003]. In other labs, CMS (consecutively matching striae) as defined by Biasotti [1959] is used as a measure to quantify the similarity of two lands. In virtually all labs, the assessment of individual characteristics of bullet markings is done by visual inspection.

1. Identification
Agreement of a combination of individual characteristics and all discernible class characteristics where the extent of agreement exceeds that which can occur in the comparison of toolmarks made by different tools and is consistent with the agreement demonstrated by toolmarks known to have been produced by the same tool.
2. Inconclusive
(a) Some agreement of individual characteristics and all discernible class characteristics, but insufficient for an identification.
(b) Agreement of all discernible class characteristics without agreement or disagreement of individual characteristics due to an absence, insufficiency, or lack of reproducibility.
(c) Agreement of all discernible class characteristics and disagreement of individual characteristics, but insufficient for an elimination.
3. Elimination
Significant disagreement of discernible class characteristics and/or individual characteristics.
4. Unsuitable
Unsuitable for examination.

Table 1: AFTE Rules of Toolmark Identifications [AFTE Criteria for Identification Committee, 1992].

The identification process – i.e. the assessment of whether two samples come from the same source (were made by the same tool, the same shoe, the same finger, shot through the same barrel) or from different sources – is quite complex. For firearms and toolmark evidence to be admissible in court, it must be possible to explicitly characterize the accuracy of the examination process [Giannelli, 1993].

The need for scientific validation and experimentally-determined error rates has been identified in several reports evaluating the discipline of forensic science in the United States as critical for addressing problems within the field and improving the overall performance of the justice system. According to the National Research Council Committee on Identifying the Needs of the Forensic Sciences Community [2009]:

much forensic evidence - including, for example, bite marks and firearm and toolmark identifications - is introduced in criminal trials without any meaningful scientific validation, determination of error rates, or reliability testing to explain the limits of the discipline.

The President’s Council of Advisors on Science and Technology (PCAST) [2016] identified two important gaps:

(1) the need for clarity on the scientific meaning of “reliable principles and methods” and “scientific validity” in the context of certain forensic disciplines, and (2) the need to evaluate specific forensic methods to determine whether they have been scientifically established to be valid and reliable.

When jurors are left to form their own conclusions about the error rates of forensic pattern disciplines, they often come up with estimates which are far lower than empirical studies, suggesting that jurors consider forensic evidence (in the absence of error rates determined from scientific studies) as of more determinative value than is warranted by the evidence. The PCAST Report summarizes the effect:

In an online experiment, researchers asked mock jurors to estimate the frequency that a qualified, experienced forensic scientist would mistakenly conclude that two samples of specified types came

from the same person when they came from two different people. The mock jurors believed such errors are likely to occur about 1 in 5.5 million for fingerprint analysis comparison; 1 in 1 million for bitemark comparison; 1 in 1 million for hair comparison; and 1 in 100 thousand for handwriting comparison.

Studies in the form of designed experiments serve as the basis for assessing the accuracy of the identification process. Knowing ground truth means that the experimenter knows whether two samples come from the same source or from different sources. Casework can not be used to assess the accuracy of the identification process, because ground truth is not knowable. However, in scientific studies of forensic examination, ground truth is available to the experimenter because the test and the “evidence” were explicitly designed.

In designed experiments, any physical factors that might affect the outcome of a subsequent assessment of pattern similarity can be controlled or systematically varied [Spiegelman and Tobin, 2013]. In the specific situation of firearms evidence, the type of firearm and the ammunition used are of particular interest.

In addition to the physical evidence-generation process, experimenters must carefully control the information participants have about the study design and structure. We provide an example of the effect of extraneous information on estimated error rates in Section 3. Studies that attempt to gain insight into the identification process must navigate these complexities, ideally without becoming too complicated themselves.

In the remainder of this section, we introduce terms commonly used in experiments with a specific focus on the forensic identification process:

**Reference and Questioned Samples** Most studies are set up to mirror casework, i.e. there is a set of samples of known origin, used for reference, and a set of questioned samples of unknown origin. The main objective of a study is to determine for a pair of known and questioned samples if they share the same source or come from different sources.

**Closed vs. Open Set** In a closed set study all questioned samples originate from the same source as one (set of) known sample(s). Conversely, an open set study is one in which questioned samples may originate from sources outside of the known samples provided. Similarly, not all known samples might have a match among the questioned samples.

**White Box vs. Black Box** In a white-box study the experimenter attempts to understand *why* an examiner came to a specific decision, in contrast to a black box study, which only evaluates the correctness of the decision without assessing the reasoning behind it.

**Blind testing** A blind (or blinded) study is one in which the participant (in this case the examiner) does not know that they are being tested<sup>i</sup>; that is, a study which appears to be part of a case, rather than research. Blind testing is often recommended [President’s Council of Advisors on Science and Technology (PCAST), 2016, Spiegelman and Tobin, 2013, Koehler, 2013], because the error rates from blind testing better generalize to casework: many studies in a variety of disciplines have shown that people behave differently when they know they are being tested.

**Number of knowns from different sources** Some studies [Keisler et al., 2018] provide only one known exemplar (or a set of multiple exemplars from the same source). Other studies, such as [Brundage,

---

<sup>i</sup>There is some variability in what the term “double-blind” refers to in the context of forensic studies. The use in Bunch and Murphy [2003], Stroman [2014], Duez et al. [2018] does not match the use in President’s Council of Advisors on Science and Technology (PCAST) [2016], Spiegelman and Tobin [2013], Koehler [2013]. At least some of the confusion stems from the use of the term in medical contexts, where “double-blind” means that neither the patient nor the doctor knows whether the treatment received is the control or the treatment under investigation. In both uses of “double-blind”, the underlying goal is to remove any biases which may cause the evaluation of the available evidence differently. In drug trials, the doctor judges whether the patient has improved or not, and thus, the doctor cannot know which treatment the patient has received, because the knowledge may affect his decision subconsciously. In forensic studies, the examiner assesses the provided evidence and comes to a decision; thus, in a double-blind forensic study, the examiner cannot know that the evidence is part of a designed study rather than casework, because knowing that extra information may subconsciously affect the examiner’s assessment. Kerkhoff et al. [2018] provide a more thorough discussion of the different ways the term “blind testing” has been used in forensic error rate studies.

1998, Hamby et al., 2009, 2019] and the Houston FSC and Phoenix studies from Vanderplas et al. [2020], include multiple different sources as knowns.

**Study length** Most crime labs are understaffed and maintain a fairly large backlog of cases; as a consequence, examiner time is limited. While examiners should participate in these studies, they must balance the competing demands of a large and consequential workload and the benefit to FTE community. Studies that require examiners to make a large number of comparisons may be less likely to find a sufficient number of participants to generate an acceptable sample size.

Many of these considerations can be boiled down to limiting the examiner’s knowledge about the study as much as possible, so that (ideally) the only information provided is the information in each pairwise comparison. Blind studies, for instance, remove the knowledge that evidence is from a designed test and not casework. Open set studies, which are preferable to closed set studies, remove the knowledge an examiner might have about whether a comparison is guaranteed to match one of the provided knowns. Studies with designs that limit the number of comparisons by allowing only a single known source and a single unknown source prevent the examiner from using any sort of deductive reasoning. However, open set studies provide only limited additional information even in the case where multiple unknown sources are compared to a single known source.

Once a study is designed and test samples have been assessed by forensic examiners, error rates can be calculated. In the remainder of this paper, we discuss the experimentally determined error rates reported in studies in firearms and toolmark examination. We examine the variability in the methods for calculating and reporting error rates, and the different meanings and utility of different types of error rates. Using a set of different accuracy and error rates, we assess the state of firearms and toolmark analysis, and the implications of currently available data on the legal assessment of the reliability of testimony about firearms and toolmark related evidence.

## 2 Calculating Error Rates

Before approaching the actual studies, we define a framework and some basic notation to facilitate a comparison of studies with different experimental designs and explore the logic behind the numerical calculations.

### 2.1 Classification Framework

When firearms and toolmark examiners assess a set of evidence (generally, a pair of items, one of known provenance and one of unknown provenance) examiners decide whether the evidence is an identification, inconclusive, or an elimination. Figure 1 is a sketch showing the relationship between evidence (dots) and examiners’ decisions (areas). The color for both evidence and decisions corresponds to the source of evidence (different/same) and the type of decision made, respectively. The different combinations of colored dots on top of colored regions correspond to the six potential outcomes of evidence assessment and examiner decision. Obviously, not all of these outcomes are correct.

Deciding whether evidence is an identification, exclusion, or inconclusive provides a *classification* of the evidence, allowing us to make use of elements of the larger *classification framework*. We will introduce the classification framework and the respective error rates using as an example a generic experiment: let us assume that a representative sample of independent forensic toolmark examiners were asked to complete a total of  $N$  comparisons, consisting of  $S$  same-source comparisons and  $D$  different source comparisons. An aggregation of the examiners’ evaluations can then be reported as shown in Table 2. Note that the layout of the table is nothing but a summary of all the possible combinations of examiners’ decisions and the actual state of the evidence as sketched out in Figure 1. The pieces of Figure 1 are shown along with letters  $a$  to  $f$  in Table 2. These letters will be used throughout this section to calculate various accuracy rates and probabilities.

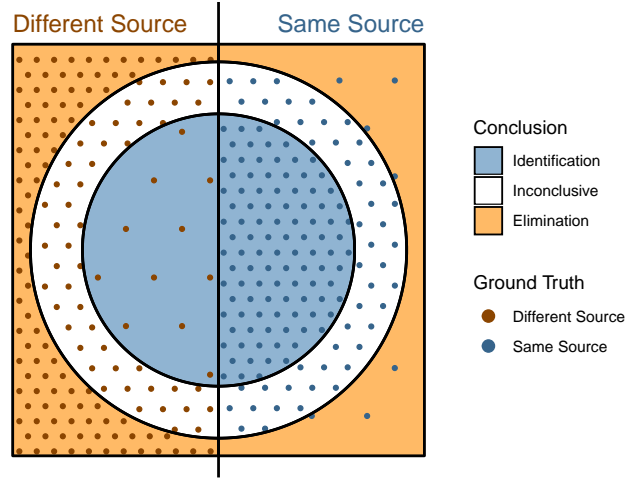








Figure 1: Sketch of the relationship between ground truth of evidence (dots) and examiners' decisions (shaded areas). In a perfect scenario dots only appear on the shaded area of the same color. Any dots on differently colored backgrounds indicate an error in the examination process.

Table 2: An example results table for a generic experiment where comparisons are pairs known to be either from the same source or from different sources, and examiners classify each comparison as an identification, an inconclusive, or an elimination, as specified in the AFTE rules of identification. Let  $S$  be the total number of same source comparisons, then  $S$  is the sum of  $a$ ,  $b$ , and  $c$ . Similarly,  $D$ , the total number of different source comparisons, can be written as  $D = d + e + f$ . The sum of  $S$  and  $D$  is the total number of comparisons,  $N$ .

Experiment Count Data				
	Identification	Inconclusive	Elimination	Source Total
Same Source	$a$ 	$b$ 	$c$ 	$S = a + b + c$
Different source	$d$ 	$e$ 	$f$ 	$D = d + e + f$
Conclusion Total	$a + d$	$b + e$	$c + f$	$N = S + D$

The *joint probability* of any source condition and any examiner conclusion can be found by taking the corresponding cell in Table 2 and dividing it by  $N$ , the total number of comparisons in the study.<sup>ii</sup> Thus, the joint probability of a same-source comparison and examiner identification is  $a/N$ , while the joint probability of a different-source comparison and examiner identification would be  $d/N$ . The accuracy (or error rate) of a classification method can then be determined by assessing how often the process produces a correct (or incorrect) result.

## 2.2 What makes an error?

In a classification problem, we would normally take the results in Table 2 and calculate various measures of accuracy and classification error. In firearms and toolmark examination, however, this is complicated by a mismatch between the physical source of the evidence and the set of examiners' decisions. There are two possibilities to describe the physical state of the evidence: two pieces of evidence are either from the same source or from different sources, but there are three primary outcomes from an examiner's decision. The examiner can make an identification, an elimination, or an inconclusive determination.<sup>iii</sup> The difference in the number of possible source categories and resulting decisions raises the question about how to deal with inconclusive results when calculating error rates. Under AFTE guidelines, an inconclusive result is an acceptable outcome of a comparison and therefore can not be considered as an error made by the examiner. It has been argued, however, that inconclusive decisions are systematic errors that occur during the evaluation process, because the final decision does not match the known information [Koehler, 2007]. While these two statements seem to be contradictory, we provide a foundation under which the two approaches can be partially reconciled in a way that provides additional insight into the examination process.

Dror and Langenburg [2018] suggest treating inconclusive results as equivalent to a "decision with certainty that the quantity and quality of information are not sufficient to draw any conclusion regarding the source". Under this framework, an examiner's assessment starts with inconclusive and is refined to identification or elimination given sufficient evidence in either direction. Statistically, this would suggest that the decision about the amount of evidence available (inconclusive or not) would be independent of the source of the evidence (same or different source). That is, inconclusive decisions should be equally likely in same-source and different-source assessments.

There are three main ways that inconclusive decisions can be treated in calculating error rates: inconclusive decisions can be (1) excluded from error calculations, (2) included as correct results [Duez et al., 2018], or (3) included as incorrect results [Chumbley et al., 2010]; each of these decisions has an impact on the actual value of the error rate as well as the interpretation of the resulting error rates. Table 3 shows an overview of these three approaches to calculating the error rate based on the general structure of Table 2, with a fourth option which will be addressed in more detail in Section 4.

Each of the approaches to calculating the error rate has a different meaning and interpretation; as a result, it is important to consider which error rate best suits the purpose of a study and its interpretation in the larger framework of pattern evidence. Error calculations in (1) arise in two scenarios: in Chumbley et al. [2010] examiners were asked to make either an identification or an elimination, and were not given the option of making an inconclusive decision. This results in cell values  $b$  and  $e$  of zero, i.e. the calculation of the overall error rate simplifies to a binary, symmetric decision process consisting of two actual states and two possible decisions. In this case, the overall error rate is the sum of the false positives and the false negatives, divided by the number of overall comparisons.

<sup>ii</sup>The joint probability of events A and B is the probability that both A and B occur. Here, the joint probability of an examiner identification and a same-source comparison would be equivalent to  $a/N$ .

<sup>iii</sup>The fourth category, unsuitable for examination, should be used when evaluating single pieces of recovered evidence; it does not result from the comparison of an unknown source to a known source.

Table 3: Different ways to calculate the overall error rate. An arrangement of two rows of three boxes resembles the interior cells of Table 2. We use the following convention:

<span style="display: inline-block; width: 15px; height: 15px; background-color: white; border: 1px solid black;"></span>	correct decisions, cell values are included in the denominator
<span style="display: inline-block; width: 15px; height: 15px; background-color: lightgray; border: 1px solid black;"></span>	cell value is zero or excluded from the calculation
<span style="display: inline-block; width: 15px; height: 15px; background-color: red; border: 1px solid black;"></span>	incorrect decision, cell values are included in the numerator and denominator of the error ratio.

