Planning, design and logistics of a decision analysis study: The FBI/Ames study involving forensic firearms examiners

Keith L. Monson, Ph.D. [a,*], Erich D. Smith, M.S. [a], Stanley J. Bajic, Ph.D. [b]

[a] *FBI Laboratory, Quantico, VA, 22135, USA*
[b] *Ames Laboratory, Ames, IA, 50011, USA*

ARTICLE INFO

ABSTRACT

This paper describes design and logistical aspects of a decision analysis study to assess the performance of qualified firearms examiners working in accredited laboratories in the United States in terms of accuracy (error rate), repeatability, and reproducibility of decisions involving comparisons of fired bullets and cartridge cases. The purpose of the study was to validate current practice of the forensic discipline of firearms/toolmarks (F/T) examination. It elicited error rate data by counting the number of false positive and false negative conclusions. Preceded by the experimental design, decisions, and logistics described herein, testing was ultimately administered 173 qualified, practicing F/T examiners in public and private crime laboratories. The first round of testing evaluated accuracy, while two subsequent rounds evaluated repeatability and reproducibility of examiner conclusions. This project expands on previous studies by involving many F/T examiners in challenging comparisons and by executing the study in the recommended double-blind format.

1. Background

1.1. Forensic examination of firearms/toolmarks (F/T) evidence

This paper details the planning, design, and administration of a validation study of determinations that F/T examiners make of whether an evidentiary bullet or cartridge case was fired by a specific firearm. Terminology, manufacturing processes, and methods in the forensic examination of firearms evidence have been extensively covered elsewhere [1–9]. Briefly, the firearm in question is fired under laboratory conditions into a box of cotton batting or a water tank to capture the bullets and the recovered (K)nown (reference) bullets and/or cartridge cases are compared to the (Q)uestioned evidence specimens. Utilizing a comparison microscope, pairs of K and Q bullets or cartridge cases are mounted on the separate stages for microscopic comparison. This instrument consists of two microscopes conjoined by an optical bridge that allows a split-screen view, simultaneously displaying K and Q specimens, as the operator manipulates the position of the specimens to bring the microscopic marks into alignment. Informed by his or her training and experience about the nature of these microscopic marks, the examiner draws a conclusion about the extent of correspondence.

The AFTE Theory of Identification, first adopted in 1992 by the

Association of Firearm and Tool Mark Examiners [10–12], is the foundation for reaching and reporting one of four possible conclusions when comparing toolmarks: identification, inconclusive, elimination, or unsuitable for comparison, as described by the AFTE Range of Conclusions (section 2.10). The Theory allows for an opinion of common origin (identification) when the surface contours of two toolmarks are in "sufficient agreement." Sufficient agreement is decided when the level of microscopic agreement is similar to the microscopic agreement seen from specimens known to have originated from the same source and exceeds microscopic agreement occurring between the Best-Known Non-Match (worst case scenario). The analysis of F/T evidence involves two levels of analysis. Level 1 analysis is the objective examination of evidence for class characteristics. Class characteristics are features of design or manufacture that are common to a class of items, e. g., caliber, number of lands and grooves, twist direction of rifling, or machining marks that are present throughout some or all of a production run. At Level 1 analysis, identification cannot occur, but an elimination can occur if there is a difference in observed class characteristics between two specimens. If, after the Level 1 analysis there is no discernible difference observed in class characteristics, the examination proceeds to Level 2 analysis. Level 2 analysis is the subjective microscopic comparison of individual characteristics, the randomly produced marks on

---

Table 1
Reported accuracy of F/T examiners in selected studies.

Author(s)	No.	Manufacturer	Seq.	Examiners	Design	Comparisons		Error rate (FP/FN), percent	
						Cartr. Cases	Bullets	Cartr. Cases	Bullets
Brundage, 1998	10	Ruger	C	30	SB, C		1020		0/0
Bunch & Murphy, 2003	10	Glock	C	8	DB, O	360		0/0	
E. Smith, 2005	9	Ruger	R	8	SB, C	360	360	0/0	0/0
Fadul, 2011	10	Glock	C	183	SB, C		2745		0.4[a]/na
Fadul et al., 2013	10	Ruger	C	217	SB, C	3825		0.06[a]/na	
Fadul et al., 2013	10	Glock	C	165	SB, P	1650			1.2[a]/na
Cazes & Goudeau, 2013	5	HiPoint	C	69	SB, C	552		0/0	
Baldwin et al., 2014	25	Ruger	R	218	SB, O	3270		0.94[b]/0.37[b]	
Stroman, 2014	3	Smith & Wesson	R	25	DB, C	75		0/0	
Kerkhoff et al., 2015	10	Various	R	11	DB, O	341	55	0/0	0/0
Hamby et al., 2016	1632	Glock	R	1	NB, C	13,30,896		0/na	
T. Smith et al., 2016	8	Various	R	31	DB, O	2046	2046	0.14/0.43	0/0.11
Keisler et al., 2018	9	Various	R	126	SB, C	2520		0/0	
Kerkhoff et al., 2018	1	Sig Sauer	R	10	DB, O	344		0/0	
	39	Glock	R						
Hamby et al., 2019	10	Ruger	C	697	SB, C	10,455			0.05[b]/na
J. Smith, 2020	35	Various	C	74	SB, O		7420		0.08/0.16
Law & Morris, 2021	20	Various	R	17	SB, O	340		0.28/0	
This Study, 2021	23	Beretta	C	173	DB, O	10,110	10,020	0.93[b]/1.87[b]	0.66[b]/2.87[b]
	4	Beretta	R						
	10	Ruger	C						
	10	Jimenez	C						

Seq: C/R (consecutive/Random manufacturing sequence).
Design: SB/DB/NB (Single/Double/No) blinding; C/O/P (Closed/Open/Partly Open) set design.
[a] False positive discovery rate.
[b] Maximum likelihood estimate.

the surface of the evidence specimens (i.e., bullets and cartridge cases) that arise during manufacture, firing, or cycling through the action of a firearm. Level 2 analysis involves side-by-side comparison to determine if there is sufficient correspondence in the microscopic marks of value to conclude that they were produced by the same source, i.e., identification. If, during the Level 2 analysis, the comparison of the microscopic marks indicates insufficient correspondence of the microscopic marks of value, then the decision is either inconclusive or an elimination. (The option to assert elimination based on individual characteristics is a matter of individual laboratory policy.)

In the forensic context, false positive conclusions (mistaken identifications) are generally regarded as far more significant than false negative ones. This is in accord with the presumption of innocence in adversarial legal systems and in common law dating to Roman times, as opined by the 1895 U.S. Supreme Court opinion that, "it is better to let the crime of a guilty person go unpunished than to condemn the innocent [false positive]" [[13], p. 454]. More recently, the *Mitchell* court declared that [[14], ¶240],

> … in the courtroom the rate of false negatives is immaterial to the *Daubert* admissibility of latent fingerprint identification offered to prove positive identification because it is not probative of the reliability of the *testimony for the purpose for which it is offered* (i.e., for its ability to effect a positive identification). Moreover, evidence of the false negative rate is often equivocal. While it might suggest a generally error-prone method, it is equally consistent with a very conservative method with a low false positive error rate.

1.2. Genesis and context for the study

In 2012, FBI researchers began to design a decision analysis study that would assess the performance of F/T examiners. It was motivated by several factors: recent challenges to the admissibility of F/T testimony based on *Daubert* [15], recommendations cited in the 2009 Report by the National Research Council (NRC) of the National Academy of Science, *Strengthening Forensic Science in the United States: A Path Forward* [16], criticisms of previous F/T reliability studies, and the effectiveness

with which challenges were addressed by a decision analysis study of latent fingerprint examiners [17,18].

Expert opinion based on the examination of bullets and cartridges to make a determination of whether they were fired from (or cycled through the action of) a particular gun has been accepted by the courts for decades [19]. Notwithstanding long-standing practice and judicial precedent for acceptance of F/T testimony, the foundations of the discipline have sometimes been criticized in recent years, particularly due to lack of reliable error rates, but F/T testimony is consistently found admissible [20–35]. A 2008 NRC Report asserted that, "the validity of the fundamental assumptions of uniqueness and reproducibility of firearm-related toolmarks has not yet been fully demonstrated" [1, p, 3]. Accordingly, a subsequent NRC committee recommended that the forensic sciences develop research "to address the accuracy, reliability, and validity in the forensic science disciplines" [16, p. 22]. This recommendation was later echoed in a 2016 PCAST report calling for "foundational validity" [36, p. 43], i.e., that a "forensic-science method requires that it be shown, based on empirical studies, to be repeatable, reproducible, and accurate, at levels that have been measured and are appropriate to the intended application" [36, p. 4] and charged "… that firearms analysis currently falls short of the criteria for foundational validity, because there is only a single appropriately designed study to measure validity and estimate reliability. The scientific criteria for foundational validity require more than one such study, to demonstrate reproducibility." [36, p. 112]. The PCAST report further emphasized the necessity for decision analysis ("black box") studies to validate subjective feature-comparison methods such as F/T analysis [36, pp. 49, 112]. A recent U.S. Department of Justice statement took issue with several assertions of the PCAST report, including that black box studies should not be taken as the *only* validation method of feature-comparison methods [37]. Black box studies, which consider only the input (evidence) and output (conclusion) without considering intervening cognitive (or other) processes, are further discussed in section 2.6.

Several compilations summarizing research studies on various aspects of the reliability of the F/T discipline are available [12,38–40]. Numerous studies have been published documenting the ability of F/T examiners to correctly identify breech face marks after repetitive firings

of the same firearm [2,41–48] and to differentiate and identify those produced from consecutively manufactured slides or breech bolts [2, 49–62]. Other studies have investigated the ability of F/T examiners to correctly identify bullets fired from the same barrel [41,46,63–67] and to distinguish those fired from consecutively manufactured barrels [2, 40,49,68–72]. Consecutively manufactured items are chosen for study because they are universally acknowledged to present the greatest challenge to distinguish due to their similarity in individual characteristics and likelihood of exhibiting subclass characteristics. Subclass characteristics are "features that may be produced during manufacture that are consistent among items fabricated by the same tool in the same approximate state of wear. These features are not determined prior to manufacture … " [73]. Ability of F/T examiners to identify cartridge cases in the absence of a firearm or supporting information has also been demonstrated [74–77], as has their ability to identify bullets and cartridge cases despite the presence of subclass characteristics [40,47,55, 57,68,75,78–86]. The elements and reported findings of selected studies assessing the accuracy of F/T examiners are summarized in Table 1, compiled from Refs. [53,56,62,68,70,71,74–77,87–92]. Results are broadly comparable, despite differences in study designs.