#	Type	Illustration	$P(Error) =$						
1	No Inconclusives or Incl. Ignored	<table><tr><td>a</td><td>b</td><td>c</td></tr><tr><td>d</td><td>e</td><td>f</td></tr></table>	a	b	c	d	e	f	$\frac{P(DS \ \& \ Ident.) + P(SS \ \& \ Elim.)}{P(Ident.) + P(Elim.)} = \frac{c + d}{a + c + d + f}$
a	b	c							
d	e	f							
2	Inconclusives as Correct	<table><tr><td>a</td><td>b</td><td>c</td></tr><tr><td>d</td><td>e</td><td>f</td></tr></table>	a	b	c	d	e	f	$P(DS \ \& \ Ident.) + P(SS \ \& \ Elim.) = \frac{c + d}{N}$
a	b	c							
d	e	f							
3	Inconclusives as Incorrect	<table><tr><td>a</td><td>b</td><td>c</td></tr><tr><td>d</td><td>e</td><td>f</td></tr></table>	a	b	c	d	e	f	$\begin{aligned} P(DS \ \& \ Ident.) + P(SS \ \& \ Elim.) + P(Inconclusives) &= \\ &= \frac{b + c + d + e}{N} \end{aligned}$
a	b	c							
d	e	f							
4	Inconclusives as Eliminations	<table><tr><td>a</td><td>b</td><td>c</td></tr><tr><td>d</td><td>e</td><td>f</td></tr></table>	a	b	c	d	e	f	$\begin{aligned} P(DS \ \& \ Ident.) + P(SS \ \& \ Elim.) + P(SS \ \& \ Inconcl.) &= \\ &= \frac{b + c + d}{N} \end{aligned}$
a	b	c							
d	e	f							

Under Scenario 1, inconclusive decisions are treated as a valid decision category but restricted from the calculation of the error rate. This solution entirely ignores inconclusive decisions. However, in practice, inconclusive results are a substantial part of the identification process and are included in examiner testimony (often phrased as “could not be excluded”). Thus, this elimination from consideration is not appropriate if the goal of a study is to assess the error rate of the entire evaluation process. Koehler [2013] recommends using the approach of option 1 for calculating error rates while also tracking rates of inconclusive identifications “for other purposes”.

Under AFTE’s Theory of Identification, inconclusive results are acceptable outcomes and not considered errors. This approach is captured in Option 2: inconclusive decisions are treated as correct regardless of the actual state. What this means in real terms, is that inconclusive decisions are counted as identifications if they are actually from the same source, but counted as eliminations if they are actually from a different source. Note that this approach is exploitable in a strange way: to give a very extreme example, an examiner could report inconclusive decisions for every evaluation over the rest of their career and never make an error. Dror [2020] point out this discrepancy in many forensic disciplines and describe the effects of this policy on an individual assessment level as well as a system level. However, when the goal is to assess the error rate of the examiner *under the prevailing guidelines of the AFTE Theory of Identification*, Option 2 may be a reasonable option for assessing error rates of individual firearm and toolmark examiners.

Option 3 provides an approach to view inconclusive results and the corresponding error rates within the wider framework of the legal system: if a firearm does not mark well, a firearms examiner might not be able to make an identification or an elimination for reasons having nothing to do with the examiner’s skill. From the perspective of the *identification process*, though, an inconclusive result does not result in a decision of identification or elimination and can therefore not be considered a successful assessment: two pieces of evidence are either from the same source or originate from different sources and an inability to distinguish between those options is an error.

Practically, Option 3 reflects the error of the *examination process* rather than the examiner as an individual. In Option 4, which will become more relevant in Section 4, inconclusive results are treated the same as eliminations. This option places the primary focus of the examiner’s assessment on identification, and treats any inability to make an identification as equivalent to an elimination.



Table 4: Source-specific probabilities, calculated using the quantities introduced in Table 2. Cells in the main body of the table show conditional probabilities of conclusion  $Y$  (one of Identification, Inconclusive, or Elimination) given known source  $X$  (same, different). In the final column of the table, the total number of comparisons is shown - this number forms the denominator for each cell when calculating source-specific probabilities. These marginal total comparison numbers are determined by the experimental design.

Source-specific Probabilities				
	Identification	Inconclusive	Elimination	Total # Comparisons
Same source	$a/S$	$b/S$	$c/S$	$S$
Different source	$d/D$	$e/D$	$f/D$	$D$

The errors introduced in this section summarize the *overall* error rate of a study. While these errors are informative, most studies provide a deeper insight into different aspects of the examination process by taking additional information into account. In the next two sections we introduce two additional types of error rates: *source-specific* and *decision-specific* error rates. Both of these approaches are additional assessments of the reliability of a classification method that leverage *conditional probabilities*, i.e. probabilities which are updated to account for known (or assumed) information. Rather than summarizing an experiment in a single number reflecting error or accuracy like before, these conditional probabilities allow us to provide a more specific assessment of the error accounting for known information, such as the examiner’s decision or the source of the evidence. This allows us in particular to calculate separate accuracy rates for comparisons of same-source or different-source evidence. As discussed in Dror [2020], additional statistics that provide context to error rates can be extremely beneficial in court. There is an additional advantage: some of these conditional probabilities do not require inconclusive results to be explicitly handled as errors or correct decisions.

### 2.3 Source-specific assessment

The most common way to assess the success (or accuracy) of a classification method is to calculate its *sensitivity* and *specificity*. Sensitivity, also called the true positive rate (TPR), is an examiner’s ability to make an identification when examining same-source evidence (SS). Sensitivity can be expressed as the conditional probability  $P(\text{Identification} | SS)$ . Conversely, specificity, or the true negative rate (TNR), is an examiner’s ability to make an elimination given evidence from different sources (DS),  $P(\text{Elimination} | DS)$ . We calculate the true negative rate as the conditional probability of the examiner making an elimination *given* the pair under examination is from different sources. A conditional probability of an event *given* another event is the probability that the two events both occur, divided by the probability of the second event, specifically for the true negative rate:

$$TNR = P(\text{Elimination} | DS) = \frac{P(\text{Elimination and } DS)}{P(DS)}.$$

Complementary to accuracy rates, we define two error rates of a classification method: The *false positive rate* (FPR) – or failed elimination rate – is the probability that an examination of different source evidence results in an error. Similarly, the *false negative rate* (FNR) – or missed identification rate – is the probability that an examination of same-source evidence results in an error.

As discussed in Section 2.2, there are multiple ways to define an error, but the concepts of FPR and FNR hold for all of the options of treating inconclusive decisions outlined in Table 3.

In all of these assessments, the rates are calculated by conditioning on the *origin of the evidence*. This process is shown graphically in Figure 2 using a decision tree to describe the comparisons which can be made conditioned on ground truth. Calculations for each value are shown in Table 4.

A google sheets workbook which performs all of these calculations given counts  $a, b, c, d, e$ , and  $f$  is available at <http://bit.ly/FTE-error-rate-worksheet>.



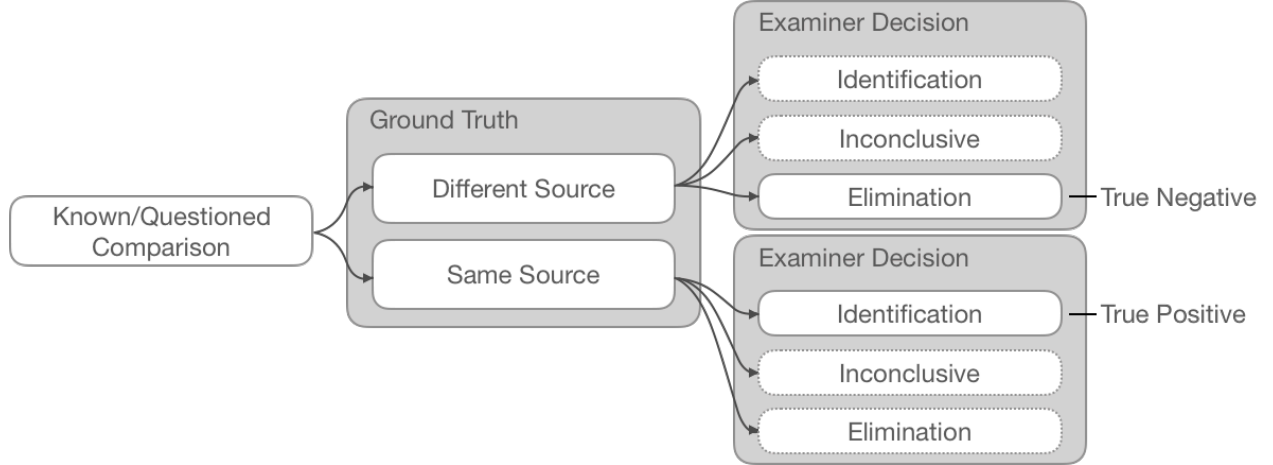


Figure 2: Tree Diagram illustrating the logic of source-specific error rates. Information flow happens from left to right: we are assessing an examiner’s decision given that someone (not the examiner) knows the source of the evidence. Source-specific rates give the probability of an examiner’s decision (Identification, Inconclusive, or Elimination) given same-source and different-source evidence.

In the context of Table 4, conditional probabilities can be interpreted as follows: given a same-source pair of evidence, the examiner makes an identification with probability  $a/S$ , the true positive rate (TPR). Conversely, for a different-source evidence pair, the examiner makes an elimination with probability  $f/D$ , the true negative rate (TNR).

## 2.4 Decision-specific assessment

An alternate way to evaluate a classification method is to assess the probability of a same-source or different-source comparison, given that the examiner has made a decision (or, more generically, a specific general conclusion). The *positive predictive value* PPV is the probability that a pair of evidence items have the same source, given that the examiner has made an identification, i.e.  $P(SS|Identification)$ . Its complement, the *false discovery rate* FDR, is the probability that the evidence is from different sources, given that the examiner has made an identification, .

Similarly, we can define predictive values when an examiner has made exclusions: the *negative predictive value* NPV is the probability that evidence is actually from different sources given that the examiner has made an exclusion:  $NPV = P(DS| Elimination)$ . Its complement, the *false omission rate* FOR is the probability that evidence is actually from the same source<sup>iv</sup> given an examiner has made an elimination. As before, we can map out the conditional hierarchy using a decision tree, shown in Figure 3. To reduce confusion with the naming of rates and predictive values we have labeled nodes in the trees of Figure 2 and Figure 3 corresponding to the rates defined in the text.

Using the quantities in Table 2, we can calculate the decision-specific or conclusion-specific probabilities from Figure 3 as shown in Table 5.

Predictive value assessments are used in assessing the implications of a particular classification label; they inform us about the *likely state of the physical evidence* given the expert’s decision or testimony. This makes the PPV, NPV, FDR, and FOR particularly useful when evaluating casework (or in a courtroom setting), because the true state of the evidence is unknown.

<sup>iv</sup>In Song et al. [2018], the FDR is referred to as the false identification error rate, and the FOR is called the false exclusion error rate.

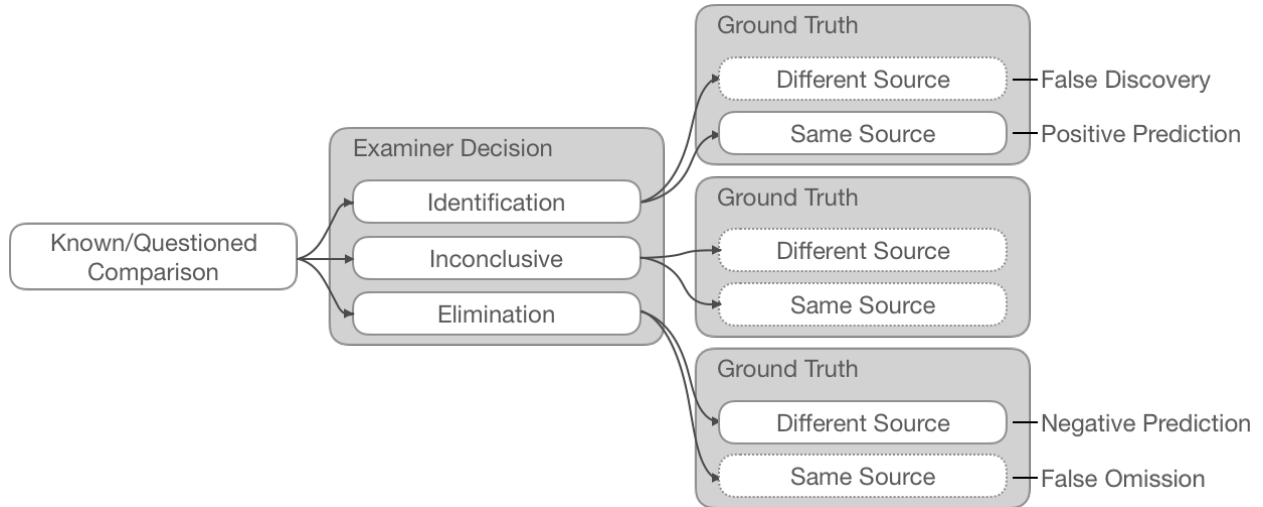


Figure 3: Tree Diagram illustrating the logic of positive and negative predictive value calculations. Decision-specific values give the probability of evidence coming from the same source or different sources given an examiner’s testimony.

Table 5: Decision-specific probabilities, calculated using the quantities introduced in Table 2. In each cell in the main body of the table, probabilities shown are the probability of source  $X$ , if the examiner’s conclusion (one of Identification, Inconclusive, or Elimination) is  $Y$ . In the final row of the table, the total number of comparisons is shown - this number forms the denominator for each cell when calculating decision-specific probabilities.

**Decision-Specific Probabilities**

	Identification	Inconclusive	Elimination
Same source	$a/(a + d)$	$b/(b + e)$	$c/(c + f)$
Different source	$d/(a + d)$	$e/(b + e)$	$f/(c + f)$
Total # Decisions	$a + d$	$b + e$	$c + f$

Note that we still cannot directly calculate the probabilities for any individual case or evidence: ground truth is still not knowable; however, the results of designed studies are used to give information tailored to a specific decision scenario. While it is reasonable to condition on unknown quantities in a mathematical sense, in practice it is much more informative to condition on quantities with known values: in this case, the examiner’s decision is the known information and the state of the physical evidence is unknown.

Conversely, source-specific assessments are useful when evaluating the performance of individuals on a designed test, because ground truth is known (and thus, can be conditioned upon) in designed studies.

## 2.5 The practical interpretation of error rates

The primary difference between the decision tree diagrams in Figures 2 and 3 is the variable on which we condition. The source-specific error rates, FPR and FNR, are calculated by conditioning on the source of the evidence (same-source/different source) and give the probability of an examiner’s decision. PPV and NPV are calculated by conditioning on an examiner’s conclusion and give the probability of evidence being same-source or different-source.

From a practical point of view, whether source-specific or decision-specific rates are of interest depends upon where we are in the overall (legal) decision process:

- As a firearms and toolmark examiner or as a lab director, we are interested in assessing accuracy based on ground truth, i.e. we are interested in source-specific rates to answer questions such as: given this evidence, are trained firearms examiner more likely than trainees to arrive at the correct conclusion? (Case studies later show that yes, that is the case.) Alternatively, lab directors might want to know if all of the examiners employed by the lab are performing above a certain threshold.
- As part of courtroom proceedings, we – the jury, lawyers, and judges – are interested in assessing the accuracy of the testimony offered by an expert, i.e. we are interested in conclusion-specific rates to answer questions of the form: given that an expert testified to the evidence as being an identification, what is the probability that the evidence comes from the same source? In this situation, we can update the overall error rate by conditioning on the examiner’s conclusion, providing more specific evidence about the scenario under scrutiny in court. As we only know the examiner’s testimony (and not ground truth), we must work with the decision-specific rates.

These two forms of questions look deceptively similar but lead to different mathematical formulations. This difficulty is not exclusive to evaluations of firearms evidence – it is a known and pervasive pitfall in other application areas, such as medical diagnostic testing [Casscells et al., 1978, Manrai et al., 2014]. The distinction between these two sources of error rates is perhaps a bit more approachable in the setting of a medical context:

- A given diagnostic test has an estimated sensitivity (true positive rate, TPR) and a specificity (true negative rate, TNR) that combined determine the overall accuracy rate. The accuracy of the test is established via extensive experimentation. These quantities are useful for an accrediting agency such as the FDA, or a company; both are interested in whether the test works as claimed.
- A patient who has received a negative test result is more interested in a different value: the probability that the patient has the disease even though the test was negative.

At the individual level, we want to use tailored information to guide our specific actions; at the system level, we are more interested in gross measures of performance.

The two conditional probabilities that are referred to in the diagnostic test example are the probability of a positive or negative test result given the disease is present or not:  $P(\text{Test result}|\text{Disease status})$  and the probability that somebody has the disease or not, given a specific test result:  $P(\text{Disease status}|\text{Test result})$ . Using Bayes’ rule, and assuming we also know the prevalence of the disease in the population, we can convert between the two probabilities. The parallel we draw here is the following: the examiner’s decision is

equivalent to a diagnostic test, and the origin of the evidence is equivalent to the disease status. In contrast to the medical example, however, we do not have information about the sensitivity and specificity of the field of firearms examination as a whole, let alone about a specific examiner. For that reason, we need designed studies to estimate the error rates of examinations in situations where ground truth is known. As we rely on these studies to generalize to casework, it is of great importance that these studies are well designed so that it is possible to estimate any of the success and error rates discussed above.

## 3 Studies

In this section, we examine the results from several studies designed to assess the error rates of firearms examination. For each study, we provide the reported results and present error rates and conditional probabilities discussed in the previous section along with a visual comparison for each study. Using these comparisons we assess the state of these error rate studies and provide an overview of the current status quo of firearms examination.

### 3.1 Error Rate Studies

The studies included in this section are not an exhaustive list of every error rate study ever performed; rather, we have included the most central and most cited studies, as well as several of the studies mentioned in the PCAST and NRC reports concerning firearm and toolmark analysis. A brief description of each study is provided here for reference purposes, with summary tables of results and conditional probabilities provided in Appendix A. A spreadsheet with the same study-specific error rates and calculations is available online at <http://bit.ly/fte-study-error-rates> for reference purposes.

**Baldwin** [Baldwin et al., 2014] is a study of Ruger SR9 cartridge cases. This is an open-set study. Each participant is asked to evaluate 15 test sets. Each test set consists of 3 reference cartridge cases and 1 questioned cartridge case. In ten of the 15 test sets the questioned cartridge case matched the references, in five tests the questioned cartridge case was from a different source.

**Keisler** [Keisler et al., 2018] is a study of cartridge case comparisons. It is an open set study. Participants were asked to evaluate 20 test sets. Each test set consisted of one reference and one questioned cartridge cases. In twelve of the 20 sets the questioned cartridge case matched the reference, in the other eight sets the questioned cartridge case came from a different source. A total of 126 examiners participated.

**Brundage-Hamby** [Hamby et al., 2019] is a closed-set study of 10 consecutively manufactured Ruger P-95 barrels. Sets consisting of 10 pairs of reference bullets and 15 questioned bullets were provided to 507 study participants over the course of several decades of data collection.

**Lyons** [Lyons, 2009] is a closed-set study of 10 consecutively manufactured Colt 1911A1 extractors. Participants received a set of 10 pairs of reference cartridge cases and 12 questioned cases. The design of this study is similar to [Hamby et al., 2019].

**Bunch** [Bunch and Murphy, 2003] is an open-set study of consecutively manufactured Glock breech faces. In this study, 8 participants from the FBI laboratory received 10 cartridge cases each and were required to evaluate all possible pairwise comparisons. The number of same-source and different-source cartridge cases varied by test kit, ranging from 10 same-source cartridges to 10 different-source cartridges.

**Fadul** [Fadul Jr. et al., 2012] is a closed-set study of the breech face striations/impressions produced by 10 consecutively manufactured slides. Participants received a set of 10 pairs of reference cartridge cases and 15 questioned cases. The design of this study is similar to [Hamby et al., 2019].

**Duez** [Duez et al., 2018] is an open-set study of breech face comparisons using virtual microscopy. Each of 56 participants (46 trained, 10 trainees, primarily from the US and Canada) were asked to evaluate the same two test sets consisting of three reference scans and four questioned scans. In one set, all

questioned scans were from the same source as the reference scans; in the second set, there were two different-source questioned scans and two same-source questioned scans.

**VCMER** [Chapnick et al., 2021] is an open-set study of breech face comparisons using virtual microscopy. Each of 76 participants (primarily from the US and Canada) received 16 test sets each consisting of two reference scans and one questioned scans. In 17 sets, the questioned bullet matched the references, and in 23 sets, the questioned bullet was fired from a different weapon.

**Mattijssen** [Mattijssen et al., 2020] is an open-set study of firing pin aperture mark images and scans from 200 Glock pistols. Sixty of the resulting pairwise comparisons were shown to 77 firearms examiners, who provided assessments based on a likelihood-ratio scale initially, and were then asked whether they would have returned an inconclusive decision after the fact. The results from the study were reported by the authors in both AFTE and LR formats, which allows us to include it in this assessment. A large majority (58) of the participants in this study report case work conclusions categorically (e.g. according to AFTE or similar scales).

**FAID-2009** [Pauw-Vugts et al., 2013] is an open-set study of bullets and cartridge cases using castings. Each of 64 participants evaluated 10 test sets (5 bullets, 5 cartridge cases). Each test set consisted of three reference castings and 1 questioned casting. In three of the bullet sets and two of the cartridge case sets, the questioned material matched the references. The examiners in this study are primarily from Europe, and the study used a 5-class categorical scale. Note that this study was part of a proficiency test, and as such, examiners may have faced different pressures than in the other error-rate studies.

Of the studies listed above, we have examined all but [Fadul Jr. et al. \[2012\]](#), which contained insufficient detail about examiner results and conclusions to calculate the quantities discussed in the previous section. Note that this paper is limited in scope to *error rate studies* conducted using firearms and toolmark examiners. There are several promising automatic analysis tools [see [Hare et al., 2017](#), [Vanderplas et al., 2020](#), [Chumbley et al., 2010](#), [Chu et al., 2013](#), [Song et al., 2018](#), [Tai and Eddy, 2018](#)] which calculate error rates based on 2d images and 3d scans of bullet and cartridge cases. These types of studies are not easily comparable to the black-box studies under consideration here: none of the automatic matching algorithms employs a category of ‘inconclusive’ as a result, so the discussions on how to deal with inconclusive results are not applicable. This implies that for algorithms we cannot distinguish between errors stemming from insufficient markings on evidence and errors inherent to the algorithm. We have included algorithmic results from [Mattijssen et al. \[2020\]](#), which included algorithmic analysis as well as examiner evaluations in the appendix for comparison purposes, but will defer additional examination of algorithmic studies to a future paper.

In the next sections, we discuss the design and results of each study, including calculation and comparison of source- and decision-specific error rates.

### 3.2 Consequences of Study Designs for Error Rate Estimation

Our survey of the most commonly cited studies reveals a list of experimental design concerns similar to those identified in the 2017 addendum to the PCAST report [[President’s Council of Advisors on Science and Technology \(PCAST\), 2017](#)].

As described in the PCAST report, “set-based” approaches can inflate examiners’ performance by allowing them to take advantage of internal dependencies in the data. The most extreme example is the “closed-set design”, in which the correct source of each questioned sample is always present; studies using the closed-set design have underestimated the false-positive and inconclusive rates by more than 100-fold. This striking discrepancy seriously undermines the validity of the results and underscores the need to test methods under appropriate conditions. Other set-based designs also involve internal dependencies that provide hints to examiners, although not to the same extent as closed-set designs.

The PCAST response, issued in 2017, identifies only one study [[Baldwin et al., 2014](#)] as appropriately

designed to evaluate the validity and reliability of firearms analysis methods. Shortly after the report was issued, additional studies were published which also meet the criteria for reliable studies in the report: [Keisler et al. \[2018\]](#), [Duez et al. \[2018\]](#), VCMER [[Chapnick et al., 2021](#)], and [Mattijssen et al. \[2020\]](#). All “good” studies used slightly different designs but share one essential element: each kit is made of multiple sets consisting of one or more reference samples (from the same source) and one sample of questioned origin, with samples from different sets having different origins.

These designs substantially reduce or eliminate the internal dependencies which provide “hints” to examiners by ensuring that each comparison of samples is considered independently. Also, the design of these studies ensures that it is possible to exactly enumerate the number of same-source and different-source comparisons.<sup>v</sup> All studies with reliable designs are also limited to cartridge case comparisons (one exception to that is FAID09 [[Pauw-Vugts et al., 2013](#)], which was published before the release of the PCAST report, but is a European-focused study). We could not identify any studies that assess the error rates of bullet or toolmark examination in a manner that would produce reliable error rate estimates.<sup>vi</sup>

As a result, studies that use less reliable designs, such as [Lyons \[2009\]](#), [Hamby et al. \[2019\]](#) (and the similarly designed [Fadul Jr. et al. \[2012\]](#)), and [Bunch and Murphy \[2003\]](#) have historically been frequently referenced in admissibility hearings. Of these, the study by [Bunch and Murphy](#) is perhaps statistically the most interesting design, but it has been superseded by studies with cleaner experimental designs, such as [Baldwin et al. \[2014\]](#) and [Keisler et al. \[2018\]](#). While [Bunch and Murphy \[2003\]](#) has internal dependencies due to the inclusion of multiple knowns, and as a result, the number of same-source and different-source comparisons cannot be fully determined, the study does have one desirable feature not found in even the well-designed studies which makes it worth discussing here. Specifically, [Bunch and Murphy](#) varies the composition of the test kits: no kit had the same composition in terms of same and different-source comparisons. This ensures that even if examiners discuss the tests, they cannot gain any additional information from such discussions.

The primary problem with [[Bunch and Murphy, 2003](#)] is that it includes multiple known exemplars. Consequently, it is possible to use logical reasoning to reduce the set of comparisons. For example, if an examiner is comparing unknown source A to known exemplars from sources 1 to 10, and A matches exemplars from source 2, it is not necessary to make comparisons to sources 3 - 10. This design ensures that it is not possible to count up the total number of different-source comparisons performed; as a result, we cannot compute the decision-specific error rates or the source-specific error rates for different-source comparisons. Any study which includes multiple known sources in a single comparison set will have this limitation. In [C.1](#), we use statistical simulation to explore the likely number of different-source comparisons performed in [[Bunch and Murphy, 2003](#)]; this assessment includes both the effect of deductive reasoning and the sampling procedure used to create the test kits. We also perform a similar assessment for [[Hamby et al., 2019](#)] in [C.2](#); the design of this study does not include the variability introduced by the test kit assembly procedure used in [Bunch and Murphy \[2003\]](#).

While studies that use multiple known sources have been conducted for more than 25 years, such studies are inherently flawed. These studies have an inherent bias, because they provide evidence that examiners can make identifications accurately (and do not make many false eliminations), but do not allow a quantification of the probability that examiners make eliminations correctly. As a result, when error rates derived from these studies are cited, they do not include errors that result from different source comparisons, which prevents evaluation of examiners on their ability to distinguish between different sources. This fundamentally biases these studies so that in court they provide useful (but misleading) information to the prosecution while offering nothing useful to the defense. That is, the underlying structure of these studies only provides information about how accurate examiners are when providing evidence against the defendant, producing a

---

<sup>v</sup>At least, it is possible with the additional assumption that a comparison between a multiple items from the same source and an unknown should count as a single comparison.

<sup>vi</sup>We have included FAID09 [[Pauw-Vugts et al., 2013](#)], which examines bullets and cartridge cases, but uses European examiners (who are trained to different standards and some of whom work in non-adversarial systems), and a 5-point categorical range of conclusions. For these reasons, this study is useful but is not necessarily a good representation of studies using the AFTE scale with US and Canadian examiners.



systematic bias when these rates are presented in court.

The design used in Hamby et al. [2019] was originally found in Brundage [1994], and dates back to 1994. Its longevity makes it the most extensive study in forensic error rates in firearms (and possibly across pattern disciplines). Unfortunately, the study is flawed in two major respects: it uses multiple known sources, and it is a closed set study. The multiple known source problem was described above, but this problem is magnified when combined with the closed-set study design.

In a closed-set design, the structure of the study helps examiners make the correct conclusion: if unknown source A is most similar to exemplars from source 8, then after the 10 comparisons are done, the examiner would rationally make an identification for source 8. In an open-set study, the examiner might rate the same pairwise comparison as inconclusive, because there is no additional information suggesting that the unknown must match one of the provided knowns. In both the multiple known and closed set study designs, the use of deductive reasoning artificially reduces the error rates calculated using the study results, as discussed in the PCAST report and addendum [President’s Council of Advisors on Science and Technology (PCAST), 2016, 2017].

It might be possible to modify the answer sheet so that examiners record each pairwise comparison, which would allow for assessment of the number of different-source comparisons performed. However, revising the answer sheet would not prevent the use of deductive reasoning, which gives examiners an advantage in error rate studies that does not exist in casework. Currently, studies patterned after Brundage [1994] and Hamby et al. [2019] use an answer sheet where examiners are only asked to report which of the 10 knowns the questioned bullet matches. Because the study does not ask examiners to report eliminations, there is no way to assess elimination error rates or the overall error rate consisting of both missed identifications and false eliminations.

In addition to the design issues which plague the closed-set studies, the execution of some of these studies makes the data even less reliable. Methodological issues, such as those described in Lyons [2009], where test sheets were returned to participants who misunderstood instructions, cast even more doubt on the utility of error rates derived from such studies and their ability to generalize to casework.

Finally, some design constraints are common to most studies. Of the studies assessed in this paper, only Bunch and Murphy [2003] uses a variable proportion of same and different source comparisons; without varying this percentage, examiners could compare notes and gain additional information about the study results. Many studies [but not all, Baldwin et al., 2014] pre-screened kit components to ensure that sufficient markings were present to allow for identification - this differs from casework and artificially inflates the rate of successful identifications and eliminations<sup>vii</sup>. Studies that tightly control the evidence quality and presence of individualizing marks would under-estimate the process error (though the estimates of examiner error are not necessarily affected). In almost all studies, participants use lab rules for determining whether eliminations could be made on individual characteristics; while Baldwin et al. [2014] instructed participants to use a uniform set of rules, many participants reported that they did not or could not adhere to these guidelines.

### 3.3 Assessment of Classification Error

Figure 4 shows the false positive and false negative rates for the six studies investigated in this paper. Both AFTE errors (which do not consider inconclusive decisions as errors and represent examiner errors) and process errors (which include inconclusive decisions and represent errors attributable to evidence quality and lab policy) are shown; examiner error rates are low for both missed elimination and missed identifications in most studies (though intervals reveal the effect of sample size, which was also noted in e.g. President’s

---

vii

In Baldwin [Baldwin et al., 2014], the only pre-screening of usability was to ensure cartridge cases were caught with the cartridge catching device. Additionally, for 3234 comparisons, FTEs evaluated how many of the known cartridge cases were usable for an evaluation: all three specimens were used in 3018 cases, two were used in 207 cases, and only one was used in nine cases. The Baldwin study is fairly unique in this regard: most studies make some effort to control the quality of the evidence sent out for assessment.