Repeatability of conclusions (the ability to obtain the same result when the same specimens are presented later in "blind" fashion to the same examiner) and reproducibility (where the same specimens are presented at a later date to another examiner) have seldom been tested in any of the forensic disciplines. Three studies of latent fingerprint examiner performance are exceptions [18,93,94]. Although proficiency tests provide periodic evaluation of inter-laboratory reproducibility, they are limited in scope and typically do not assess individual competency or repeatability [95]. A study designed to test the effect of bias due to suggestive case-specific information (there was no evidence of such) indirectly addressed the topics of repeatability and reproducibility in the F/T discipline [96]. When the same examiner performed second comparisons of six resubmitted bullets at a later date (repeatability), and when a second examiner performed independent comparisons from 153 previously examined case work files (reproducibility), there were numerous differences in reported conclusions in terms of identifications vs. inconclusives (but not for identifications vs. eliminations). A related study involving 8 examiners and 97 blind, peer verification reviews of one another, where the second examiner reviewed bullet comparison photos from the first examiner, reported only two disagreements in evidential strength, but none in terms of support for same vs. different source conclusions [97]. A round-robin proficiency test involving 64 laboratories who compared polymer casts of the same five bullets and five cartridge cases to two reference specimens provided for each unknown (also polymer casts) reported 4% false identifications and 2% false eliminations [98]. Another recent study used polymeric double casts of cartridge cases to assess reproducibility and degree of consensus in conclusions reached by 18 practicing examiners [92]. The examiners received 20 comparison sets containing sets of 1 Q and 3 K replicate specimen, of which 13 sets were ground truth exclusions and 7 sets were identifications. One examiner withdrew from the study after making 5 false positive conclusions. Among the remaining 17 examiners, for the 7 ground truth identifications there were no false exclusions, all examiners correctly declared identifications for 4 sets, and 3 sets were deemed identification by some examiners but inconclusive by others. None of the 13 ground truth exclusions elicited the same conclusion from all 17 examiners; most of the variability was in level of inconclusive, although there was one false positive identification.

2. Prospective planning: overall design principles

2.1. Definition of scope

Scope defines the boundaries of what a study will or will not include, how comprehensive it will be, who will do it, and how long it will likely take. For this study, the requirement was to develop and execute an extensive validation study of the reliability of forensic F/T source conclusions. The project assessed the performance of numerous qualified firearms examiners working in accredited laboratories in the United States in terms of overall accuracy (error rate), repeatability, and reproducibility of decisions involving forensic comparisons of simulated firearms evidence (bullets and cartridge cases). Examiner error rate was estimated by counting both the number of false positive and false negative conclusions. Deliverables included several peer-reviewed publications describing the results.

2.2. Hypotheses

This project evaluated whether examiners perform their examinations with a high degree of accuracy. It was designed to validate the discipline of F/T examination of firearms evidence by examining the following hypotheses:

H1. Toolmarks reproduced on fired cartridge cases by the same slide/chamber are identifiable as coming from a common source by qualified F/T examiners, while toolmarks produced by different slides/chambers are not judged as coming from a common source.

H2. Toolmarks reproduced on fired bullets by the same barrel are identifiable as coming from a common source by qualified F/T examiners, while toolmarks on bullets fired by different barrels are not judged as coming from a common source.

2.3. Experimental plan

A written detailed plan should precede any but the most facile experimental work. Unplanned modification of the design based on initial results after experimentation has begun, especially for a validation study, should be avoided if possible, so as not to compromise the validity of the study [36, p. 52 [99], p. 9]. A pilot study is advisable to understand the effectiveness of test distribution and practicality of test specimens which will help to preclude later plan changes (section 4.4). Additionally, the plan serves both to guide the research and to justify the time and resources that will be required. Its level of detail depends on the scope of the proposed study. If some portion of the project will be performed under contract, the experimental plan will help to define the contract's statement of work and its context within the overall project. In general, a plan will include:

- Statement of the study's goals
- Purpose of the study (e.g., to solve a problem, advance the science, address legal challenges)
- Brief summaries of esoteric topics to assist the non-specialist reader (e.g., the funding entity)
- Previous work that is related, or upon which the proposed study will depend
- Detailed research approach (methods, number/types of specimens, logistics, responsibilities of participants, and data analysis techniques anticipated)
- Resources required (equipment, personnel, time, funding). A Gantt chart is often helpful to estimate scheduling and reveal dependencies of various activities for a project.

Key design features of this study included:

1. Evaluation of the accuracy, repeatability, and reproducibility of F/T examiners' decisions regarding common source. Can a qualified F/T reach the proper decision for common source when applying the AFTE Theory for Identification?
2. Study participants shall be anonymized, qualified F/T examiners who are AFTE members working in accredited laboratories
3. Firearms from three different manufacturers and multiple examples of the chosen models, including those that are deemed

relatively difficult to compare. Firearms and ammunition will be selected that tend to produce limited microscopic marks for comparison and no aperture shear, but present subclass characteristics.

4. Groups of consecutively manufactured slides and barrels that are collected at intervals throughout the manufacturing life of the single tool used to cut/shape them, to produce highly similar but individual (non-matching) specimens (best known non-match) and maximizing the potential for subclass similarity
5. Additional comparison slides and barrels from different production runs (known non-match)
6. Extensive firing of each firearm (~500x) to test effects of firing sequence on the reproduction and longevity of individual characteristics, thereby affecting examiner accuracy
7. Preparation and distribution of test packets and use of double-blind conditions to conduct comparisons
8. An open set design, i.e., there may not necessarily be a match for every Q specimen
9. To increase the relative test difficulty, an overall proportion of true matches of approximately 33%, but variable among test packets
10. All items in an individual comparison set shall be fired from the same make and model firearm, precluding elimination based on class characteristics
11. A break-in period of firing new firearms to normalize marks they produce [100].
12. A comparison set consisting of a single Q to be compared to two K specimens, the latter being fired from the same firearm. Providing multiple K specimens minimizes the possibility that a single K did not replicate a toolmark [87].
13. Each set represents an independent comparison unrelated to any other set in the test.
14. Survey of participants, to include laboratory accreditation, personal certification, years of experience, equipment used in comparison, and laboratory policies on inconclusive and exclusionary decisions
15. Discourage collaboration or verification of decisions by a second examiner
16. Preclude sharing of results on individual packets and the possibility that participants may infer test design by coding/relabeling the contents before their submission to another examiner or resubmission to the same examiner
17. Pilot testing to evaluate study design

Testing was administered in three rounds to 173 qualified, practicing F/T examiners employed by federal, state, or local crime laboratories. The first round evaluated accuracy and reliability. The second and third rounds, administered to a subset of the initial examiners due to attrition, evaluated repeatability and reproducibility of their conclusions. The second and third rounds could be conducted sequentially or concurrently with the first. As recommended by PCAST [36], the project was conducted in double-blind format (section 2.7) as a joint effort involving both Government and Contractor personnel, each with defined responsibilities. In general, the Contractor developed sampling and specimen marking plans; developed specimen distribution plans for each round; developed a statistical analysis plan; labeled, assembled, and distributed test packets to participants; and tabulated, analyzed, and published results jointly with the Government. Fired bullets and cartridge cases were the test materials; approximately 28,250 of each item had to be labeled, distributed, and accounted for. The Government provided all firearms and ammunition for testing; performed all live firing; furnished fired bullets and cartridge cases to Contractor for distribution to test participants.

| | Ground Truth | |
Examiner Decision	Matching (KM)	Non-Matching (KNM)
Identification (ID)	A	B
Inconclusive	C	D
Elimination (Elim)	E	F

False positive error rate	$= \dfrac{Incorrect\ ID}{all\ KNM}$	$= \dfrac{B}{B+D+F}$
False negative error rate	$= \dfrac{Incorrect\ Elim}{all\ KM}$	$= \dfrac{E}{A+C+E}$
Sensitivity	$= \dfrac{Correct\ ID}{all\ KM}$	$= \dfrac{A}{A+C+E}$
Specificity	$= \dfrac{Correct\ Elim}{all\ KNM}$	$= \dfrac{F}{B+D+F}$

Fig. 1. Formulae to describe examiner identification outcomes.

2.4. Experimental variables

Prior to pilot testing, the following potential independent variables were identified:

- firearm manufacturer (2 levels, later modified, section 4.1)
- cartridge case material (2 levels, hard steel and soft brass; this variable later removed, section 4.3)
- primer material (2 levels, hard and soft; this variable later removed, section 4.3)
- stage in manufacture relative to tool lifetime (3 levels)
- shot sequence (semi-ordinal, in groups of 50)
- known match status (2 levels, yes/no)
- time examiner spent on each comparison
- responses to several survey questions (laboratory accreditation, personal certification, years of experience, equipment used in comparison, and laboratory policy on inconclusive and exclusionary decisions)

The dependent variable is the examiner's decision:

- identification, inconclusive, elimination, or unsuitable (section 2.10)

Testing of examiner performance on comparisons of bullet and cartridge case are essentially two independent studies. Thus, firearm components (slides and barrels) were not included in the list of variables, as cartridge cases and bullets will never be presented to examiners for intercomparison. Note that a comparison set always contained items produced by a firearm of the same make and model (although not revealed to the examiners).

2.5. Performance metrics

Accuracy is the ability of examiners to arrive at the correct conclusion, often expressed in negative terms as error rate. In this study, inaccuracy is measured by the number of false positive conclusions on known non-matches and the number of false negatives on known matches. Other methods for calculating error exist (discussed later in this section and in 2.10) and the "error rate" data obtained from this study may not be directly comparable to other scenarios, depending on the methods used. *Reproducibility* is the extent of agreement when multiple examiners independently compare the same specimens, while *repeatability* reflects how consistently an individual examiner reaches

the same conclusion when unknowingly re-comparing the same specimens [18]. If individual error rates are highly variable, appropriate data analysis will require a hierarchical model in which error rates can vary by examiner. A previous study used a beta-binomial model as the basis for statistical inference for this situation, in which a mixture of binomial distributions (each of which is appropriate for an individual examiner) follows a beta distribution (representing the total variation among examiners) [87,101].

Additional performance metrics that can be calculated from the results include *sensitivity* (the ability to detect identifications when they exist, i.e., the fraction of identifications declared within the number of known matches (KM) present in a test, or "correct identification rate" [102]) and *specificity* (the fraction of eliminations detected within the number of known non-matches (KNM) present, i.e., "correct elimination rate" [102]). Calculation of these and several other metrics is illustrated in Fig. 1. The PCAST report underlines the importance of examining the true positive rate (sensitivity) in conjunction with the false positive rate (specificity) [36, p. 119]. But sensitivity and specificity are counterbalanced; one will dominate relative to the other depending on the weight allotted to false positives vs. false negatives [103,104].

Although some might propose an inconclusive decision as an unsuccessful outcome, or failure ("error") to identify a KM [105,106], such a decision rightly represents judgment that the comparison presents insufficient information (quality and/or quantity of individual characteristics) for a definitive statement (minimization of false positive being paramount, see section 1.1) [102,107–109].

The intuitive false positive discovery rate (the fraction of incorrect identifications among all identifications declared, FPDR) and the corresponding false negative discovery rate (FNDR) exclude inconclusive decisions in their calculation [93]. The fundamental shortcoming of these metrics is their dependence on the number of KMs available (prevalence) in the particular study designed to measure error [17,104, 110–112]. In a research study that contains (or a case working laboratory that tends to receive) a higher proportion of ground truth identifications, false negatives will be more common (leading to a higher FNDR and lower FPDR), and vice versa. Consequently, these metrics are not comparable across studies except for situations in which all studies being compared have, at least approximately, the same ratio of numbers of true-match evaluations to true-nonmatch evaluations. Furthermore, unlike the proportions of false positive and false negative error rates, they are statistically biased. This is because the denominators of these ratios are random, i.e., based on counts observed in the study, rather than fixed by study design. Further, the bias of these statistics is dependent on the sample size of the study. Hence, even their use in comparisons of results from studies with the same ratio of true-match to true-nonmatch comparisons is questionable unless the sample sizes of the studies are comparable. (N.B., positive and negative predictive values are the complement of positive and negative false discovery rates: PPV = 1 - FPDR and NPV = 1 – FNDR.)

Koehler provides a synoptic exposition on assessing the reliability of expert testimony and discusses many of these metrics and the pertinence and limitations of each [110]. Regardless of the metric used, a reported error rate must be considered within the context of the study that produced it. Among other things, error estimates are dependent on study design, test materials, difficulty of comparisons involved, and the participants involved [37,113]. Error rate estimates from such studies "… may give insights into forensic domains in general, but may say very little about a specific examiner's decision in a particular case" [113, p. 4]. Thus, validity of methods and techniques (referred to as foundational, base rate or aggregate validity) must be distinguished from validity as applied in a particular case [8,17,29,36,113–116].

2.6. Black box

As defined by the Organization of Scientific Area Committees for Forensic Science (OSAC) [117],

A black box study assesses the accuracy of examiners' conclusions without considering how the conclusions were reached. The examiner is treated as a 'black-box' and the researcher measures how the output of the 'black-box' (examiner's conclusion) varies depending on the input (the test specimens presented for analysis). To test examiner accuracy, the "ground truth" regarding the type or source of the test specimens must be known with certainty.

Being agnostic to process, a black box study acknowledges the discipline's "lack of a precisely defined process" [16, p. [155]] by focusing solely on results to address accuracy, reliability, and validity of the discipline. A black box approach is particularly suitable to address these concerns as they apply to F/T analysis, because toolmarks change over time and with use [42,46,47,118] due to tool wear and variations in machining conditions during manufacture of successive items [119], wear or buildup of residues during use (fouling), slight changes in physical placement of cartridges due to mechanical tolerances of the chamber and firing pin [120], and incomplete obturation (swelling) for bullets and cartridge cases, which may limit toolmark production.

2.7. Double-blind

Double-blind tests are the gold standard in medical research, where neither doctor nor patient knows whether a new drug or treatment was administered [121]. Ideally, the participants are unaware they are even being tested, but since this condition is difficult to attain in a decision analysis study involving case-working forensic examiners, such studies are sometimes qualified as being "declared" [75,89,112]. In order to guarantee anonymity, to comply with requirements of our institutional review board (section 3.1), and to obtain signed informed consent, the examiners were of necessity made aware that they were participating in a research study. It has been proposed to discontinue use of the designation "double blinding" and related adjectival qualifiers, as they are often ambiguous, uninformative, and misleading, instead simply to label a study as unblinded or blinded, accompanied in the latter case by statements or a table delineating what, how, and to whom information was blinded [122]. Conduct of a double-blind study requires the participation of a third party for labeling, selection, distribution, and record-keeping of test specimens. It satisfies the PCAST recommendations that [36, p. 17],

… the FBI Laboratory should assist in the design and execution of additional empirical "black-box" studies for subjective methods, including for latent fingerprint analysis and firearms analysis [that] … should be conducted by or in conjunction with independent third parties with no stake in the outcome.

Double-blind testing was accomplished by contracting with Ames Laboratory to handle the logistics of specimen coding and labeling; test packet assembly, mailing, and tracking; direct communication with the participants; and calculating results and statistics. To avoid possible bias, no FBI F/T examiners participated in the main study. They did so in the pilot study (section 4.4) but pilot results were not used in the actual study.

2.8. Open set

An open set design, where there may not necessarily be a match for every Q specimen, addresses criticisms that some previous studies with a closed set could possibly underestimate the false positive rate. The PCAST report disparaged prevalent "set-based" studies, in which all possible pairs of specimens in a comparison set are compared (similar to case work examinations), as tending to underestimate the rate of false positives. In a closed set, "examiners can perform perfectly if they simply match each bullet to the standard that is closest" (unlike in case work) [36, pp. 106–109]. An open set design avoids underestimation of false positives inherent in a closed set but may increase the number of

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 for examination.

Fig. 2. AFTE range of conclusions [131].

inconclusive decisions. An open set design was implemented, supplemented further by varying the ratio KM/KNM from one comparison set to another.

2.9. Challenging comparisons

A stringent test of examiner abilities was desired. Yet, arriving at the perfect balance between toolmarks that can be too readily placed into correspondence and identified, and those that are so difficult that many examiners would declare them to be inconclusive, is challenging. An overwhelming number of inconclusive calls on very difficult comparisons, although somewhat informative, would not provide a useful estimate of accuracy. Comparisons were made moderately difficult by design to increase the likelihood of making both false positive and false negative errors [123]. Groups of several consecutively manufactured slides and barrels from different manufacturers (section 4.1) were sampled at intervals throughout the life of the single tool used to cut them, to produce highly similar but non-matching toolmark patterns in the specimens while maximizing the potential for subclass similarity.

2.10. Range of conclusions

In 1992, the AFTE adopted the Range of Conclusions (Fig. 2) with the option of three levels of inconclusive [124]. A draft OSAC document further elaborates the Range [125]. The utility and import of the Range, its relation to gradations of conclusions used in other forensic disciplines, and whether inconclusive conclusions should figure into calculation of error rate, or be used to offer a presumptive link to investigators have long been a matter of discussion in various OSAC committees and elsewhere [29,34,92,106,110,113,126–130]. Some go so far as to opine that an inconclusive decision that could, if expressed differently, provide exclusionary power, is as consequential as a false identification [129, p. 198]. In the present study, inconclusive conclusions will be recorded and analyzed, but shall not be considered errors.

All U.S. examiners generally conform to the current AFTE Range [131]. The extent to which an inconclusive conclusion is invoked and elaborated and whether elimination can be concluded from individual characteristics or when only class characteristics are in agreement, are matters of training, personal judgement, and laboratory policy. In some laboratory systems, inability to match individual characteristics is declared inconclusive by policy, while in others, elimination remains an option. Regardless of variations in individual or laboratory-imposed adaptations of the AFTE Range (e.g., a single choice for inconclusive, or a binary choice of identification or elimination), examiners were instructed to use the AFTE Range in this study. Doing so offers an experimental inquiry into expected consistency should the five-level Range become a standard and avoids complications of intercomparing results from examiners who use different scales.

In the previous Ames study, expression of a conclusion that there was a lack of sufficient corresponding microscopic detail was handled inconsistently: 44% of examiners judged these eliminations, 21% called them inconclusive, and 35% used a mixture of the terms [87].

2.11. Risks

Another component of planning involves anticipation of potential risks to cost, schedule, and/or successful conduct of the project and possible mitigations. There is schedule risk due to participation of examiners who collaborate at the discretion of their employing agency. Participants also may be tardy in returning their responses due to case work demands. The possibility of some degree of selection bias is nearly unavoidable in any study that is difficult, time consuming, and requires the participation of active professionals such as case-working forensic examiners. Federal statutes require informed consent by human subjects as a condition for their participation in research [132,133]; see also section 3.1. Therefore, all such research participation must be inherently and intrinsically voluntary (i.e., self-selected). Extent of participation often decreases in ensuing rounds, obliging consideration of possible attrition bias [134–136]. Since participants will be aware that they are participating in a research study, they may be either more or less diligent compared to conducting case work. Prohibition of collaboration or verification of decisions by a second examiner may inflate error counts that would otherwise be resolved via laboratory quality control. As the plan emphasizes challenging comparisons, it may lead to a preponderance of inconclusive or unsuitable decisions. Use of three levels of inconclusive (section 2.10) may lead examiners to choose inconclusive A or C [131] rather than commit to identification or elimination, reducing statistical power of estimated error rate. This tendency was recently demonstrated when latent fingerprint examiners used a similar 5-point scale rather than the prevalent 3-point scale [128]. Examiners working in an agency that permits only a single level of inconclusive may apply three levels inconsistently. Some specimens may be unsuitable for examination, as prescreening is unfeasible, but potential unsuitability of specimens is mitigated by providing a pair of K specimens. Two K specimens may not fully represent the stochastic variability in firing from the same firearm [137], but practical considerations preclude providing additional specimens in a large study. Neither the firearms nor barrel casts will be provided, which might limit the ability of examiners to diagnose subclass characteristics. This constraint is mitigated by studies showing little effect on identification performance when no gun

Fig. 3. Firearms selected for this study: a) Beretta M9-A3-FDE, b) Ruger SR-9c, c) Jimenez JA-Nine (not necessarily to scale).

is available [70,74,75,87]. Laboratory-specific latitude in declaring eliminations based on microscopic marks, although sanctioned by the discipline [131,138], complicates data interpretation in error rate studies. The survey design will aid interpretation.