Council of Advisors on Science and Technology (PCAST) [2017]. Process errors, however, occur at higher rates and are particularly concerning for missed eliminations. There is consistently a much larger discrepancy between the two error rates for eliminations than for identifications; this phenomenon is examined in more detail in Section 4. Only a portion of the discrepancy between the examiner (AFTE) and process error rates for missed eliminations can be explained by lab policies that only allow eliminations based on mismatches in class characteristics.

The Bunch study, for instance, was conducted at the FBI, where all examiners followed the same lab policy allowing eliminations based on class characteristic mismatches but precluding eliminations based on differences in individual characteristics; as a result, the study has the largest difference between the two missed elimination rates.

Part of this difference can be attributed to lab policies: in some labs exclusions based on individual characteristics are not allowed (i.e. class characteristics match). However, the effect of lab policies does not explain all: Baldwin et al. [2014] made efforts to control for the effects of lab policy in determination of inconclusive findings, the efforts were not completely successful. This is despite instructions given to participants: “A very important aspect of this work that needs to be clearly understood is that the study specifically asked participants not to use their laboratory or agency peer review process... Some [participants] indicated that the design of our study with all cartridges fired from the same model of firearm using the same type of ammunition would prohibit the use of a finding of elimination, while others used a mixture of inconclusive and elimination or did not use inconclusive at all to indicate a finding other than identification.”

In Duez et al. [2018], the authors report that “13% of examiners are not permitted to eliminate on individual characteristics (therefore, their conclusions of inconclusive are perfectly acceptable).” Interestingly, it seems that some examiners reported eliminations in contravention of lab rules. Policies that forbid elimination based on a mismatch between individual characteristics alone seem to directly contradict the AFTE Theory of Identification, which allows for eliminations based on “significant disagreement of discernible class characteristics and/or individual characteristics”. It is apparent that there is a distinct difference between the error rates and inconclusive rates in US/CA and EU-focused experiments, showing the clear impact of the confluence of training, legal system, reporting scale, and the type of error rate study (proficiency exam vs error rate assessment).

The rates reported in this section are calculated directly from information reported in the papers without adjustments accounting for the use of deductive reasoning; as such, they represent the best possible case for error rates in studies that include multiple knowns, such as Hamby et al. [2019] and Bunch and Murphy [2003]. More realistic estimates of error rates that account for the use of deductive reasoning can be found in Appendix C.

### 3.4 Source-specific error rates

The false-positive rate and false-negative rate reported in Figure 4 are the error rates typically used to characterize a classification method. These rates, however, are functions of a larger set of probabilities which are calculated conditional on the known source of the evidence. In Figure 5, we see the full set of source-specific conditional probabilities: the probability that an examiner will make an identification, elimination, or inconclusive determination given that the source of the evidence is different (top panel) or the same (bottom panel). Probabilities are shown with 95% Pearson-Clopper confidence intervals, which provide a visual indication of the estimate’s variability. It is apparent that examiners are extremely good at working with same-source evidence: there are relatively few inconclusive determinations, almost no missed identifications, and very few false eliminations. It is also apparent that examiners have much more difficulty when examining evidence that arises from different sources. As an example, examiners who participated in the Bunch study [Bunch and Murphy, 2003] had a significantly higher probability of reaching an inconclusive conclusion than an elimination when evidence items had a different source. Closed-set designs (used in Hamby et al. [2019] and Lyons [2009]) do not even allow for the estimation of different-source probabilities because it is not possible to estimate the number of different source comparisons which are made; this inherent bias has the

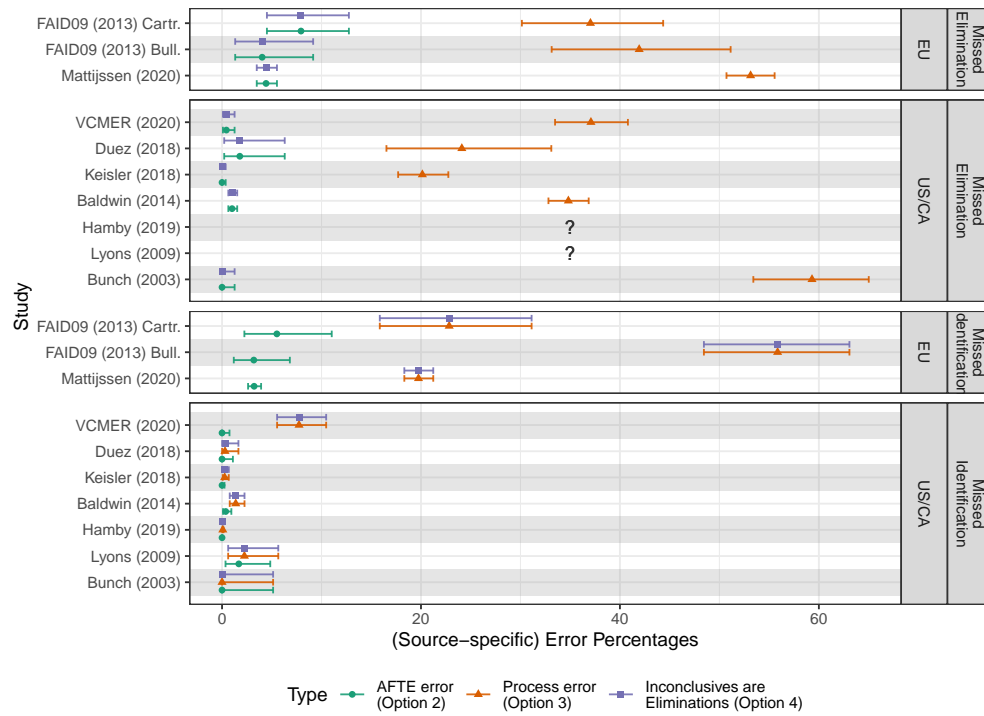


Figure 4: Percentages of missed eliminations and missed identifications by study. 95% Pearson-Clopper confidence intervals are drawn around the error estimates. Missed eliminations cannot be calculated for closed-set studies.

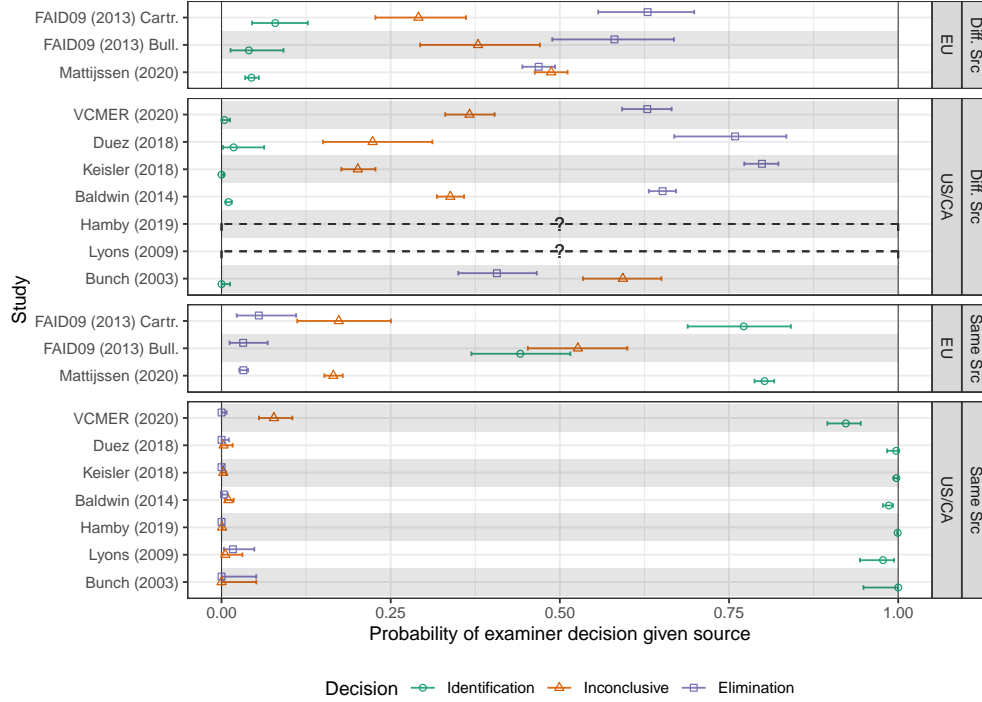


Figure 5: Pearson-Clopper 95% confidence intervals for the probability of an examiner’s conclusion given source of the evidence. No assessments can be made about different-source specific probabilities when the total number of comparisons cannot be determined. The probability of inconclusive determinations on different source evidence is quite high, much higher than for same source evidence, where the probability for an inconclusive result is close to zero. FAID09 [Pauw-Vugts et al., 2013] is unique among the rest of the studies: for both sets, the rate of inconclusive decisions is similar for same and different source evidence.

unintentional effect of masking the fact that examiners are not as accurate when evaluating different-source comparisons. While in most studies (except Bunch), the probability of making an inconclusive determination for different source evidence is below the probability of making an elimination, the probability that examiners will correctly make an elimination is still extremely low relative to the probability of a correct decision in the evaluation of same source evidence. It may be that examiners are trained to look for similarities, instead of differences, or that it is simply more difficult to classify a difference as opposed to a similarity; further, some labs inexplicably forbid eliminations unless there is a class characteristic mismatch. Regardless of the explanation, it is apparent that when presented with a different source comparison, examiners make an elimination less frequently than an identification when presented with a same-source comparison.

### 3.5 Decision-specific error rates

In courtroom testimony, we do not know ground truth - we do not know whether the comparison that is presented is from the same source or from different sources. As a result, we cannot use the source-specific error rates or probabilities in Figure 5 to justify the testimony and say something like “the probability of an identification given that the evidence is from the same source is nearly 1”. In these situations, the “data” we have is the examiner’s decision, so we wish to compute the reverse conditional probability: the probability of same-source evidence given that the examiner made an identification. Figure 6 shows the probability of same source evidence conditional on the examiner’s assessment for five published studies. Note that because there are two possible sources for the evidence (same or different), the equivalent probabilities for different-source

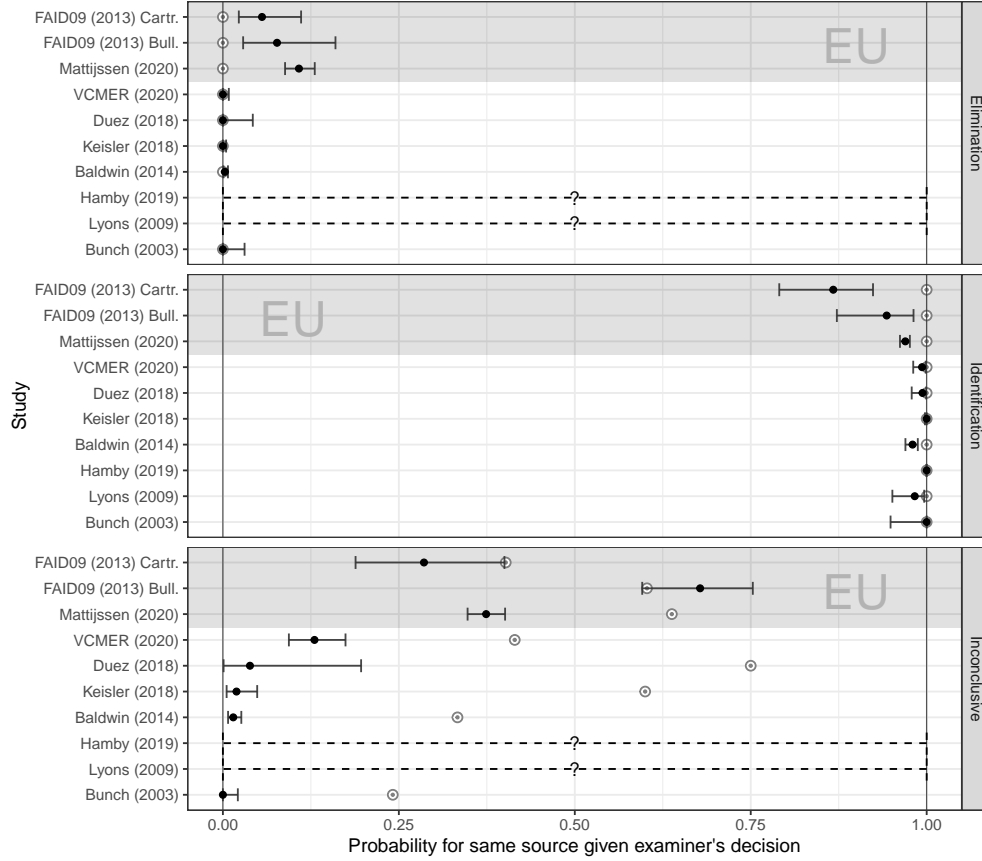


Figure 6: Pearson-Clopper 95% confidence intervals for the probability of same source evidence given an examiner's conclusion. Expected values for each of the probabilities as discussed in Section 4 are shown as grey targets. Some of the studies do not allow an assessment of all of the predictive values. Note that all 3 sets with non-AFTE conclusions [Pauw-Vugts et al., 2013, Mattijssen et al., 2020] have much higher rates of same-source elimination errors. Only FAID09 [Pauw-Vugts et al., 2013] has inconclusive rates which are similar to the overall rate of same-source evidence.

evidence can be obtained by subtracting the same-source probability from 1.

As before, the probabilities computed from the study results are shown with their respective 95% (Pearson-Clopper) confidence intervals. The circles shown in light grey in the middle panel show the expected proportion of inconclusive results which are from the same source. The calculation of these expected values will be discussed in more detail in the next section.

In Figure 6, it is clear that the probability of same-source evidence given an inconclusive evaluation is much lower than the expected proportion (target probability) for most studies. The decision-specific probabilities provide indications that the distribution of inconclusive determinations is very different from our expectations. One notable exception is FAID09 [Pauw-Vugts et al., 2013], which is a proficiency test (and thus, may be harder than the studies focused on error rates) of examiners in various European laboratories. While we cannot determine *why* this study is so different, it is clear that the discrepancy between inconclusive determinations and evidence source does not exist under all conditions. This suggests we may want to examine the source of this bias in more depth.

Table 6: Expected (target) conditional probabilities in a world where examiners do not make mistakes but do make inconclusive decisions.

$P(X Y)$	$Y = \text{Examiner Conclusion}$		
$X = \text{Ground Truth}$	Identification	Inconclusive	Elimination
Same source	$P(SS \text{Identification}) = 1$	$P(SS \text{Inconclusive}) = P(SS)$	$P(SS \text{Elimination}) = 0$
Different source	$P(DS \text{Identification}) = 0$	$P(DS \text{Inconclusive}) = P(DS)$	$P(DS \text{Elimination}) = 1$

## 4 Inconclusive Evaluations, Errors, and the Legal System

Before we discuss the results of inconclusive findings in the selected studies, we must explicitly characterize our expectations for decision-specific probabilities of inconclusive findings. That is, given that an examiner states that a finding is inconclusive, what is our expectation for the probability that the evidence is from the same source or from different sources? In the absence of any additional information, we have no reason to believe that a finding of inconclusive should be related to either same source or different source origin. An inconclusive finding, according to the AFTE rules, indicates that there is insufficient agreement or disagreement between discernible individual features to make an identification or elimination. This suggests that there is an implicit threshold of the amount of evidence necessary to arrive at a definitive conclusion [Dror and Langenburg, 2018]. We have no reason to believe that these thresholds are asymmetric; that is, the threshold to go from inconclusive to identification should be the same as the threshold to go from inconclusive to elimination. Statistically, we would express this symmetry as the independence between the source and an inconclusive finding: that is,  $P(\text{same source}|\text{inconclusive}) = P(\text{same source})$  and  $P(\text{different source}|\text{inconclusive}) = P(\text{different source})$ .