Funding

The project required significant funding, the predominant expense being for Ames Laboratory contract services. Securing the requisite funding for such a large endeavor involves presentation of the topics herein to varying levels of detail. Presentations to executive management and contract administration staff demand concise focus on business considerations (e.g., need, benefit, timeline, staffing, and cost, with an overview of the experimental aspects), while a grant proposal would also require lengthy exposition of the technical approach. Although ultimately supported solely by FBI internal research resources, the FBI investigators actively pursued, but did not receive, external funding options from other government entities and independent research foundations.

3. Addressing human factors related to conduct of the study

3.1. Institutional review board (IRB) and informed consent

Any experimentation involving humans as subjects or participants must, by U.S. law, do so under a signed agreement of informed consent [132,133]. The researchers who administer the study must submit that agreement and other documents to the IRB of their organization for review and approval before any experimentation with human subjects begins. The respective IRBs of both the FBI and Ames Laboratory contracting agency, Iowa State University, approved all aspects of this study. After accepting an invitation to participate (appendix 1, Invitation letter), participants were asked to provide informed consent via a form approved by both FBI and Ames Laboratory IRBs (appendix 2, Informed consent form). The primary consideration was to protect the participants from risk to their professional standing and reputation by making all results anonymous.

3.2. Survey of examiner-specific factors

Individual surveys of participants provided data to explore whether any of these factors were related to the reported results. Factors queried include laboratory accreditation, personal certification, years of experience, equipment used for comparisons, and laboratory policies on inconclusive and exclusionary decisions. The one-time survey form provided to each examiner is shown in appendix 3, Survey form for examiner-specific factors.

3.3. Instructions for performing comparisons and reporting results

Participants were given detailed instructions explaining the specimens provided, how to document comparisons, and return shipping procedures (appendix 4, Instructions for performing comparisons and reporting results). When an identification conclusion was rendered for bullets, they were asked to mark the land or groove impression principally used to arrive at the decision (example in appendix 4). After return of the specimens to Ames Laboratory, the markings on incorrectly identified bullets and casings were photographed. Doing so provided opportunity for the FBI team to retrospectively analyze erroneous decisions. All markings were removed from bullet and cartridge case specimens prior to their reuse in subsequent rounds.

Examiners recorded their conclusions for every comparison packet (appendix 5, Reporting form):

- Examiner ID
- IDs of either:
  o Cartridge case (coded for associated slide number and shot sequence) or
  o Bullet (coded for barrel number and shot sequence)
- Comparison decision
- Quality/quantity of marks and relative difficulty of comparison
- Time spent conducting the comparison
- Whether the consecutive matching striae (CMS) method was used

Some factors that may affect the results can only be estimated indirectly, others not at all. Recording the time to carry out each comparison offers a crude estimate of the degree of examiner diligence, albeit conflated with difficulty, conflicting demands of case work, and individual speed of decision-making. Propensity to declare a definitive conclusion or to defer to an inconclusive statement on challenging comparisons defies characterization. Although previous studies showed negligible effects due to microscope and lighting, individual certification, years of experience, or whether decisions were supported by use of the CMS method [56,60,71,76,87], these readily available data were also captured. However, since the number of errors was anticipated to be small, it is unlikely that the contribution of many of these effects can be fully characterized.

4. Specifics of experimental design: choices and rationale

Part of defining the scope of a study involves deciding how comprehensive it will be, particularly how many instances or conditions of each independent variable will be involved or tested. In designing a study, a researcher faces many preliminary decisions and inevitable compromises. Except to answer the simplest of questions, once the overall goal has been defined, the number of potential factors tested (independent variables) must be restricted so that the final result has adequate statistical power. The number of specimens and/or examiners that should be tested increases prodigiously as more independent variables are introduced, as the magnitude of the effect being studied becomes smaller, and as the desired confidence in the result increases (smaller standard deviation, confidence limits, or p-value). To detect a

small or rare effect (examiner errors in the present case) one must make a large number of observations within a representative population.

Limitations are also imposed by constraints such as funding, available personnel resources, how much time will be required to complete the study, and the anticipated degree of participation by test subjects. Post-experiment statistical analysis may reveal that some "independent" variables are highly related or have little effect on the dependent variable and may be disregarded as contributing minimal additional information.

4.1. Choice of pistols

The study included fired Beretta bullets and cartridge cases, Jimenez cartridge cases, and Ruger bullets. They were chosen to represent firearms commonly encountered in case work, with details of manufacture that present particular characteristics or challenges for comparison. Multiple pistols of each brand were used to avoid potential correlation effects from the same firearm being used in multiple sets for a single examiner.

To produce cartridge cases for comparison, Beretta M9A3-FDE (Flat Dark Earth surface finish) and Jimenez JA-Nine pistols were chosen (Fig. 3). Two previous black box studies used Ruger-fired cartridge cases as test specimens [60,87], but we wished to preclude examination of the strong aperture shear striations that Ruger pistols produce. Ruger pistols are short recoil firearms that utilize a tilting barrel to unlock the chamber. This method for unlocking produces a prominent, independent, and readily identifiable toolmark feature called aperture shear as well as firing pin drag [58,90,139]. These features are created when the firearm action begins to unlock, tilting the chamber/barrel in a downward motion while the primer is still within the firing pin aperture. The resulting operation causes the primer to scrape across the firing pin aperture and the firing pin to drag across the firing pin impression. Such a system produces two distinct striated toolmark features independent of the breech face.

To preclude comparison using strong aperture shear striations on cartridge cases, this study employed pistols with simple blowback action (Jimenez) and locking block action (Beretta) that, by their design and operation, do not produce aperture shear/firing pin drag. Since the Beretta and Jimenez pistols do not use a tilting barrel recoil mechanism, they do not produce striated aperture shear marks, so cartridge case comparisons are primarily confined to breech face marks and firing pin impressions, perhaps supplemented by extractor, ejector, chamber, or magazine marks. The Beretta breech face is aperture-formed through a stamping process to produce a concave edge. This reduces primer surface area with which the breech face can make contact, limiting the extent of toolmark reproduction. Jimenez pistols represent a low-cost firearm type commonly encountered in case work whose method of manufacture leads to an expectation of a high degree of subclass characteristics [140–142]. Because they use a simple blowback operation, they may also produce relatively weaker breech face marks [143,144]. Jimenez breech faces are machine milled, introducing the potential for repeating subclass characteristics in toolmarks among several breech faces.

To produce bullets, Beretta M9A3-FDE (the same as those used to produce cartridge cases) and Ruger SR9c pistols were chosen. Beretta barrels are triple-broached and Ruger barrels are double-broached. For both these operations the broach tool cuts parallel with the barrel bore center axis, which introduces the potential to produce similar toolmarks across more than one barrel. Both brands are therefore presumptively more likely to produce bullets that exhibit subclass characteristics. The presence of subclass characteristics was reported in other studies of double-broached Ruger barrels [145,146]. Due to wide Jimenez barrel specifications, bullets fired from them are often not identifiable, and thus were deemed unsuitable for the study.

4.2. Supervision of firearm manufacturing

In order to ensure acquisition of consecutively manufactured firearms, an FBI employee observed the relevant production operations of the Beretta M9A3-FDE, Ruger SR9c, and Jimenez JA-Nine pistols used in this study.

At the Beretta plant in Gallatin, Tennessee, flat barrel blanks imported from Italy are first drilled to diameter then successively rifled by three ganged broaches to progressively cut and finish the rifling. Broaches rotate as they are drawn through fixed barrel blanks. The machinist inspected the barrel bores visually and with multiple go/no-go and dial gauges. An entire production run of 66 barrels was witnessed; the FBI purchased 55 complete firearms with consecutive barrels (not all were used for the study; see section 5.1). Two triply-broached barrel flats for sequence numbers 20 and 55 were assessed prior to exterior finishing via coordinate measuring machine (CMM). All barrels were re-marked as successive machining operations of the barrel exterior removed temporary sequence codes, being hammer stamped after the final process. Heat treating and surface finishing finalized barrel manufacture. Slides were produced by multiple machining operations. Broaching of breech faces within the slide and pointing using a pneumatic press, which created a chamfer around the firing pin hole, were non-consecutive operations. As one of the final steps, the breech faces of the 55 slides purchased by the FBI were sequentially finished by hand filing by one person (witnessed by an FBI employee), after which each was hammer-stamped with a sequence code. After deburring and polishing the slides in a barrel tumbler, slides and barrels were Cerakote®-treated for resistance to scratching and corrosion. Non-consecutive firing pins and extractors were made in the same Beretta factory.

Barrel blanks from their Newport, New Hampshire facility were shipped to the Ruger plant in Prescott, Arizona for production. After barrel bore drilling, the exterior features were cut and the chamber was completed. Barrels were heat treated before rifling. Production of consecutively-rifled barrels was witnessed. A first ganged broach was used for final bore sizing and finishing. This bore broach is used to clean the bore and is pulled through the barrel parallel to the bore axis. Simultaneously, the barrel is rotated as the bore broach is pulled through the barrel. The surface left behind represents the tops of the lands. A second, ganged rifling broach cuts the twisted grooves. A machinist inspected every 20th barrel visually and with go/no go gauges. Barrel residual stress was relieved by heat treatment.

In Jimenez pistols, contract-manufactured bolt inserts are inserted into the frame during the casting of frames by a second outside company [140]. These operations were not witnessed. The front face of the bolt insert functions as the breech face. Consecutive breech face milling was witnessed at the Jimenez factory in Henderson, Nevada. This operation is performed to remove any flashing remaining from casting to operator-controlled extent, depending upon how many burrs remain [140–142]. Although bullets fired from them were not used in this study, the rifling of ten consecutive Jimenez barrels was also witnessed. Firing pins and extractors were also made under contract and are not consecutive.

Sometimes overlooked in validation studies for firearms identification is the break-in period—the time when the first test fires are produced in a firearm. The need to document such "settle-in" effects was emphasized by the National Research Council [1, pp. 79–82], although studies indicate that it is of minor consequence, particularly for breech face marks [39]. Initially, the manufacturing marks can be coarser, thus producing stronger microscopic marks for comparison. However, as a firearm is continually fired, the microscopic features tend to reach a steady state (normalize). When comparing evidence, the microscopic features have typically normalized.

All firearms were test-fired, first by the manufacturer for function and safety, then by the FBI Firearms-Toolmarks Unit (FTU), to establish a break-in period before collection of bullets and cases for the study. Beretta and Ruger pistols were test-fired a total of 60 times and the

Jimenez pistols, 30 times.

4.3. Choice of ammunition

Several metallic configurations of ammunition were considered as candidates for toolmark reproduction. Initially, our intent was to use both brass- and steel-cased ammunition in the study. In order to remove two independent variables (case type and primer type, see section 2.4) and due to limited reproduction of microscopic marks of value for pattern examination (desirable for the study), the decision was made after pilot testing (section 4.4) to use only steel-cased ammunition. Although less commonly encountered in case work, steel-cased ammunition is in wide military use by former Warsaw Pact nations [147], is cheaper to produce than brass, and is imported and sold at lower relative cost in the U.S. consumer market. More importantly to the present study, breech face and other marks are less prominent in steel cartridge cases than those in softer brass, therefore more challenging to compare. Also due to hardness, since steel cases are anticipated to obturate to a lesser extent than brass, a regimen of cleaning every 250 rounds was instituted to remove consequently increased sooty deposits.

The ammunition chosen for this study was Wolf Polyformance 9 mm Luger (9 × 19mm). The cartridges have polymer-coated steel cases, brass Berdan primers, and 115 grain full metal jacket (FMJ) bullets, supplied in boxes of 50 cartridges. The Wolf bullet consists of a lead core with a copper plated steel jacket [148,149]. Sufficient ammunition was purchased and fired to accommodate supplying the planned 300 participants with packets of test specimens—30,000 rounds.

We explored the possibility of using fewer fired specimens by generating exact replicas by two-step double casting [150–154]. Replication offers the advantage of exact duplicates, simplified marking (since the same physical specimens need not be repeatedly resent), and shorter turnaround time for examination if replicates are sent to more than one examiner at a time. That option was abandoned due to time and technical limitations in generating large numbers of sufficiently high-quality double castings.

4.4. Pilot study

The FBI conducted a pilot test to determine suitable firearms and ammunition that would prove challenging for examiners. FBI researchers solicited for participants, collected test materials, generated a small-scale black box test, and assembled examiner evaluation/feedback test data. Test specimens consisting of 100 fired cases and 100 fired bullets were collected using Beretta Model 92, Hi-Point Model C, and Ruger SR9c firearms. The caliber selected was 9 mm Luger (9 × 19 mm) and consisted of brass and steel cartridge cases and copper- and steel-jacketed lead bullets. To fine tune the design, 5 sets of one Q and two K bullets and 5 sets of one Q and two K cartridge cases were distributed to four state F/T examiners and three FBI examiners. Feedback was requested on study design; specimen comparison difficulty; quality, quantity, reproducibility of toolmarks; informed consent form; survey form; and any other information the participants wanted to provide. Also provided were proposed versions of the questionnaire, examination instructions, and reporting form, upon which the examiners were invited to comment:

- Are the instructions clear describing what is expected from the examiner for this study? If not, how can they be improved?
- Is the packaging and labeling of the cartridge cases, bullets, and sets adequate? If not, can you suggest how the packaging can be improved?
- Do you wish to comment on the design of the test? (Please keep in mind that the purpose of this study is NOT to evaluate your facility's procedures but to determine the false positive and false negative rates for individual examiners comparing bullets and cartridge cases.)

- Please comment on the Reporting Form provided in this pilot study. Is the form laid out in an understandable manner? If not, where is it unclear? Should anything be added?
- The range of possible findings (Identification, Elimination, three Inconclusive levels, or Unsuitable) are from the AFTE Glossary, "Range of Conclusions Possible when Comparing Toolmarks." Should any other finding options be added? Why?
- Can you provide the approximate total time it took for the completion of all examinations? Could there be more, or should there be fewer comparisons in a packet? (The study will consist of 4–6 of these packets being sent to an examiner over a ~2-year period.)
- Is the Informed Consent Form clear and understandable?
- Is there anything else regarding this study about which you'd like to comment?

Improvements were made to the study materials based on the volunteers' comments. With input from the Ames team, the outline for the final study was refined and decided upon by the FBI team, who selected the firearms and ammunition to use, developed a firing plan for the collection of specimens, designed and tested a method for holding and mounting bullet specimens. The Ames team developed methods of specimen randomization to avoid dependence among examiner evaluations.

5. Project execution: testing and analysis

5.1. Specimen collection: sampling

Sampling was designed to satisfy several goals: to evaluate the production of toolmarks early, midway, and late in the life of the tool; changes related to firing order; marks characteristic of various firearms; and multiple intervals within a single manufacturing run of the same firearm.

For generation of fired cartridge cases, we used 10 new Jimenez and 23 new Beretta firearms with consecutively manufactured slides (or for the Berettas, consecutive clusters within a single manufacturing run) plus four Berettas from the FBI Reference Firearms Collection (RFC). The RFC firearms are of unknown provenance, derived from adjudicated cases. When a Jimenez internal mechanical part, necessary for operation but unrelated to identification of fired ammunition, failed during specimen collection, parts from a Bryco 59 from the RFC were used to replace parts necessary to permit continued test firing.[1] Within consecutively manufactured runs, clusters of Beretta slides were chosen to represent the beginning, middle, and end of the useful life of the final cutting tool.

- 10 Jimenez slides, consecutively machine-finished under manual control
- 23 Beretta slides, breech faces consecutively hand-finished
- 4 Beretta slides taken from the FBI-RFC (previously used, unrelated to one another)

Each Jimenez or Bryco was fired 850 times, resulting in 9350 cartridge cases. Each Beretta was fired 700 times, resulting in 18,900 Beretta cartridge cases.

Bullet specimens were collected in a similar fashion, using 10 new Ruger and the same 23 new Beretta guns, all with sequentially manufactured barrels, plus 1 Ruger and 4 Berettas from the RFC. Clusters of Beretta barrels were chosen to represent manufacturing intervals. Bullets that became deformed from impacting the walls of the water tank were not used in this study.

---

[1] The Jennings JA-Nine is essentially a renamed Bryco 59, being manufactured with the same molds and equipment [140] A.K. Welch, History and manufacturing process of the Jennings/Bryco/Jimenez Arms pistols, AFTE J 45 (3) (2013) 260–266.

Fig. 4. Water tank used to collect bullets in groups of 50. An enclosure net was used to catch expended cartridge cases (Caldwell Shooting Supplies, Columbia, MO).

- 23 Beretta barrels (with sequence numbers 1–5, 16–19, 31–35, 46–50, 62–66) within production by the same three successive broaches
- 4 Beretta barrels taken from the FBI-RFC (previously used, unrelated to one another)
- 10 Ruger barrels (with sequence numbers 1–9, 33) within production by the same two successive broaches

Each Ruger was fired 850 times, resulting in 9350 bullets. Each Beretta was fired 700 times, resulting in 18,900 Beretta bullets. Beretta bullets were collected concurrently with the Beretta cartridge cases.

All ammunition was fired into a water tank (Fig. 4) and collected in groups of 50. Firing order was tracked to enable conclusions concerning the effect of firearm wear on examiners' analysis results. Ejected cartridge cases were recovered from a fired brass trap fitted to the tank (Caldwell Shooting Supplies, Columbia, MO). When airborne lead rose to a level of concern midway through specimen collection, an exhaust chamber for lead remediation was added to enclose the firing area (not shown). A dry cloth patch, without solvent, was used to clean all barrels and breech faces after every group of 250 shots.

Fired cartridge cases were returned to their original packaging, including 50-holed plastic holders that prevented individual cartridge cases from contacting each other and potentially acquiring additional marks. The containers were labeled with the gun serial number and the sequential firing order (within a range of 50 fired specimens) prior to shipment to Ames Laboratory. The FBI placed collected bullets in small manila envelopes that were labeled with the gun serial number and the sequential firing order (within a range of 50 fired specimens) prior to shipment to Ames Laboratory.

Separate cardboard boxes, one for each of the Jimenez, Ruger, and Beretta firearms, were used to contain all the specimens for a given serial number firearm and firing sequence, so 11 boxes of Jimenez cartridge cases, 11 boxes of Ruger bullets, and 27 boxes of Beretta cartridge cases and 27 boxes of Beretta bullets were shipped by the FBI to Ames. For example, each Jimenez cardboard box contained 17 labeled boxes of 50 cartridge cases (850 rounds fired per gun), so the collection of Jimenez boxes totaled 187 boxes of 50 fired cases (9350 specimens).

5.2. Mounting and labeling specimens for distribution

To maintain the integrity of the study, specimens distributed to examiners must be mounted and labeled such that they: 1) prevent participants from inferring test design, 2) preclude sharing or recall of results on individual packets when submitted to another examiner or resubmitted to the same examiner, and 3) minimize specimen handling

to avoid introduction of extraneous marks. In addition, conduct of a double-blind study requires the participation of a third party for labeling, selection, and record-keeping of test specimens. The FBI contracted Ames Laboratory to fulfill these logistics functions, to help fine-tune the study design, particularly related to statistical sampling of test components, and to conduct data analysis and reporting. Although the FBI was responsible for overall design, funding, and materials, this contractual relationship accords with recommendations that decision analysis studies be conducted in conjunction with independent, non-law enforcement, entities not engaged in performing forensic F/T examinations [29,36,155].

Within Ames Laboratory, two independent groups were established. The communication group, which had contact with the participating firearm examiners; maintained a list of names and addresses; collected consent forms; and arranged for shipping and receiving of test packets to and from the examiners; and the experimental/analysis group, which was responsible for constructing the cartridge case and bullet sample sets to be analyzed; assigning a primary tracking code for each packet; scoring and verification of the results; repackaging analyzed packets for subsequent mailings of the study; and performing statistical analysis of the reported results. Communication between the two Ames groups was restricted to the exchange of packets through the use of a secondary identification code. The secondary code was used to identify packets that either needed to be sent to examiners or had been returned by examiners. The secondary code was located on the outside of a sealed container containing the assembled test packets. The secondary code created a firewall allowing transfer of packets between the two groups while protecting details of their true makeup. Packets returned by examiners were logged and inspected upon arrival by the communication group so that any examiner-specific identifying information could be removed. Then the sealed bag containing the analysis results was transferred to the experimental/analysis group for scoring, database entry, and verification of the results.

Multiple stages of labeling of individual bullets and cartridge cases, of test sets (1Q+2 K), and of each mailing packet (all test sets for examination in a single round of testing) were used to: identify each item and its source, remain securely attached to it yet prevent inference of test design or correlation of the item with its source, and maintain blinded administration of testing by isolating the knowledge available to different groupings of researchers (FBI researchers, Ames communications group, and Ames experimental/analysis group). Two-dimensional QR codes were chosen to label individual bullets and cartridge cases because they can code extensive information in a small area and discourage attempts to decipher their contents (which consisted of unique randomized alphanumeric characters nonetheless) or to recognize the same specimens if resubmitted for repeatability studies. Considerable effort was expended in attempting to laser engrave QR codes directly onto specimens. Even after constructing custom specimen holding jigs, problems due to curved surfaces (cartridge cases, bullet ogive) and in balancing laser power levels to engrave plastic bullet mounts without melting them resulted, so that direct engraving had to be abandoned. Additional information on labeling procedures is provided in appendix 6, Specimen labeling details.