This implies that we can calculate the “target” probability for each decision. For same-source pairs, the target for identification is 100%: all same source pairs should result in an identification. Similarly, the target for elimination is 0%: not even a single same-source pair should result in an elimination. Same-source pairs should be assessed as inconclusive with frequency proportional to the number of same-source comparisons in the study. Conceptually, Figure 1 demonstrates this characteristic: the rate of inconclusives is the same for both same-source and different-source evidence. That is, under relatively ideal conditions (where there are inconclusive evaluations, but no errors according to the AFTE process), the expected conditional probabilities in an experiment would match those shown in Table 6.

In case work we are not able to assess the probability of same source or different source comparisons. However, those probabilities are easily accessible in formal studies or blinded proficiency testing, as  $P(SS)$  and  $P(DS)$  in these studies are determined by the experimental design.

As an example, in the Baldwin study [Baldwin et al., 2014], we would expect the probabilities for same and different source given an inconclusive result to be:

$$\begin{aligned} P(\text{same source} | \text{Inconclusive}) &= P(\text{same source}) = \frac{1}{3}, \\ P(\text{different source} | \text{Inconclusive}) &= P(\text{different source}) = \frac{2}{3}. \end{aligned}$$

This expectation is based on the study’s design: 5 of the 15 comparisons were among same-source pairs and 10 were among different-source pairs. However, in Baldwin, as in the other studies, the probability of same-source evidence given an inconclusive is lower than the expectation.

This finding is similar to the implications of Biedermann et al. [2019], though the authors do not follow their reported results to the logical conclusion. Specifically, they show that there is information in the finding of an inconclusive that is useful: inconclusive determinations are more likely to occur when the evidence originates from different sources. In the studies we examined, the vast majority of paired comparisons for which source

could not be conclusively established by the participant should have been eliminations. Therefore, we propose the following: if inconclusive determinations are essentially another way to say “different source”, then it makes sense to collapse inconclusive evaluations and eliminations, at least when calculating error rates.

Calculations are shown below for Baldwin et al. [2014]:

$$\begin{aligned}
P(\text{same source} \mid \text{elimination}) &= 4/1425 = 0.0028 \\
P(\text{same source} \mid \text{inconclusive or elimination}) &= (11 + 4)/(748 + 1425) = 0.0069 \\
P(\text{different source} \mid \text{identification}) &= 22/1097 = 0.0201.
\end{aligned} \tag{1}$$

Examining the experimentally determined error rates for firearms and toolmark identification leads to a disturbing realization. There is an order of magnitude difference between the probability of same-source evidence given an elimination and the probability of different-source evidence given an identification in studies conducted using examiners who are primarily from the US and Canada (e.g. excluding Mattijssen et al. [2020], Pauw-Vugts et al. [2013]). In fact, even if inconclusives and eliminations are combined as under Option 4 in Baldwin et al. [2014], the probability of a false elimination is 0.0069, only a third of the probability of a false identification. These observations are shown visually in Figure 7 and explicitly calculated in Equation (1) for Baldwin et al. [2014], the only study from the US or Canada with sufficient error counts. It is necessary to have at least one observation in each cell in order to bound the probabilities as well as sufficient comparisons to be able to differentiate between e.g. 0.02 and 0.2; this is particularly important when we deal with very low and very high probabilities, such as success and error rates. If we repeat these calculations using another large study, Mattijssen et al. [2020], however, we find that this issue is not universal.

$$\begin{aligned}
P(\text{same source} \mid \text{elimination}) &= 95/879 = 0.1081 \\
P(\text{same source} \mid \text{inconclusive or elimination}) &= (487 + 95)/(1302 + 879) = 0.2669 \\
P(\text{different source} \mid \text{identification}) &= 74/2439 = 0.0303
\end{aligned} \tag{2}$$

While we cannot conclude that this difference is due to the location of the examiners, or the protocol by which the study was conducted, it is clear that the studies conducted primarily on US examiners show a bias that does not exist in Mattijssen et al. [2020]; that is, we do not have to be content with this bias as the status quo. More studies are needed to determine which factors (training, location, legal system, system of reporting conclusions, study protocol) are most important in contributing to the clear bias in the distribution of inconclusives across same-source and different-source comparisons.

Note that while the probability of same source given an elimination is affected by lab policies towards the treatment of inconclusives, these policies do not explain the continued discrepancy in error probabilities when inconclusives and eliminations are considered together. The fact that all of the AFTE studies and none of the non-AFTE studies display this trend suggests that while lab policies do not explain the discrepancy, the training examiners receive, the scale used to evaluate evidence, and the underlying culture of the legal system (e.g. adversarial vs. non-adversarial) might be factors relevant to this discrepancy. The non-AFTE studies were conducted using worldwide or European participants, who have different training and ongoing evaluation requirements, as well as fundamentally different legal systems.

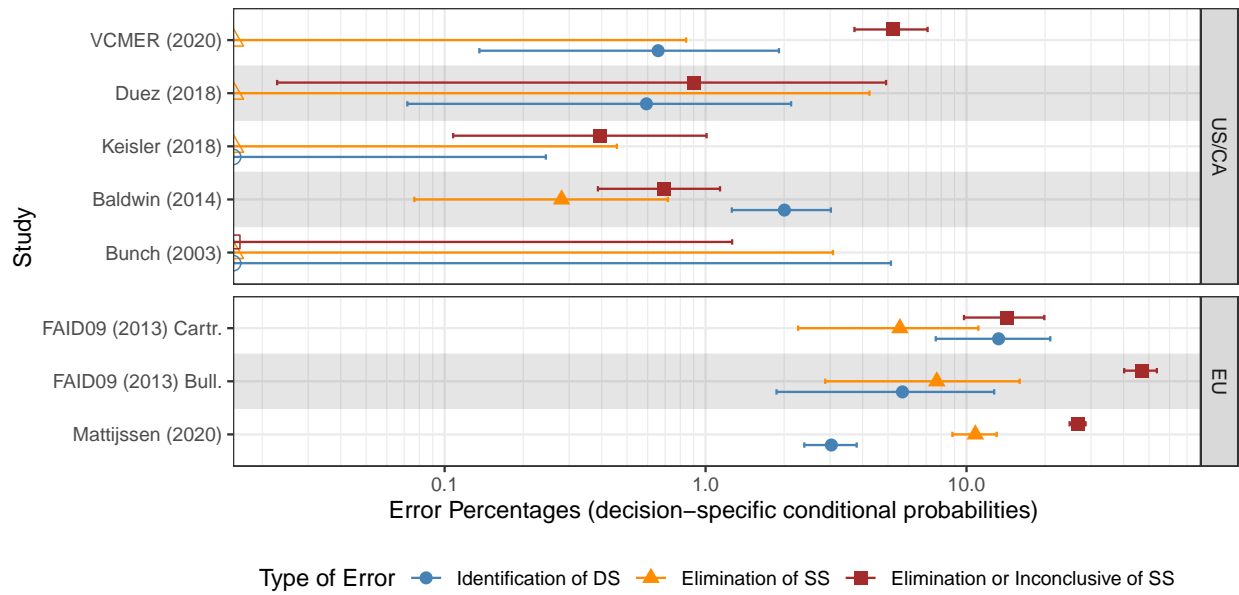


Figure 7: Decision-specific error rates. Note that the error percentages are shown on a log scale to allow visualizing differences in errors for both European studies and studies from the US and Canada. The European studies exhibit errors at an order of magnitude higher than US/CA based studies. For all US/CA studies in which any errors happened, errors in identification are more likely than errors in eliminations. This difference is significant in Baldwin et al. [2014], the only US/CA study with a large number of participants and evaluations. In all but the VCIMER (2020) study, counting inconclusives as eliminations does not significantly increase the corresponding error rate, and makes error rates from both DS and SS materials more similar.



Figure 7 focuses on the larger studies with designs which allow for estimation of these error rates, and shows the estimated probabilities (and the corresponding confidence intervals) of an incorrect identifications and eliminations. We will ignore [Mattijssen et al. \[2020\]](#) for the moment, as it was conducted using likelihood ratio based reasoning (with inconclusives as an afterthought) and thus is not completely comparable to studies which use an AFTE-like scale with inconclusives by default. In addition, we will exclude the Keisler study from consideration here because it did not record any definitive (not-inconclusive) errors; as a result, we cannot precisely estimate the rate of same source eliminations or different-source identifications. In the three remaining studies conducted using the AFTE range of conclusions that have at least one definitive error, the point estimate of the probability of a false identification is higher than the probability of a false elimination. In the case of [Baldwin et al. \[2014\]](#), the difference is statistically significant (Baldwin is the only study which observed definitive false identifications and false eliminations). While the trend is the same in the three studies, [Baldwin et al. \[2014\]](#) suggests that in the absence of definitive information, examiners tend to more often conclude identification than elimination. In the courtroom (or before, in a plea bargain situation), this results in a bias in favor of the prosecution (or against the defense).

There are many ways this bias could arise: there are, of course, well documented motivational and cognitive biases [[Giannelli, 2010](#), [Robertson and Kesselheim, 2016](#)] which may affect examiners. In addition, it may be easier to identify similarities than to explicitly identify dissimilarities [[Bagnara et al., 1983](#), [Ashby and Perrin, 1988](#)]. Examiner training primarily focus on making identifications, rather than spending an equal amount of time on making eliminations. In the case of bullet examination, there is an objective criterion for sufficient similarity: 6 consecutively matching striae, according to [Biasotti \[1959\]](#); similar criteria have been proposed for algorithmic assessment of cartridge cases [[Song et al., 2014](#)]. However, in both instances, there is not a corresponding threshold for dissimilarity. Here, we do not attempt to identify the source of the bias; we only identify that there is observable bias in the distribution of inconclusive results and suggest that this bias must be addressed. It is important to note that when we speak of biases, our conclusions apply to error rate studies, not to case work. In case work, fragmented or damaged evidence may cause an examiner to make an inconclusive determination; the studies we have examined here do not have this additional complication. While it would be interesting to design studies that included these types of comparisons, it is unclear how one would systematically collect such evidence.

## 5 Recommendations and Conclusions

It seems clear from our assessment of the currently available studies that there is significant work to be done before we can confidently state an error rate associated with different components of firearms and toolmark analysis. In particular, there is a need for studies that are both large (many examiners and many evaluations) and that meet the following design criteria:

- Single known source for all comparisons
- Open set, so that some questioned items do not originate from any of the sources included in the study and so that some sources do not contribute any questioned samples.
- Random allocation of different test kits to participants, to minimize the possibility of information exchange.
- Clear instructions, so that every participant follows the same protocol.
- Questioned samples that span the range of difficulty, including challenging comparisons.

All of the published AFTE scale studies that meet at least some of the design criteria focus only on cartridge cases; it is imperative that similarly well designed studies examining bullets and toolmarks be conducted and published to validate the entire discipline of firearms and toolmark analysis.

While published studies vary in their treatment of inconclusive decisions when calculating and reporting error rates or accuracies, most studies report what we have termed source-specific conditional probabilities.

We propose that error rates be calculated for the examiner and the process separately. Examiner-specific error rates should not count inconclusives as errors, while process-specific error rates should have to count inconclusives as errors. Process-specific error rates would then reflect the inability of the examiner to make a correct determination due to insufficient identifying information on the evidence. These two separate measures will be used differently: examiner error rates can be used for evaluation within the forensics lab setting (for e.g. qualification purposes in proficiency tests), but process error rates are essential for the proper use of firearms evidence in court: process error rates are germane to the question of whether evidence can inform about the suspect’s guilt.

We also argue that a simpler alternative to this treatment of error rates in legal settings would be to report the decision-specific conditional error rates: the probability that, given an examiner’s conclusion, that conclusion is incorrect. An error rate computed in this way provides the most relevant characterization of error in a court setting, where ground truth is not known, but the examiner’s decision is given. Further, decision-specific error rates avoid the need to treat inconclusives differently than identification or elimination.

In practice, using decision-specific error rates depends on the availability of information that is not currently collected. Specifically, we would need to know examiner-specific probabilities of incorrect conclusions as well as examiner-specific historical decisions. We might be able to learn the former from reliable proficiency testing, while the latter could be obtained from past lab reports and testimony.

Specifically, in court, we suggest that when the admissibility of an examiner’s testimony is assessed, the examiner is asked to state the lab policy on making eliminations and the rate at which the examiner makes inconclusive decisions, along with any relevant error rates specific to the lab and to themselves. We also suggest that during cross-examination after the testimony was presented, the defense ask about the decision-specific probability of an error (relative to the examiner’s decision). This information will provide error rates which are specific and relevant to the presented testimony.

In the process of examining these different error rate calculations, we discovered a series of systematic biases in error rate studies and the examination process which all work against the defendant. For example, in many studies, the study is designed such that it is possible to estimate false elimination rates more precisely than false identification rates. In fact, in some common study designs, it is only possible to estimate the false elimination error rate; the rate of false identifications cannot be estimated at all. This ensures that there is no way to invalidate the examiner’s testimony on the basis that it might contribute to a false conviction of the defendant. In addition, examiners working under the AFTE range of conclusions appear to have a lower threshold for identification than for elimination; when evidence originates from different sources, examiners are more likely to arrive at an inconclusive decision than they are when the evidence has the same source. As a result, there is a systematically higher probability of different-source evidence given an examiner’s identification than the probability of same-source evidence given an examiner’s elimination. This fundamentally contradicts the principles which form the foundation of our legal system. These biases mirror biases in the admissibility of forensic evidence which are also more likely to favor the prosecution [Moreno, 2004, Epstein, 2014, Roach, 2009]. This paper briefly examines two non-AFTE studies, FAID09 [Pauw-Vugts et al., 2013] and Mattijssen et al. [2020]; these studies do not demonstrate this bias, but it is not clear whether the difference is the population (examiners from outside the US and Canada), the scales used by the examiners, the legal systems, or some other factor. Additional studies should be performed to investigate possible explanations for this difference and to determine whether the perceived distinction between study types holds up under scrutiny.

An effective approach to eliminate the biases associated with the current approach to firearm and tool examination is to rely on automatic, objective algorithms to compute a *degree of similarity* between two items. Research and development of algorithms are underway [Chumbley et al., 2010, Song et al., 2018, Hare et al., 2017, Tai and Eddy, 2018], but more work must be conducted before they can be implemented in real case work. These algorithms generally include positive and negative criteria, which may contribute to their ability to make unbiased decisions about the similarity or dissimilarity of the evidence. Algorithms generally are symmetric in the assessment of positive and negative criteria – e.g. if a high number of consecutively

matching striae is considered evidence in favor of an identification, a low number of consecutively matching striae is consequently evidence in favor of an elimination. Through awareness of the biases that already exist in the evaluation of firearms and toolmark evidence, we can avoid the pitfalls of developing machine learning algorithms using biased training data [Howard and Borenstein, 2018, Lehr and Ohm, 2017-2018], ensuring that any automatic processes are as fair as possible. It is important that before these algorithms are applied to casework, we examine the distribution of automated scores and cutoff-based classifications compared to examiner decisions to determine whether the algorithms are, in fact, less biased than examiners. Even though the use of algorithms is not far in the future, we should also work to resolve the biases associated with the current practice of firearm and toolmark identification. At minimum, we should re-consider the definition and calculation of error rates and modify some of the factors which contribute to the uneven distribution of inconclusives.