5.3. Test packet composition, assembly, and distribution

Test packets and comparison sets were assembled using the following parameters:

1. A test set consisted of a single Q to two Ks, the latter fired from the same firearm.
2. Only cartridge cases and bullets fired from the same make and model firearm were compared. This was not revealed to the examiners.
3. The ratio of non-Beretta to Beretta specimens (for cartridge cases and bullets) was 1:2.
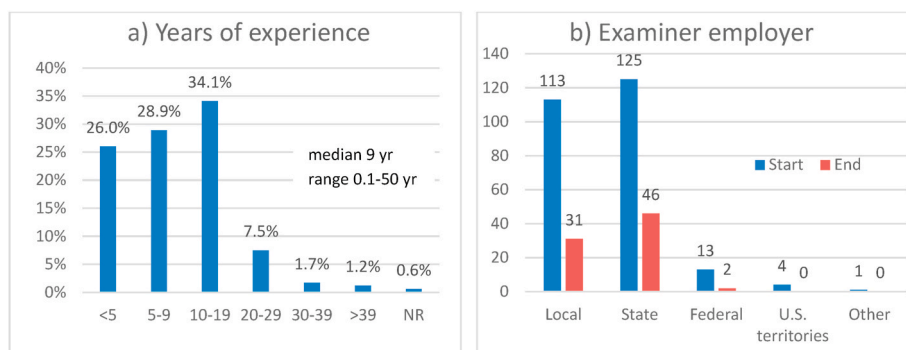
Fig. 5. Examiner demographics: a) Years of experience (NR = no response) and b) employer of those participating at the start and at the conclusion of all three rounds of the study.

4. An open set design was utilized, i.e., there was not necessarily a match for every Q.
5. Each test set represents an independent comparison unrelated to any other set.
6. The overall ratio of known matches was 1:2, but varied across test sets.

Each of the six mailings that a firearm examiner received consisted of 30 comparison specimen sets made up of 15 comparisons of 2 knowns to 1 questioned cartridge case and 15 comparisons of 2 knowns to 1 questioned bullet. The cartridge case comparisons consisted of 5 sets of Jimenez and 10 sets of Beretta cartridge cases. Bullet comparison sets comprised 5 sets of Ruger and 10 sets of Beretta bullets. An overall ratio of known match to non-match (KM/KNM) ratio of 1:2 was implemented. The ratio was disguised from the participants by varying the ratio for both cartridge case and bullet comparison sets among the test packets prepared for distribution (see below). Participants were instructed not to share or discuss the contents of their packets or their reported results to minimize the risk of revealing details of the experimental design. Details of the design were not shared with anyone outside the group of Ames researchers assembling the comparison sets and FBI project managers. Only those researchers in the experimental/analysis group at Ames Lab knew the ground truth for the assembled test packets. Before reusing specimen sets in subsequent rounds, the specimens were visually examined and gently cleaned, and the packets (but not individual specimens) were relabeled. After all results were recorded and accounted for, all codes that related individual examiners to their results were destroyed to assure anonymity.

To minimize the possibility that participants might deduce a pattern in the total number of KM sets in a packet, confer, and therefore affect subsequent analyses, a repeating 25-element matrix was used to assemble the test sets of cases and bullets for analysis. For each subgroup of 5 mailing sets within a matrix of 25 mailing sets, the numbers of known-match sets varied from 0 to 4 Jimenez and from 1 to 5 Beretta cartridge case sets, and from 0 to 4 Ruger and from 1 to 5 Beretta bullet sets, such that each packet had between 3 and 7 known-match cartridge case and between 3 and 7 known-match bullet sets. The numbers of KM cartridge case and KM bullet sets were offset with respect to each other to vary the total number of KM sets in the packets. Thus, the total number of KM sets ranged from 6 to 14 but an overall ratio of 1:2 kM to KNM sets for the group of 25 packets was maintained. In addition, all 25 possible permutations of combining 3–7 kM cartridge case with 3–7 kM bullet comparison sets were achieved for these packets. All specimens were returned to the FBI at the conclusion.

5.4. Number of participants and recruitment

The study was designed to provide a representative estimate of the performance of F/T examiners who testify to their conclusions in court.

Participation was limited to fully qualified examiners who were currently conducting firearm examinations, were members of AFTE, and were employed in the firearms section of an accredited public crime laboratory within the U.S. or a U.S. territory. No incentives were offered. European examiners were excluded due to limitations on shipping ammunition that required mutilation to render it unusable for reloading [87] and because many countries use alternative means of expressing their judgments in terms of degree of support for same- or different source [82,96,139,156–158], for which comparison or conversion to error rate is ambiguous. A decision was made to exclude any examiners currently employed by the FBI to avoid potential bias, even though some were not involved in the specimen preparation process.

The initial plan was for participation of 300 examiners in the study. Broad calls for volunteers were made through the AFTE website, via announcements by FBI personnel at national forensic meetings, and through e-mail lists maintained by AFTE. The letter of invitation is shown in appendix 1. These methods resulted in 270 respondents who contacted the Ames communication group. Exclusion of FBI examiners reduced the initial starting number of respondents to 256 participants. Once the first mailing of specimen packets was distributed and volunteers became fully aware of the amount of work required, many examiners decided to drop out of the study without analyzing the first test packet. Additional examiners withdrew from the study over the course of the data collection period. Other examiners joined the study after it was underway. A total of 173 examiners working in 41 states returned evaluations and were active in some part of the study. At the conclusion of the study, only 79 participants had finished all six mailings of test-packet analyses. Despite the high dropout rate, the average number of packets completed by an individual participant was between three and four; in other words, the average examiner contributed over 100 specimen-set comparisons to the study. Approximately 95% of participating examiners were employed by state and local crime labs, with the remainder at federal and U.S. territory labs, with a range in years of experience (Fig. 5).

6. Outcome to be reported in subsequent papers

Analysis of experimental results will be presented in a series of forthcoming publications. Topics include accuracy, reproducibility and repeatability, observations due to different firearms, manufacturing sequence, firing order; and trends in application of the AFTE Range of Conclusions by examiners.

CRediT authorship contribution statement

Keith L. Monson: Conceptualization, Supervision, Visualization, Investigation, Methodology, Resources, Project administration, Writing – original draft, Writing – review & editing. **Erich D. Smith:** Conceptualization, Supervision, Investigation, Methodology, Visualization,

Writing – review & editing. **Stanley J. Bajic:** Investigation, Writing – review & editing.

Declaration of competing interest

None.

Acknowledgments

This is manuscript 22.03 of the FBI Laboratory. Names of commercial manufacturers are provided for identification purposes only, and inclusion does not imply endorsement of the manufacturer or its product or services by the FBI. Any reproduction, or other use of these presentation materials without the express written consent of the FBI is prohibited. The views are those of the author(s) and do not necessarily reflect the official policy or position of the FBI or the U.S. Government.

Appendices.

1. Letter of invitation
2. Informed consent
3. Survey of examiner-specific factors
4. Instructions for performing comparisons and reporting results
5. Reporting form
6. Specimen labeling details

References

[1] NRC, Ballistic imaging, in: D. Cork, J. Rolph, E. Meieran, C. Petrie (Eds.), National Research Council, National Academy of Sciences, 2008. https://www.nap.edu/login.php?record_id=12162&page=https%3A%2F%2Fwww.nap.edu%2Fdownload%2F12162.

[2] K. Bonfanti, J. De Kinder, The influence of manufacturing processes on the identification of bullets and cartridge cases—a review of the literature, Sci. Justice 39 (1) (1999) 3–10, https://doi.org/10.1016/s1355-0306(99)72008-3.

[3] R.S. Bolton-King, Preventing miscarriages of justice: a review of forensic firearm identification, Sci. Justice 56 (2) (2016) 129–142, https://doi.org/10.1016/j.scijus.2015.11.002.

[4] G. Gerules, S. Bhatia, D. Jackson, A survey of image processing techniques and statistics for ballistic specimens in forensic science, Sci. Justice 53 (2) (2013) 236–250, https://doi.org/10.1016/j.scijus.2012.07.002.

[5] C. Monturo, Forensic Firearm Examination, Academic Press, London, 2019.

[6] R. Nichols, Firearm and Toolmark Identification: the Scientific Reliability of the Forensic Science Discipline, Academic Press, London, 2018.

[7] R. Saferstein, in: Forensic Science: from the Crime Scene to the Crime Lab, fourth ed., Pearson, New York, 2019.

[8] S.G. Bunch, E.D. Smith, B.N. Giroux, D.P. Murphy, Is a match really a match? A primer on the procedures and validity of firearm and toolmark identification, Forensic Sci Comm, Federal Bureau of Investigation, 2009. https://www2.fbi.gov/hq/lab/fsc/backissu/july2009/review/2009_07_review01.htm.

[9] OSAC, Firearms Process Map, Firearms & Toolmarks Subcommittee, Organization of Scientific Area Committees for Forensic Science, in: https://www.nist.gov/document/osac-firearms-process-mapjan2021, 2021.

[10] AFTE, Theory of Identification as it relates to toolmarks, AFTE J 43 (4) (2011) 287.

[11] AFTE, AFTE Theory of Identification, AFTE J 24 (2) (1992) 336–340.

[12] R. Grzybowski, J. Miller, B. Moran, J. Murdock, R. Nichols, R. Thompson, Firearms/toolmark identification: passing the reliability test under federal and state evidentiary standards, AFTE J 35 (2) (2003) 209–241.

[13] Coffin vs. U.S., 156 U.S. 432, U.S., 1895.

[14] U.S.v. Mitchell, 365 F.3d 215, 3d Cir, 2004.

[15] Daubert v. Merrell Dow Pharmaceuticals, Inc., rrell Dow Pharmaceuticals, Inc., 509 U.S. 579, U.S, in: Me, 1993.

[16] NRC, Strengthening Forensic Science in the United States: A Path Forward, National Research Council, National Academy of Sciences, 2009. https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf.

[17] B. Ulery, R. Hicklin, J. Buscaglia, M. Roberts, Accuracy and reliability of forensic latent fingerprint decisions, Proc. Natl. Acad. Sci. Unit. States Am. 108 (19) (2011) 7733–7738, https://doi.org/10.1073/pnas.1018707108.