## References

- AFTE Criteria for Identification Committee. Theory of identification, range striae comparison reports and modified glossary definitions. *AFTE Journal*, 24(3):336–340, 1992.
- F. Gregory Ashby and Nancy A. Perrin. Toward a unified theory of similarity and recognition. *Psychological Review*, 95(1):124–150, 1988. ISSN 1939-1471(Electronic),0033-295X(Print). doi:[10.1037/0033-295X.95.1.124](https://doi.org/10.1037/0033-295X.95.1.124).
- Sebastiano Bagnara, David B. Boles, Francesca Simion, and Carlo Umiltà. Symmetry and similarity effects in the comparison of visual patterns. *Perception & Psychophysics*, 34(6):578–584, November 1983. ISSN 0031-5117, 1532-5962. doi:[10.3758/BF03205914](https://doi.org/10.3758/BF03205914).
- David P Baldwin, Stanley J Bajic, Max Morris, and Daniel Zamzow. A Study of False-Positive and False-Negative Error Rates in Cartridge Case Comparisons. Technical report, Ames Lab IA, Performing, Fort Belvoir, VA, April 2014.
- A. A. Biasotti. A statistical study of the individual characteristics of fired bullets. *Journal of Forensic Sciences*, 4:34–50, 1959.
- Alex Biedermann, Silvia Bozza, Franco Taroni, and Joëlle Vuille. Are Inconclusive Decisions in Forensic Science as Deficient as They Are Said to Be? *Frontiers in Psychology*, 10, March 2019. ISSN 1664-1078. doi:[10.3389/fpsyg.2019.00520](https://doi.org/10.3389/fpsyg.2019.00520).
- David J. Brundage. *The Identification Of Consecutively Rifled Gun Barrels*. PhD thesis, Southern Illinois University at Carbondale, 1994. URL [https://vufind.carli.illinois.edu/vf-sic/Record/sic\\_1201372/Description](https://vufind.carli.illinois.edu/vf-sic/Record/sic_1201372/Description).
- David J. Brundage. The Identification of Consecutively Rifled Gun Barrels. *AFTE Journal*, 30(3):438–444, 1998.
- Stephen Bunch and Douglas Murphy. A comprehensive validity study for the forensic examination of cartridge cases. *AFTE Journal*, 35(2):201–203, 2003.
- Ward Casscells, Arno Schoenberger, and Thomas B Graboys. Interpretation by Physicians of Clinical Laboratory Results. *New England Journal of Medicine*, 299(18):999–1001, November 1978. doi:[10.1056/NEJM197811022991808](https://doi.org/10.1056/NEJM197811022991808).
- Chad Chapnick, Todd J. Weller, Pierre Duez, Eric Meschke, John Marshall, and Ryan Lilien. Results of the 3D Virtual Comparison Microscopy Error Rate (VCMER) Study for firearm forensics. *J Forensic Sci*, 66(2):557–570, Mar 2021. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/1556-4029.14602>.

- Wei Chu, Robert M. Thompson, John Song, and Theodore V. Vorburger. Automatic identification of bullet signatures based on consecutive matching striae (CMS) criteria. *Forensic Science International*, 231(1-3): 137–141, 2013. ISSN 03790738. doi:[10.1016/j.forsciint.2013.04.025](https://doi.org/10.1016/j.forsciint.2013.04.025).
- L. Scott Chumbley, Max D. Morris, M. James Kreiser, Charles Fisher, Jeremy Craft, Lawrence J. Genalo, Stephen Davis, David Faden, and Julie Kidd. Validation of Tool Mark Comparisons Obtained Using a Quantitative, Comparative, Statistical Algorithm. *Journal of Forensic Sciences*, 55(4):953–961, July 2010. ISSN 1556-4029. doi:[10.1111/j.1556-4029.2010.01424.x](https://doi.org/10.1111/j.1556-4029.2010.01424.x).
- Itiel Dror. The Error in ‘Error Rate’: Why Error Rates Are So Needed, Yet So Elusive. SSRN Scholarly Paper ID 3593309, Social Science Research Network, Rochester, NY, May 2020. URL <https://papers.ssrn.com/abstract=3593309>.
- Itiel E Dror and Glenn Langenburg. “Cannot Decide”: The Fine Line Between Appropriate Inconclusive Determinations Versus Unjustifiably Deciding Not To Decide. *Journal of Forensic Sciences*, 64(1):10–15, May 2018. doi:[10.1111/1556-4029.13854](https://doi.org/10.1111/1556-4029.13854).
- Pierre Duez, Todd Weller, Marcus Brubaker, Richard E. Hockensmith II, and Ryan Lilien. Development and Validation of a Virtual Examination Tool for Firearm Forensics. *Journal of Forensic Sciences*, 63(4): 1069–1084, October 2018. doi:[10.1111/1556-4029.13668](https://doi.org/10.1111/1556-4029.13668).
- Jules Epstein. Preferring the Wise Man to Science: The Failure of Courts and Non-Litigation Mechanisms to Demand Validity in Forensic Matching Testimony. *Widener Law Review*, 20(1):81–118, 2014. URL <https://heinonline.org/HOL/P?h=hein.journals/wlsj20&i=87>.
- T.G. Fadul Jr., G.A. Hernandez, Stephanie Stoiloff, and S. Gulat. An empirical study to improve the scientific foundation of forensic firearm and tool mark identification utilizing 10 consecutively manufactured slides. *AFTE Journal*, 45(4):376–389, December 2012.
- Paul C. Giannelli. Daubert: Interpreting the Federal Rules of Evidence Scientific Evidence after the Death of Frye. *Cardozo Law Review*, 15(6):1999–2026, 1993. URL <https://heinonline.org/HOL/P?h=hein.journals/cdozo15&i=2026>.
- Paul C. Giannelli. Independent crime laboratories: The problem of motivational and cognitive bias. *Utah Law Review*, page 247, 2010.
- James E. Hamby, David J. Brundage, and James W. Thorpe. The Identification of Bullets Fired from 10 Consecutively Rifled 9mm Ruger Pistol Barrels: A Research Project Involving 507 Participants from 20 Countries. *AFTE Journal*, 41(2):99–110, 2009.
- James E. Hamby, David J. Brundage, Nicholas D. K. Petraco, and James W. Thorpe. A Worldwide Study of Bullets Fired From 10 Consecutively Rifled 9MM RUGER Pistol Barrels—Analysis of Examiner Error Rate. *Journal of Forensic Sciences*, 64(2):551–557, 2019. ISSN 1556-4029. doi:[10.1111/1556-4029.13916](https://doi.org/10.1111/1556-4029.13916). 00000.
- Eric Hare, Heike Hofmann, and Alicia Carriquiry. Automatic matching of bullet land impressions. *Ann. Appl. Stat.*, 11(4):2332–2356, 12 2017. doi:[10.1214/17-AOAS1080](https://doi.org/10.1214/17-AOAS1080).
- Ayanna Howard and Jason Borenstein. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and engineering ethics*, 24(5):1521–1536, 2018.
- Mark A. Keisler, Stacey Hartman, Angela Kilmon, Melissa Oberg, and Mitzi Templeton. Isolated pairs research study. *AFTE Journal*, 50(1):56–58, November 2018.
- W. Kerkhoff, R. D. Stoel, E. J. A. T. Mattijssen, C. E. H. Berger, F. W. Didden, and J. H. Kerstholt. A part-declared blind testing program in firearms examination. *Science & Justice*, 58(4):258–263, July 2018. ISSN 1355-0306. doi:[10.1016/j.scijus.2018.03.006](https://doi.org/10.1016/j.scijus.2018.03.006).

- Jonathan J. Koehler. Fingerprint Error Rates and Proficiency Tests: What They are and Why They Matter Symposium. *Hastings Law Journal*, 59(5):1077–1100, 2007. URL <https://heinonline.org/HOL/P?h=hein.journals/hastlj59&i=1117>.
- Jonathan J. Koehler. Proficiency tests to estimate error rates in the forensic sciences. *Law, Probability and Risk*, 12(1):89–98, March 2013. ISSN 1470-840X. doi:[10.1093/lpr/mgs013](https://doi.org/10.1093/lpr/mgs013).
- David Lehr and Paul Ohm. Playing with the data: What legal scholars should learn about machine learning. *U.C. Davis Law Review*, 51:653, 2017-2018.
- Ryan Lilien. Firearm Forensics Black-Box Studies for Examiners and Algorithms using Measured 3D Surface Topographies. NIJ Award 2017-IJ -CX-0024 254338, Cadre Forensics, 2019. URL <https://www.ncjrs.gov/pdffiles1/nij/grants/254338.pdf>.
- D. J. Lyons. The identification of consecutively manufactured extractors. *AFTE Journal*, 41(3):246–256, 2009.
- Arjun K Manrai, Gaurav Bhatia, Judith Strymish, Isaac S Kohane, and Sachin H Jain. Medicine’s Uncomfortable Relationship With Math. *JAMA Internal Medicine*, 174(6):991–993, June 2014. doi:[10.1001/jamainternmed.2014.1059](https://doi.org/10.1001/jamainternmed.2014.1059).
- Erwin J. A. T. Mattijssen, Cilia L. M. Witteman, Charles E. H. Berger, Nicolaas W. Brand, and Reinoud D. Stoel. Validity and reliability of forensic firearm examiners. *Forensic Science International*, 307, February 2020. ISSN 0379-0738. doi:[10.1016/j.forsciint.2019.110112](https://doi.org/10.1016/j.forsciint.2019.110112).
- Joelle Anne Moreno. What Happens When Dirty Harry Becomes an (Expert) Witness for the Prosecution. *Tulane Law Review*, 79(1):1–54, 2004. URL <https://heinonline.org/HOL/P?h=hein.journals/tulr79&i=15>.
- National Research Council Committee on Identifying the Needs of the Forensic Sciences Community. Strengthening Forensic Science in the United States: A Path Forward. *National Academies Press*, 2009.
- P. Pauw-Vugts, A. Walters, L. Øren, and L. Pfoser. FAID 2009: Proficiency test and workshop. *AFTE Journal*, 45(2):115–127, 2013.
- President’s Council of Advisors on Science and Technology (PCAST). Report on forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods, 2016. URL [https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast\\_forensic\\_science\\_report\\_final.pdf](https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf).
- President’s Council of Advisors on Science and Technology (PCAST). Addendum to the Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature Comparison Methods. Technical report, January 2017. URL [https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast\\_forensics\\_addendum\\_finalv2.pdf](https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_addendum_finalv2.pdf).
- Kent Roach. FORENSIC SCIENCE AND MISCARRIAGES OF JUSTICE: SOME LESSONS FROM COMPARATIVE EXPERIENCE. *Jurimetrics*, 50(1):67–92, 2009. ISSN 0897-1277. URL <https://www.jstor.org/stable/41550027>.
- C.T. Robertson and A.S. Kesselheim. *Blinding as a solution to bias: Strengthening biomedical science, forensic science, and law*. Elsevier Science, 2016. ISBN 978-0-12-802633-5.
- John Song, Wei Chu, Mingsi Tong, and Johannes Soons. 3d topography measurements on correlation cells: a new approach to forensic ballistics identifications. *Measurement Science and Technology*, 25(6), 2014. ISSN 0957-0233. doi:[10.1088/0957-0233/25/6/064005](https://doi.org/10.1088/0957-0233/25/6/064005).
- John Song, Theodore V. Vorburger, Wei Chu, James Yen, Johannes A. Soons, Daniel B. Ott, and Nien Fan Zhang. Estimating error rates for firearm evidence identifications in forensic science. *Forensic Science International*, 284:15–32, March 2018. ISSN 03790738. doi:[10.1016/j.forsciint.2017.12.013](https://doi.org/10.1016/j.forsciint.2017.12.013).

- C. Spiegelman and 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. *Law, Probability and Risk*, 12(2):115–133, June 2013. doi:[10.1093/lpr/mgs028](https://doi.org/10.1093/lpr/mgs028).
- A. Stroman. Empirically Determined Frequency of Error in Cartridge Case Examinations Using a Declared Double-Blind Format. *AFTE Journal*, 46(2):157–174, 2014.
- Xiao Hui Tai and William F Eddy. A fully automatic method for comparing cartridge case images. *Journal of Forensic Sciences*, 63(2):440–448, 2018.
- Susan Vanderplas, Melissa Nally, Tylor Klep, Cristina Cadevall, and Heike Hofmann. Comparison of three similarity scores for bullet LEA matching. *Forensic Science International*, page 110167, 2020. ISSN 0379-0738. doi:<https://doi.org/10.1016/j.forsciint.2020.110167>.

## A Study Summaries and Results

**Baldwin** The Baldwin study [Baldwin et al., 2014] was designed such that each test kit consists of 15 sets of 3 known cartridge cases and 1 questioned cartridge case. In 5 of the 15 sets the questioned cartridge was from the same source as the knowns, while the other 10 questioned cartridges were from different sources as their respective knowns.

25 firearms were used for the study, such that within each kit no firearm was re-used for either knowns or questioned cartridge cases, i.e. no additional information could be gained by comparing any cartridge cases across sets.

Results are summarized in Table 7.

Table 7: Baldwin study results, conclusion-specific and source-specific probabilities, and reported overall error rates.				
<div>A Study of False-Positive and False-Negative Error Rates in Cartridge Case Comparisons <i>Baldwin et al. ( 2014)</i></div>				
Study Type	Test Set		Participants	
	# SS Comparisons	# DS Comparisons		
Open set	5	10	218 examiners	
Experiment Count Data				
	Identification	Inconclusive	Elimination	Source Total
Same Source	1075	11	4	1090
Different source	22	735+2 <sup>a</sup>	1421	2180
Conclusion Total	1097	748	1425	3270
Conclusion-Specific Probabilities				
	Identification	Inconclusive	Elimination	<i>Experiment Count Data are divided by Conclusion Totals (3rd row)</i>
Same source	0.9799	0.0147	0.0028	
Different source	0.0201	0.9853	0.9972	
Total # Comparisons	1097	748	1425	
Source-Specific Probabilities				
	Identification	Inconclusive	Elimination	Total # Comparisons
Same source	0.9862	0.0101	0.0037	1090
Different source	0.0101	0.3381	0.6518	2180
<i>Experiment Count Data are divided by Source Totals (last column)</i>				
Overall Error Rates				
Opt.	Meaning	Missed Identification	Missed Elimination	Total
2	FTE error	0.0037	0.0101	0.0080
3	Process error	0.0138	0.3482	0.2367
4	Inconcl. = Elim.	0.0138	0.0101	0.0113
<sup>a</sup> Two comparisons were not reported and are considered to be inconclusives.				



**Keisler** The Keisler Study [Keisler et al., 2018] was designed so that each test kit consisted of sets of 20 pairs of cartridge cases from Smith & Wesson pistols, where 12 of the pairs were from the same source and 8 pairs were from different sources.

Kits were assembled using only 9 Smith & Wesson pistols (i.e. there is a potential to gain additional information by making comparisons across sets). However, participants were instructed to only compare single pairs.

Results from the study are shown in Table 8.

Table 8: Isolated Pairs Research Study				
<i>Keisler et al. (2018)</i>				
A study of Smith & Wesson cartridge cases.				
Study Type	Test Set		Participants	
	# SS Comparisons	# DS Comparisons		
Open set	12	8	126 examiners	
Experiment Count Data				
	Identification	Inconclusive	Elimination	Source Total
Same source	1508	4	0	1512
Different source	0	203	805	1008
Conclusion Total	1508	207	805	2520
Conclusion-Specific Probabilities				
	Identification	Inconclusive	Elimination	<i>Experiment Count Data are divided by Conclusion Totals (3rd row)</i>
Same source	1.0000	0.0193	0.0000	
Different source	0.0000	0.9807	1.0000	
Total # Comparisons	1508	207	805	
Source-Specific Probabilities				
	Identification	Inconclusive	Elimination	Total # Comparisons
Same source	0.9974	0.0026	0.0000	1512
Different source	0.0000	0.2014	0.7986	1008
<i>Experiment Count Data are divided by Source Totals (last column)</i>				
Overall Error Rates				
Opt.	Meaning	Missed Identification	Missed Elimination	Total
2	FTE error	0.0000	0.0000	0.0000
3	Process error	0.0026	0.2014	0.0821
4	Inconcl. = Elim.	0.0026	0.0000	0.0016

**Duez** The Duez study [Duez et al., 2018] used virtual microscopy to evaluate scans of cartridge cases. Each participant was asked to make eight evaluations of breech face impressions. These consisted of two sets:

CCTS1 set of three knowns, four questioned breech face impressions. All questioned breech face impressions are from the same source as the knowns.

CCTS2 set of three knowns, four questioned breech face impressions. Two questioned breech face impressions are from the same source, two are from different sources.

Both sets were evaluated by 56 participants (46 fully certified examiners and 10 trainees). CCTS1 resulted in 56 x 4 correct identifications. The design of the experiment does not allow us to quantify all of the quantities to evaluate examiner performance unless we aggregate performance over both sets. Table 9 shows the results of this aggregation.

Table 9: Duez study results, conclusion-specific and source-specific probabilities, and reported overall error rates. For the conditional probabilities, the differences in certified examiners and trainees are not large, so only the aggregate results are shown. Certified examiners have a perfect identification rate, only evaluations of different source pairs lead to inconclusives.				
Development and Validation of a Virtual Examination Tool for Firearm Forensics <i>Duez et al. ( 2018)</i>				
Study Type	Test Set		Participants	
	# SS Comparisons	# DS Comparisons		
Open set	6	2	46+10 examiners	
Experiment Count Data				
	Identification	Inconclusive	Elimination	Source Total
Same source	276+59	0+1	0+0	276+60
Different source	0+2	12+13	80+5	92+20
Conclusion Total	276+61	12+14	80+5	448
Conclusion-Specific Probabilities				
	Identification	Inconclusive	Elimination	<i>Experiment Count Data are divided by Conclusion Totals (3rd row)</i>
Same source	0.9941	0.0385	0.0000	
Different source	0.0059	0.9615	1.0000	
Total # Comparisons	337	26	85	
Source-Specific Probabilities				
	Identification	Inconclusive	Elimination	Total # Comparisons
Same source	0.9970	0.0030	0.0000	336
Different source	0.0179	0.2232	0.7589	112
<i>Experiment Count Data are divided by Source Totals (last column)</i>				
Overall Error Rates				
Opt.	Meaning	Missed Identification	Missed Elimination	Total
2	FTE error	0.0000	0.0179	0.0045
3	Process error	0.0030	0.2411	0.0625
4	Inconcl. = Elim.	0.0030	0.0179	0.0067

**Brundage-Hamby** The Brundage-Hamby study consists of sets of 20 test fires from known barrels and 15 questioned bullets. The 20 known test fires are 2 bullets from each of ten consecutively manufactured barrels. The Brundage-Hamby study is a closed set study, i.e. the 15 questioned bullets are known to be fired from one of these ten barrels. Participants (firearm examiners) are asked to identify which of the knowns a questioned bullet matches.

The study was originally reported on by Brundage in 1998 [Brundage, 1998]. Updates on the study with increasing number of responses have been published several times since Hamby et al. [2009, 2019]. The design of the Brundage-Hamby is well known in the forensics community and has been copied in Fadul Jr. et al. [2012] for a study of cartridge cases of the same firearms. Slight modifications of the study design are also common [Lyons, 2009].

Results reported in the 2019 paper are shown in Table 10.

Table 10: Evaluations, conditional error rates, and overall error rates from the combined Brundage-Hamby studies. Note the focus on identifications.

A Worldwide Study of Bullets Fired From 10 Consecutively Rifled 9MM RUGER Pistol Barrels - Analysis of Examiner Error Rate <i>Hamby et al. ( 2019)</i>				
Study Type	Test Set		Participants	
	# Knowns	# Unknowns		
Closed set	10	15	507 examiners	
Experiment Count Data				
	Identification	Inconclusive	Elimination	Source Total
Same Source	10447	8	0	10455
Different source	0	?	?	47047.5 <sup>a</sup>
Conclusion Total	10447	?	?	57502.5
Conclusion-Specific Probabilities				
	Identification	Inconclusive	Elimination	
Same source	1.0000	?	?	<i>Experiment Count Data are divided by Conclusion Totals (3rd row)</i>
Different source	0.0000	?	?	
Total # Comparisons	10447	?	?	
Source-Specific Probabilities				
	Identification	Inconclusive	Elimination	Total # Comparisons
Same source	0.9992	0.0008	0.0000	10455
Different source	?	?	?	?
<i>Experiment Count Data are divided by Source Totals (last column)</i>				
Overall Error Rates				
Opt.	Meaning	Missed Identification	Missed Elimination	Total
2	FTE error	0.0000	?	?
3	Process error	0.0008	?	?
4	Inconcl. = Elim.	0.0008	?	?
<sup>a</sup> This number is imputed based on the average number of pairwise comparisons an examiner would have to do to complete a Hamby study. Details can be found in the supplemental material.				

**Bunch** [Bunch and Murphy, 2003] used a study design that is not conducive to a quick summary. The study consisted of 8 test kits of varying composition; the test kits were evaluated by 8 examiners at the FBI laboratory.

Table 11: Results and summary tables for the Bunch study. Values in this table are as reported in the study; however, due to the study’s structure, it is possible to perform fewer comparisons, because not all comparisons are independent. Results for independent comparisons are shown in Table 19

A Comprehensive Validity Study for the Forensic Examination of Cartridge Cases <i>Bunch and Murphy (2003)</i>				
Study Type	Test Set		Participants	
	# Knowns	# Unknowns		
Open set	variable	variable	8 FBI examiners	
Reported (Nominal) Count Data				
	Identification	Inconclusive	Elimination	Source Total
Same source	70	0	0	70
Diff source	0	172	118	290
Conclusion Total	70	172	118	360
Conclusion-Specific Probabilities				
	Identification	Inconclusive	Elimination	<i>Experiment Count Data are divided by Conclusion Totals (3rd row)</i>
Same source	1.0000	0.0000	0.0000	
Different source	0.0000	1.0000	1.0000	
Total # Comparisons	70	172	118	
Source-Specific Probabilities				
	Identification	Inconclusive	Elimination	Total # Comparisons
Same source	1.0000	0.0000	0.0000	70
Different source	0.0000	0.5931	0.4069	290
<i>Experiment Count Data are divided by Source Totals (last column)</i>				
Overall Error Rates				
Opt.	Meaning	Missed Identification	Missed Elimination	Total
2	FTE error	0.0000	0.0000	0.0000
3	Process error	0.0000	0.5931	0.4778
4	Inconcl. = Elim.	0.0000	0.0000	0.0000

**Lyons** The Lyons study [Lyons, 2009] examined marks made by 10 consecutively manufactured extractors (manufactured by Caspian Arms Ltd).

Kits were assembled from 32 cartridge cases: 20 known cartridge cases from pairs of 2 cartridges from each of the 10 extractors (the knowns) and 12 questioned cartridges, such that each known corresponded to at least one questioned, with some replication in most of the kits (one kit accidentally only had ten questioned cartridges). Thus, the setup of this study is similar to the Brundage-Hamby study. This study suffers from the same problems as the Brundage-Hamby study: it is a closed set study with multiple knowns and asks for identifications only. We can therefore only estimate a fraction of the relevant error rates. It is also not possible to determine the total number of independent different source comparisons. The study results are shown in Table 12, along with computed conclusion-specific and source-specific probabilities and error rates.

Table 12				
The Identification of Consecutively Manufactured Extractors <i>Lyons ( 2009)</i>				
Study Type	Test Set		Participants	
	# Knowns	# Unknowns		
Closed set	10	10 or 12	15 examiners	
Experiment Count Data				
	Identification	Inconclusive	Elimination	Source Total
Same Source	174 <sup>a</sup>	1	3	178
Different source	3	?	?	?
Conclusion Total	177	?	?	?
Conclusion-Specific Probabilities				
	Identification	Inconclusive	Elimination	<i>Experiment Count Data are divided by Conclusion Totals (3rd row)</i>
Same source	0.9831	?	?	
Different source	0.0169	?	?	
Total # Comparisons	177	?	?	
Source-Specific Probabilities				
	Identification	Inconclusive	Elimination	Total # Comparisons
Same source	0.9775	0.0056	0.0169	178
Different source	?	?	?	?
<i>Experiment Count Data are divided by Source Totals (last column)</i>				
Overall Error Rates				
Opt.	Meaning	Missed Identification	Missed Elimination	Total
2	FTE error	0.0169	?	?
3	Process error	0.0225	?	?
4	Inconcl. = Elim.	0.0225	?	?
<sup>a</sup> Lyons [2009] reports 175 correct identifications, but it is clear from the discussion that one of those same source identifications was in fact an inconclusive. Twelve answer sheets with 12 correct identifications, one with 10 (out of 10) correct identifications, one with 9 correct identifications and 3 errors, and one with 11 correct identifications and one inconclusive. So 12·12 + 10 + 9 + 11 = 174.				

**VCMER** The VCIMER study [Chapnick et al., 2021] used virtual microscopy to evaluate scans of cartridge cases. Each participant was asked to make sixteen evaluations of breech face impressions, selected from a total of forty sets in a balanced incomplete block design. Each set consisted of two exemplars and one questioned bullet. Data from this study was also reported in Lilien [2019].

Table 13: VCIMER study results, conclusion-specific and source-specific probabilities, and reported overall error rates.

Firearm Forensics Black-Box Studies for Examiners and Algorithms using Measured 3D Surface Topographies <i>Chapnick et al. ( 2021)</i>				
Study Type	Test Set		Participants	
	# SS Comparisons	# DS Comparisons		
Open set	17	23	76 examiners	
Experiment Count Data				
	Identification	Inconclusive	Elimination	Source Total
Same source	453	38	0	491
Different source	3	254	436	693
Conclusion Total	456	292	436	1184
Conclusion-Specific Probabilities				
	Identification	Inconclusive	Elimination	<i>Experiment Count Data are divided by Conclusion Totals (3rd row)</i>
Same source	0.9934	0.1301	0.0000	
Different source	0.0066	0.8699	1.0000	
Total # Comparisons	456	292	436	
Source-Specific Probabilities				
	Identification	Inconclusive	Elimination	Total # Comparisons
Same source	0.9226	0.0774	0.0000	491
Different source	0.0043	0.3665	0.6291	693
<i>Experiment Count Data are divided by Source Totals (last column)</i>				
Overall Error Rates				
Opt.	Meaning	Missed Identification	Missed Elimination	Total
2	FTE error	0.0000	0.0043	0.0025
3	Process error	0.0774	0.3709	0.2492
4	Inconcl. = Elim.	0.0774	0.0043	0.0346

**Mattijssen** The Mattijssen study [Mattijssen et al., 2020] examined firing pin aperture shear marks made by 200 9mm Luger Glock pistols confiscated in the Netherlands. The study considered both 3D scans and 2D images; examinations were of the resulting digital representations rather than the actual physical objects. Seventy-seven examiners from 5 continents participated in the study; of these, 75 were fully qualified examiners. Of the participants, 58 indicated that they provided categorical conclusions (exclusion/inclusion/inconclusive), 13 provided probabilistic conclusions, and 6 used a 5-step reporting scale as in FAID09 [Pauw-Vugts et al., 2013].

Participants were initially shown comparison images (aligned by computer algorithm) with varying degrees of similarity as a calibration step. Then, 60 comparison images were shown (consisting of 60 sets of either matching or non-matching shear marks, aligned by computer algorithm). Participants were first asked to determine the similarity on a 5-point scale from (almost) no similarity to (almost) total similarity. Once these evaluations were complete, participants were shown each comparison image again, and were asked to answer 3 additional questions assessing 1) conclusion consistent with same/different source, 2) degree of support for this conclusion, in 6 stages corresponding to approximate likelihood ratios, and 3) whether the examiner would have provided an inconclusive conclusion in casework. The addition of these 3 questions allowed the authors to frame this study in a way that is consistent with the AFTE theory of identification, and thus, in a way which is compatible with the way study results are analyzed in this analysis (though of course the methodology is different).

Table 14: Results for the study’s examination of error rates in firearm examiners				
Validity and Reliability of Forensic Firearms Examiners <i>Mattijssen et al. ( 2020)</i>				
Study Type	Test Set		Participants	
	# Knowns	# Unknowns		
Open set	38	22	77 examiners	
Experiment Count Data				
	Identification	Inconclusive	Elimination	Source Total
Same Source	2365	487	95	2947
Different source	74	815	784	1673
Conclusion Total	2439	1302	879	4620
Conclusion-Specific Probabilities				
	Identification	Inconclusive	Elimination	<i>Experiment Count Data are divided by Conclusion Totals (3rd row)</i>
Same source	0.9697	0.3740	0.1081	
Different source	0.0303	0.6260	0.8919	
Total # Comparisons	2439	1302	879	
Source-Specific Probabilities				
	Identification	Inconclusive	Elimination	Total # Comparisons
Same source	0.8025	0.1653	0.0322	2947
Different source	0.0442	0.4871	0.4686	1673
<i>Experiment Count Data are divided by Source Totals (last column)</i>				
Overall Error Rates				
Opt.	Meaning	Missed Identification	Missed Elimination	Total
2	FTE error	0.0322	0.0442	0.0366
3	Process error	0.1975	0.5314	0.3184
4	Inconcl. = Elim.	0.1975	0.0442	0.1420

Interestingly, Mattijssen et al. [2020] also examined the conclusions drawn from a cross-correlation based



similarity assessment in both 2D (images) and 3D (scans). We include the reference comparison tables for these as well, because this allows us to compare the results from an objective algorithm with the results from examiners. Note that the algorithms do not produce inconclusive results.

Table 15: Results from a set of comparisons of 2D images evaluated with an automatic algorithm.

Validity and Reliability of Forensic Firearms Examiners <i>Mattijssen et al. ( 2020)</i>				
Study Type	Test Set		Participants	
	# Knowns	# Unknowns		
Open set	200	79600	Computer (2D algorithm)	
Experiment Count Data				
	Identification	Inconclusive	Elimination	Source Total
Same Source	198	0	2	200
Different source	1012	0	78588	79600
Conclusion Total	1210	0	78590	79800
Conclusion-Specific Probabilities				
	Identification	Inconclusive	Elimination	<i>Experiment Count Data are divided by Conclusion Totals (3rd row)</i>
Same source	0.1636	-	0.0000	
Different source	0.8364	-	1.0000	
Total # Comparisons	1210	-	78590	
Source-Specific Probabilities				
	Identification	Inconclusive	Elimination	Total # Comparisons
Same source	0.9900	-	0.0100	200
Different source	0.0127	-	0.9873	79600
<i>Experiment Count Data are divided by Source Totals (last column)</i>				
Overall Error Rates				
Opt.	Meaning	Missed Identification	Missed Elimination	Total
2	FTE error	0.0100	0.0127	0.0127
3	Process error	0.0100	0.0127	0.0127
4	Inconcl. = Elim.	0.0100	0.0127	0.0127

Table 16: Results from a set of comparisons of 3D images evaluated with an automatic algorithm.