[18] B. Ulery, R. Hicklin, J. Buscaglia, M. Roberts, Repeatability and reproducibility of decisions by latent fingerprint examiners, PLoS One 7 (3) (2012), e32800, https://doi.org/10.1371/journal.pone.0032800.

[19] SWGGUN, SWGGUN Admissibility Resource Kit (ARK), Scientific Working Group for Firearms and Toolmarks. https://afte.org/resources/swggun-ark, 2019.

[20] U.S.v. Darryl Green, et al., 405 F. Supp. 2d 104, D. Mass., 2005.

[21] U.S.v. Tibbs, Memorandum Opinion, 2016 CF1 19431, D.C. Sup., 2019.

[22] Williams v. U.S., 210 A.3d 734, D.C. App., 2016.

[23] U.S.v. Amando Monteiro, et al., 407 F. Supp. 2d 351, D. Mass, 2006.

[24] U.S.v. Chaz Glynn, No. 06 CR 580(JSR), S.D.N.Y., 2008.

[25] U.S.v. Edgar Diaz, et al., No. 05-00167 WHA, N.D. Cal., 2007.

[26] U.S.v. Tayon Mouzone, WDQ-08-086, D. Md., 2009.

[27] U.S.v. Jermaine, Dore and Dwayne Barrett, 12 Cr, vol. 45, S.D.N.Y., 2013.

[28] AFTE, Court Citations, 2020. (Accessed 7 January 2020).

[29] J.J. Koehler, Forensics or fauxrensics: ascertaining accuracy in the forensic sciences, Ariz. State Law J. 49 (2017) 1369–1416.

[30] A. Schwartz, A systemic challenge to the reliability and admissibility of firearms and toolmarks identification, Columbia Law Rev. (2005) 1–42. https://journals.library.columbia.edu/index.php/stlr/article/download/3776/1578.

[31] W. Tobin, P. Blau, Hypothesis testing of the critical underlying premise of discernible uniqueness in firearms-toolmarks forensic practice, Jurimetric J 53 (2) (2013) 121–142.

[32] C. Spiegelman, W. Tobin, Analysis of experiments in forensic firearms/toolmarks practice offered as support for low rates of practice error and claims of inferential certainty, Law Probab. Risk 12 (2) (2013) 115–133, https://doi.org/10.1093/lpr/mgs028.

[33] P.C. Giannelli, Daubert Challenges to Firearms ("ballistics") Identifications, Case W Res Law Faculty Publ, 2007, pp. 548–568. Case Western Reserve School of Law, http://scholarlycommons.law.case.edu/cgi/viewcontent.cgi?article=1153&context=faculty_publications.

[34] D.H. Kaye, Firearm-mark evidence: looking back and looking ahead, Case W Res Law Rev (2017) 723–746. https://scholarlycommons.law.case.edu/caselrev/vol68/iss3/13.

[35] Louisiana v. Goodwin-Bey, 1531-CR00555-01, Cct. Ct, Green Co. Div. V, 2016.

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

[37] U.S. DOJ, in: U.S.D.o. Justice (Ed.), U.S. DOJ, U.S. Department of Justice Statement on the PCAST Report, 2021. https://www.justice.gov/olp/page/file/1352496/download.

[38] SWGGUN, Testability of the scientific principle. http://www.swggun.org/swg/index.php?option=com_content&view=article&id=5:testability-of-the-scientific-principle&catid=9:ark&Itemid=18, 2011.

[39] R. Nichols, Defending the scientific foundations of the firearms and tool mark identification discipline: responding to recent challenges, J. Forensic Sci. 52 (3) (2007) 586–594, https://doi.org/10.1111/j.1556-4029.2007.00422.x.

[40] J. Hamby, D. Brundage, J. Thorpe, The identification of bullets fired from 10 consecutively rifled 9mm Ruger pistol barrels: a research project involving 507 participants from 20 countries, AFTE J 41 (2) (2009) 99–110.

[41] Y. Ogihara, M. Kubota, M. Sandra, K. Fukuda, T. Uchiyama, J. Hamby, Comparison of 5,000 consecutively fired bullets and cartridge cases from a 45 caliber M1911A1 pistol, AFTE J 15 (3) (1983) 127–140.

[42] J. Gouwe, J. Hamby, S. Norris, Comparison of 10,000 consecutively fired cartridge cases from a Model 22 Glock .40 S&W caliber semiautomatic pistol, AFTE J 40 (1) (2008) 57–63.

[43] T. Uchiyama, Toolmark reproducibility on fired bullets and expended cartridge cases, AFTE J 40 (1) (2008) 3–46.

[44] A. Azahidi, Breechface recess marks on cartridge cases discharged from 9mm Walther P99 series pistols and their persistence, AFTE J 44 (3) (2012) 244–247.

[45] A. Saribay, A. Hannam, C. Tarimci, An investigation into whether or not the class and Individual characteristics of five Turkish manufactured pistols change during extensive firing, J. Forensic Sci. 54 (5) (2009) 1068–1072, https://doi.org/10.1111/j.1556-4029.2009.01107.x.

[46] C. Wong, The inter-comparison of 1,000 consecutively-fired 9mm Luger bullets and cartridge cases from a Ruger P89 pistol utilizing both pattern matching and quantitative consecutive matching striae as criteria for identification, AFTE J 45 (3) (2013) 267–272.

[47] F. Vinci, R. Falamingo, C.P. Campobasso, J. Bailey, Morphological study of class and individual characteristics produced by firing 2500 cartridges in a .45 caliber semi-automatic pistol, AFTE J 37 (4) (2005) 368–373.

[48] S. Kirby, Comparison of 900 consecutively fired bullets and cartridge cases from a 455 caliber S&W revolver, AFTE J 15 (3) (1983) 113–126.

[49] D. Brundage, The identification of consecutively rifled gun barrels, AFTE J 30 (3) (1998) 438–444.

[50] J. Hamby, J. Thorpe, The examination, evaluation and identification of 9mm cartridge cases fired from 617 different Glock Model 17 & 10 semiautomatic pistols, AFTE J 41 (4) (2009) 310–324.

[51] W. Matty, Raven .25 automatic pistol breech face tool marks, AFTE J 16 (3) (1984) 57–60.

[52] E. Thompson, Phoenix Arms (Raven) breech face toolmarks, AFTE J 26 (2) (1994) 134–135.

[53] S. Bunch, D. Murphy, A comprehensive validity study for the forensic examination of cartridge cases, AFTE J 35 (2) (2003) 201–203.

[54] A. Coody, Consecutively manufactured Ruger P-89 slides, AFTE J 35 (2) (2003) 157–160.

[55] B. Coffman, Computer Numerical Control (CNC) production tooling and repeatable characteristics on ten Remington Model 870 production run breech bolts, AFTE J 35 (1) (2003) 49–54.

[56] T. Fadul, G. Hernandez, E. Wilson, S. Stoiloff, S. Gulati, An Empirical Study to Improve the Scientific Foundation of Forensic Firearm and Tool Mark Identification Utilizing 10 Consecutively Manufactured Glock EBIS Barrels with the Same EBIS Pattern, National Institute of Justice, 2012. www.ncjrs.gov/pdffiles1/nij/grants/237960.pdf.

[57] P. Lardizabal, Cartridge case study of the HK USP, AFTE J 27 (1) (1995) 49–51.

[58] T. Weller, A. Zheng, R. Thompson, F. Tulleners, Confocal microscopy analysis of breech face marks on fired cartridge cases from 10 consecutively manufactured pistol slides, J. Forensic Sci. 57 (4) (2012) 912–917, https://doi.org/10.1111/j.1556-4029.2012.02072.x.

[59] C.E. Castro, S.A. Norris, B.K. Setume, J.E. Hamby, The examination, evaluation, and identification of .40 S&W calibre cartridge cases fired from 1079 different Glock semiautomatic pistols manufactured over a six-year period, J. Can. Soc. Forensic. Sci. 47 (2) (2014) 65–73, https://doi.org/10.1080/00085030.2014.904105.

[60] T. Fadul, G. Hernandez, S. Stoiloff, S. Gulati, An empirical study to improve the scientific foundation of forensic firearm and tool mark identification utilizing 10 consecutively manufactured slides, AFTE J 45 (4) (2013) 376–393.

[61] D. LaPorte, An empirical and validation study of breechface marks on .380 ACP caliber cartridge cases fired from ten consecutively finished Hi-Point Model C9 pistols, AFTE J 43 (4) (2011) 303–309.

[62] M. Cazes, J. Goudeau, Validation study results from Hi-Point consecutively manufactured slides, AFTE J 45 (2) (2013) 175–177.

[63] J. Hamby, Identification of projectiles, AFTE J 6 (5/6) (1974) 22.

[64] R. Shem, P. Striupaitis, Comparison of 501 consecutively fired bullets and cartridge cases from a .25 caliber Raven pistol, AFTE J 15 (3) (1983) 109–112.

[65] D. Mikko, J. Miller, J. Flater, Reproducibility of toolmarks on 20,000 bullets fired through an M240 machine gun barrel, AFTE J 44 (3) (2012) 248–253.

[66] S. Christen, H.R. Jordi, Individuality testing of new Glock pistol barrels "Marksman Barrel", Forensic Sci. Int. 295 (2019) 64–71, https://doi.org/10.1016/j.forsciint.2018.11.028.

[67] Z. Wei, Y. Luo, P. Zhou, K. Ji, Reproducibility and persistence of individual marks on 3000 fired bullets from five Chinese Norinco QSZ-92 9 × 19 mm pistols, and the search performance of Evofinder® to a 1000 firearm database of the same model firearm, J. Forensic Sci. 66 (6) (2021) 2387–2392, https://doi.org/10.1111/1556-4029.14794.

[68] T. Fadul, An empirical study to evaluate the repeatability and uniqueness of striations/impressions imparted on consecutively manufactured Glock EBIS gun barrels, AFTE J 43 (1) (2011) 37–44.

[69] J. Miller, An examination of two consecutively rifled barrels, AFTE J 32 (3) (2000) 259–270.

[70] J.E. Hamby, D.J. Brundage, N.D.K. Petraco, J.W. Thorpe, A worldwide study of bullets fired from 10 consecutively rifled 9mm Ruger pistol barrels—analysis of examiner error rate, J. Forensic Sci. 64 (2) (2019) 551–557, https://doi.org/10.1111/1556-4029.13916.