**Validity and Reliability of Forensic Firearms Examiners** *Mattijssen et al. ( 2020)*

Study Type	Test Set		Participants		
	# Knowns	# Unknowns			
Open set	200	79600	Computer (3D algorithm)		
Experiment Count Data					
	Identification	Inconclusive	Elimination	Source Total	
Same Source	198	0	2	200	
Different source	999	0	78601	79600	
Conclusion Total	1197	0	78603	79800	
Conclusion-Specific Probabilities				<i>Experiment Count Data are divided by Conclusion Totals (3rd row)</i>	
	Identification	Inconclusive	Elimination		
Same source	0.1654	-	0.0000		
Different source	0.8346	-	1.0000		
Total # Comparisons	1197	-	78603		
Source-Specific Probabilities				<i>Experiment Count Data are divided by Source Totals (last column)</i>	
	Identification	Inconclusive	Elimination		Total # Comparisons
Same source	0.9900	-	0.0100		200
Different source	0.0126	-	0.9874		79600
Overall Error Rates					
Opt.	Meaning	Missed Identification	Missed Elimination	Total	
2	FTE error	0.0100	0.0126	0.0125	
3	Process error	0.0100	0.0126	0.0125	
4	Inconcl. = Elim.	0.0100	0.0126	0.0125	

**FAID09** The FAID09 study [Pauw-Vuğts et al., 2013] consists of 10 sets of 3 knowns and 1 unknown, in both bullets and cartridge cases. Test sets were castings of the original fired bullets and cases, rather than the objects themselves. Examiners reported conclusions on a 5-point (+ unsuitable) scale, where A corresponds to identification, B corresponds to probable identification or AFTE Inconclusive-A, C corresponds to inconclusive or AFTE Inconclusive B, D corresponds to probable exclusion, or AFTE Inconclusive C, and E corresponds to Exclusion; Z corresponds to unsuitable. There were between 62 and 64 evaluations of each set.

Table 17: FAID09 study results for bullets, conclusion-specific and source-specific probabilities, and reported overall error rates.

FAID2009 Proficiency Test and Workshop <i>Pauw-Vuğts et al. ( 2013)</i>				
Study Type	Test Set		Participants	
	# SS Comparisons	# DS Comparisons		
Open set	188	124	64 examiners	
Experiment Count Data				
	Identification	Inconclusive	Elimination	Source Total
Same source	83	99	6	188
Different source	5	47	72	124
Conclusion Total	88	146	78	312
Conclusion-Specific Probabilities				
	Identification	Inconclusive	Elimination	<i>Experiment Count Data are divided by Conclusion Totals (3rd row)</i>
Same source	0.9432	0.6781	0.0769	
Different source	0.0568	0.3219	0.9231	
Total # Comparisons	88	146	78	
Source-Specific Probabilities				
	Identification	Inconclusive	Elimination	Total # Comparisons
Same source	0.4415	0.5266	0.0319	188
Different source	0.0403	0.3790	0.5806	124
<i>Experiment Count Data are divided by Source Totals (last column)</i>				
Overall Error Rates				
Opt.	Meaning	Missed Identification	Missed Elimination	Total
2	FTE error	0.0319	0.0403	0.0353
3	Process error	0.5585	0.4194	0.5032
4	Inconcl. = Elim.	0.5585	0.0403	0.3526

Table 18: FAID09 study results for cartridges, conclusion-specific and source-specific probabilities, and reported overall error rates.

FAID2009 Proficiency Test and Workshop <i>Pauw-Vugts et al. ( 2013)</i>				
Study Type	Test Set		Participants	
	# SS Comparisons	# DS Comparisons		
Open set	127	189	64 examiners	
Experiment Count Data				
	Identification	Inconclusive	Elimination	Source Total
Same source	98	22	7	127
Different source	15	55	119	189
Conclusion Total	113	77	126	316
Conclusion-Specific Probabilities				
	Identification	Inconclusive	Elimination	<i>Experiment Count Data are divided by Conclusion Totals (3rd row)</i>
Same source	0.8673	0.2857	0.0556	
Different source	0.1327	0.7143	0.9444	
Total # Comparisons	113	77	126	
Source-Specific Probabilities				
	Identification	Inconclusive	Elimination	Total # Comparisons
Same source	0.7717	0.1732	0.0551	127
Different source	0.0794	0.2910	0.6296	189
<i>Experiment Count Data are divided by Source Totals (last column)</i>				
Overall Error Rates				
Opt.	Meaning	Missed Identification	Missed Elimination	Total
2	FTE error	0.0551	0.0794	0.0696
3	Process error	0.2283	0.3704	0.3133
4	Inconcl. = Elim.	0.2283	0.0794	0.1392

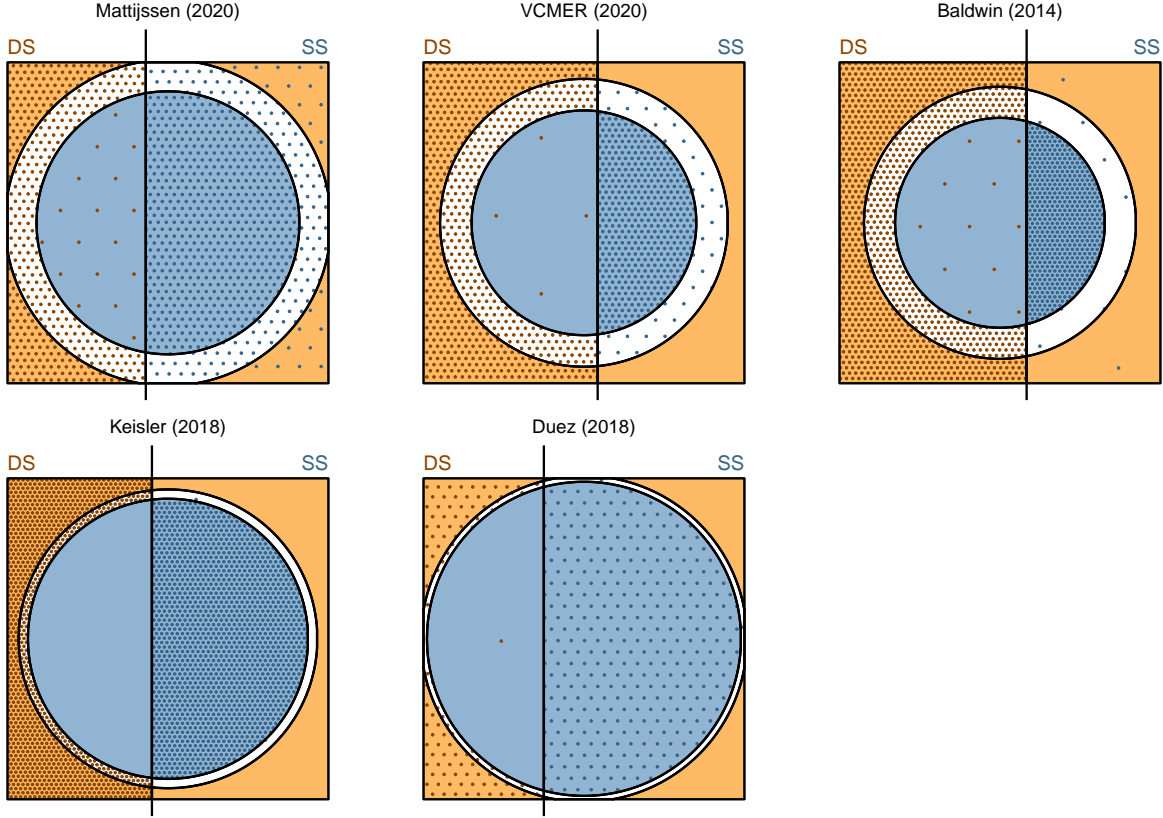


Figure 8: Sketch of the relationship between ground truth of evidence (dots) and examiners' decisions (shaded areas). In [Mattijssen et al. \[2020\]](#), each dot represents approximately 4 examiner evaluations; in [Baldwin et al. \[2014\]](#) each dot represents 2 examiner evaluations.

## B Visual Examination of Experimental Frequencies

Figure 8 shows the visual equivalents of Figure 1, but for these studies, there is generally a very clear difference between the probability of inconclusive decisions under different and same source comparisons.

Bunch (2003) is not included in this display because the proportion of identifications and the proportion of same-source comparisons are both so low that the figure could not be properly rendered.

## C Simulation of comparisons in studies with multiple knowns

### C.1 Bunch simulation

In [Bunch and Murphy \[2003\]](#), there are two sources of unknown information: the composition of the study test kits, and potential examiner use of deductive logic. While it is possible to assess the effect of deductive logic by eliminating redundant comparisons systematically, because the composition of the test kits was random, there is not a single value for the reduction in comparisons due to the use of deduction. As a result, we must first simulate the composition of a test kit, and then evaluate the minimal number of comparisons which must be completed using deductive reasoning.

The process for assembling the test kits provides sufficient information to simulate the composition of the eight test kits used in the study. Using the simulation method, we created 500,000 sets of 8 test kits; from these simulated sets, we identified any which match the reported values of 70 same-source and 290 possible different-source comparisons (287,536). We considered adding the restriction of 45 consecutively manufactured comparisons, but found that this reduced the number of simulations to 6,416; given that we do not assess the consecutively manufactured comparisons separately, adding this restriction was deemed unnecessary. Using these sets, we can then estimate the minimal number of comparisons which are necessary using deductive reasoning in addition to examination. As an independent comparison is one in which the examiner has no relevant prior information about either of the cartridges, the minimal set of comparisons necessary to evaluate all cartridges would also be the set of independent comparisons.

R code to reproduce this simulation (and the deductive reasoning algorithm) can be found at <https://gist.github.com/srvanderplas/9cb0268df99e97a9ce327bb1489f7046>.

Table 19: Results and summary tables for the Bunch study, with only independent pairwise comparisons included. 95% bootstrap intervals are provided for quantities estimated via simulation. To allocate the inconclusives and eliminations, we used the rule stated in [Bunch and Murphy \[2003\]](#): FBI examiners can exclude only on class characteristic mismatches. It should be noted that in none of the 287,536 simulations which matched the other set criteria did we find a set which had more than 109 class characteristic mismatches, suggesting that there may have been some eliminations made on individual characteristics, but as this cannot be confirmed, estimates of eliminations only include class characteristic mismatches.

### A Comprehensive Validity Study for the Forensic Examination of Cartridge Cases *Bunch and Murphy ( 2003)*

Study Type		Test Set		Participants
		# Knowns	# Unknowns	
Open set		variable	variable	8 FBI examiners
<b>Independent Comparisons</b>				
Mean (95% CI)				
	Identification	Inconclusive	Elimination	Source Total
Same source	28.7 [26, 31]	0	0	28.7 [26, 31]
Diff source	0	156.5 [134, 181]	42.5 [22, 64]	199.1 [184, 217]
Conclusion Total	28.7 [26, 31]	156.5 [134, 181]	42.5 [22, 64]	
<b>Independent Comparisons: Conclusion-Specific Probabilities</b>				
	Identification	Inconclusive	Elimination	Source Probability
Same source	1.000	0.000	0.000	0.126 [0.110, 0.142]
Diff source	0.000	1.000	1.000	0.874 [0.858, 0.890]
Conclusion Probability	0.126 [0.110, 0.142]	0.687 [0.597, 0.779]	0.187 [0.095, 0.278]	
<b>Independent Comparisons: Source-Specific Probabilities</b>				
	Identification	Inconclusive	Elimination	Source Probability
Same source	1.000	0.000	0.000	0.126 [0.110, 0.142]
Diff source	0.000	0.787 [0.682, 0.891]	0.214 [0.109, 0.318]	0.874 [0.858, 0.890]
Conclusion Probability	0.126 [0.110, 0.142]	0.687 [0.597, 0.779]	0.187 [0.095, 0.278]	
<b>Independent Comparisons: Error Rates</b>				
Opt.	Meaning	Missed Identification	Missed Elimination	Total
2	FTE error	0.000	0.000	0.000
3	Process error	0.000	0.787 [0.682, 0.891]	0.687 [0.587, 0.779]



Table 20: Number of all possible sets of questioned bullets for Brundage-Hamby sets.

# barrels	one questioned bullet matches	two questioned bullets match	three questioned bullets match	number of possibilities
I	5	5	0	$252 = \binom{10}{5} \binom{5}{5}$
II	6	3	1	840
III	7	1	2	360

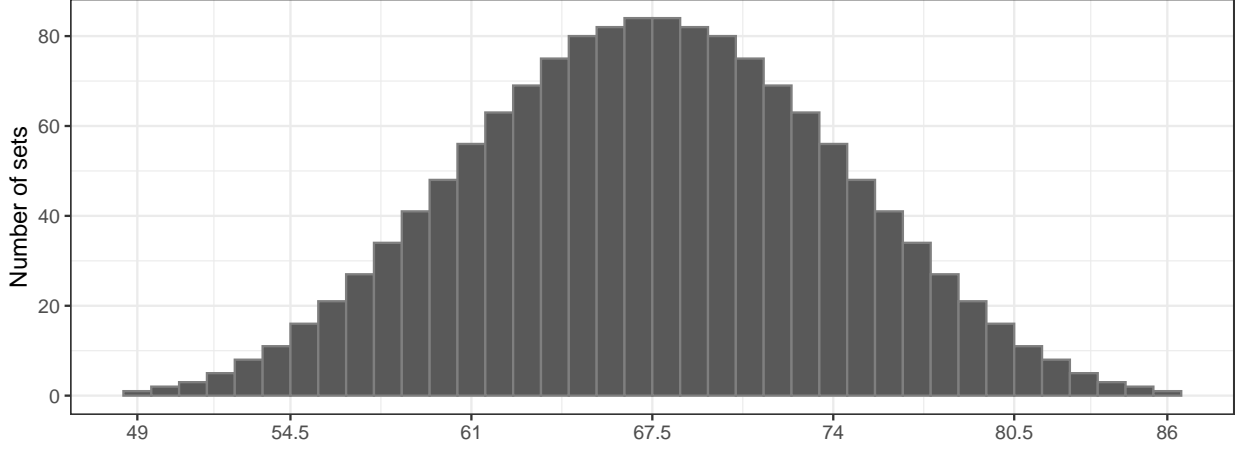


Figure 9: Histogram of the total number of different source comparisons before an identification is made.

## C.2 Enumeration of all possible Brundage-Hamby set comparisons

As discussed in Section 3.2, closed-set studies like Hamby et al. [2019] do not allow us to easily calculate the number of different-source comparisons performed. However, we can estimate the number of comparisons that might need to be performed using logic similar to that employed in Appendix C.1 (but without the variability introduced by the unique experimental design in Bunch and Murphy [2003]).

Brundage-Hamby sets consist of a set of 20 known bullets (two each from the ten consecutively manufactured barrels) and 15 questioned bullets. Brundage [1998] outlines the construction of sets of questioned bullets in detail as follows: ten bullets are chosen, one from each of the ten barrels. The remaining five bullets are picked at random from the barrels, such that at most three questioned bullets are from the same barrel.

Using this strategy, a total of 1452 different sets can be constructed. These sets have questioned bullets of three main forms, listed in Table 20 from the perspective of the number of barrels with one, two, or three matching questioned bullets.

Figure 9 shows the histogram of the total number of different source comparisons before and identification is made.