[71] J.A. Smith, Beretta barrel fired bullet validation study, J. Forensic Sci. 66 (2) (2021) 547–556, https://doi.org/10.1111/1556-4029.14604.

[72] C.S. DeFrance, M.D. Van Arsdale, Validation study of electrochemical rifling, AFTE J 35 (1) (2003) 33–37.

[73] AFTE, Glossary, AFTE standardization and training committee, Association of Firearm and Tool Mark Examiners. https://afte.org/uploads/documents/AFTE_Glossary_Version_6.110619_DRAFT_.PDF, 2013.

[74] M.A. Keisler, S. Hartman, A. Kilmon, M. Oberg, M. Templeton, Isolated pairs research study, AFTE J 50 (1) (2018) 56–58.

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

[76] T.P. Smith, A.G. Smith, J.B. Snipes, A validation study of bullet and cartridge case comparisons using samples representative of actual casework, J. Forensic Sci. 61 (4) (2016) 939–946, https://doi.org/10.1111/1556-4029.13093.

[77] E.D. Smith, Cartridge case and bullet comparison validation study with firearms submitted in casework, AFTE J 37 (4) (2005) 130–135.

[78] S. Owens, An examination of five consecutively rifled Hi-Point 9mm pistol barrels with three lands and grooves left twist rifling to assess identifiability and the presence of subclass characteristics, AFTE J 49 (4) (2017) 208–215.

[79] R. Nichols, Subclass characteristics: from origin to evaluation, AFTE J 50 (2) (2018) 68–88.

[80] G. Rivera, Subclass characteristics in Smith & Wesson SW40VE sigma pistols, AFTE J 39 (3) (2007) 247–253.

[81] L. Lightstone, The potential for and persistence of subclass chracteristics on the breech faces of SW40VE Smith & Wesson Sigma pistols, AFTE J 42 (4) (2010) 308–322.

[82] F. Riva, R. Hermsen, E. Mattijssen, P. Pieper, C. Champod, Objective evaluation of subclass characteristics on breech face marks, J. Forensic Sci. 62 (2) (2017) 417–422, https://doi.org/10.1111/1556-4029.13274.

[83] L. Lopez, S. Grew, Consecutively machined Ruger bolt faces, AFTE J 32 (1) (2000) 19–24.

[84] D. Lyons, The identification of consecutively manufactured extractors, AFTE J 41 (3) (2009) 246–256.

[85] F.A. Tulleners, J.S. Hamiel, Sub class characteristics of sequentially rifled 38 Special S&W revolver barrels, AFTE J 31 (2) (1999) 117–122.

[86] T. Uchiyama, Similarity among breech face marks fired from guns with close serial numbers, AFTE J 18 (3) (1986) 15–52.

[87] D. Baldwin, S. Bajic, M. Morris, D. Zamzow, A Study of False-Positive and False-Negative Error Rates in Cartridge Case Comparisons, Ames Laboratory, U.S. Department of Energy, 2014. https://apps.dtic.mil/dtic/tr/fulltext/u2/a611807.pdf.

[88] W. Kerkhoff, R.D. Stoel, C.E.H. Berger, E.J.A.T. Mattijssen, R. Hermsen, N. Smits, H.J.J. Hardy, Design and results of an exploratory double blind testing program in firearms examination, Sci. Justice 55 (6) (2015) 514–519, https://doi.org/10.1016/j.scijus.2015.06.007.

[89] W. Kerkhoff, R.D. Stoel, E.J.A.T. Mattijssen, C.E.H. Berger, F.W. Didden, J.H. Kersholt, A part-declared blind testing program in firearms examination, Sci. Justice 58 (4) (2018) 258–263, https://doi.org/10.1016/j.scijus.2018.03.006.

[90] J.E. Hamby, S. Norris, N.D. Petraco, Evaluation of Glock 9 mm firing pin aperture shear mark individuality based on 1,632 different pistols by traditional pattern matching and IBIS pattern recognition, J. Forensic Sci. 61 (1) (2016) 170–176, https://doi.org/10.1111/1556-4029.12940.

[91] L.S. Chumbley, M.D. Morris, S.J. Bajic, D. Zamzow, E. Smith, K. Monson, G. Peters, Accuracy, Repeatability, and Reproducibility of Firearm Comparisons Part 1: Accuracy, 2021 arXiv,, https://arxiv.org/abs/2108.04030.

[92] E.F. Law, K.B. Morris, Evaluating firearm examiner conclusion variability using cartridge case reproductions, J. Forensic Sci. 66 (5) (2021) 1704–1720, https://doi.org/10.1111/1556-4029.14758.

[93] G. Langenburg, C. Champod, T. Genessay, Informing the judgments of fingerprint analysts using quality metric and statistical assessment tools, Forensic Sci. Int. 219 (1–3) (2012) 183–198, https://doi.org/10.1016/j.forsciint.2011.12.017.

[94] H. Eldridge, M. De Donno, C. Champod, Testing the accuracy and reliability of palmar friction ridge comparisons – a black box study, Forensic Sci. Int. 318 (2021) 110457, https://doi.org/10.1016/j.forsciint.2020.110457.

[95] NCFS, Proficiency Testing in Forensic Science, National Commission on Forensic Science, 2016. https://www.justice.gov/archives/ncfs/page/file/831806/download.

[96] J. Kersholt, A. Eikelbooma, T. Dijkman, R. Stoel, R. Hermsen, B. van Leuven, Does suggestive information cause a confirmation bias in bullet comparisons? Forensic Sci. Int. 198 (1–3) (2010) 138–142, https://doi.org/10.1016/j.forsciint.2010.02.007.

[97] E.J.A.T. Mattijssen, C.L.M. Witteman, C.E.H. Berger, R.D. Stoel, Cognitive biases in the peer review of bullet and cartridge case comparison casework: a field study, Sci. Justice 60 (4) (2020) 337–346, https://doi.org/10.1016/j.scijus.2020.01.005.

[98] P. Pauw-Vugts, A. Walters, L. Oren, FAID2009: proficiency test and workshop, AFTE J 45 (2) (2013) 115–127.

[99] FDA, Adaptive designs for medical device clinical studies, U.S, Food and Drug Administration, www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm446729.pdf, 2016.

[100] J. De Kinder, Ballistic fingerprinting databases, Sci. Justice 42 (4) (2002) 197–203, https://doi.org/10.1016/s1355-0306(02)71829-7.

[101] M.E. Schuckers, Using the beta-binomial distribution to assess performance of a biometric identification device, Int. J. Image Graph. 3 (2003) 523–529, https://doi.org/10.1142/S0219467803001147, 03.

[102] T.J. Weller, M.D. Morris, Commentary on: I. Dror, N scurich "(Mis) use of scientific measurements in forensic science", Forensic Sci Int Synergy 2 (2020) 701–702, https://doi.org/10.1016/j.fsisyn.2020.10.004. https://doi.org/10.1016/j.fsisyn.2020.08.006.

[103] D.-l. Li, F. Shen, Y. Yin, J.-x. Peng, P.-y. Chen, Weighted Youden index and its two-independent-sample comparison based on weighted sensitivity and specificity, Chin. Med. J. 126 (6) (2013) 1150–1154, https://doi.org/10.3760/cma.j.issn.0366-6999.20123102.

[104] K.J. Drobatz, Measures of accuracy and performance of diagnostic tests, J. Vet. Cardiol. 11 (2009) S33–S40, https://doi.org/10.1016/j.jvc.2009.03.004.

[105] I.E. Dror, N. Scurich, (Mis)use of scientific measurements in forensic science, Forensic Sci Int Synergy 2 (2020) 333–338, https://doi.org/10.1016/j.fsisyn.2020.08.006.

[106] H. Hofmann, A. Carriquiry, S. Vanderplas, Treatment of inconclusive results in firearms error rate studies, Law Probab. Risk 19 (2020) 317–364, https://doi.org/10.1093/lpr/mgab002.

[107] A. Biedermann, K.N. Kotsoglou, Forensic science and the principle of excluded middle: "Inconclusive" decisions and the structure of error rate studies, Forensic Sci Int Synergy 3 (2021) 100147, https://doi.org/10.1016/j.fsisyn.2021.100147.

[108] DOJ, Uniform Language for Testimony and Reports (ULTR) for the Forensic Firearms/Toolmarks Discipline – Pattern Examination, U.S. Department of Justice, 2020. https://www.justice.gov/olp/page/file/1284766/download.

[109] A. Biedermann, S. Bozza, F. Taroni, J. Vuille, Are inconclusive decisions in forensic science as deficient as they are said to be? Front. Psychol. 10 (2019) 520, https://doi.org/10.3389/fpsyg.2019.00520.

[110] J.J. Koehler, Fingerprint error rates and proficiency tests: what they are and why they matter, Hastings LJ 59 (5) (2008) 1077–1100.

[111] D.H. Kaye, The validity of tests: Caveant omnes, Jurimetrics J 27 (1986) 349–361.

[112] K. Kafadar, Statistical issues in assessing forensic evidence, Int. Stat. Rev. 83 (1) (2015) 111–134, https://doi.org/10.1111/insr.12069.

[113] I.E. Dror, The error in "Error Rate": why error rates are so needed, yet so elusive, J. Forensic Sci. 65 (4) (2020) 1034–1039, https://doi.org/10.1111/1556-4029.14435.

[114] B. Budowle, M.C. Bottrell, S.G. Bunch, R. Fram, D. Harrison, S. Meagher, C.T. Oien, P.E. Peterson, D.P. Seiger, M.B. Smith, M.A. Smrz, G.L. Soltis, R.B. Stacey, A perspective on errors, bias, and interpretation in the forensic sciences and direction for continuing advancement, J. Forensic Sci. 54 (4) (2009) 798–809, https://doi.org/10.1111/j.1556-4029.2009.01081.x.

[115] A. Biedermann, S. Bozza, F. Taroni, The decisionalization of individualization, Forensic Sci. Int. 266 (2016) 29–38, https://doi.org/10.1016/j.forsciint.2016.04.029.

[116] W. Thompson, J. Black, A. Jain, J. Kadane, AAAS Forensic Science Assessments: A Quality and Gap Analysis -Latent Fingerprint Examination, American Association for the Advancement of Science, 2017. https://www.aaas.org/resources/forensic-science-assessments-quality-and-gap-analysis?utm_content=buffer80012&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer.

[117] OSAC, Draft guidance on testing the performance of forensic examiners, human factors committee, organization of scientific area committees for forensic science. https://www.nist.gov/document/drafthfcguidancedocument-may8pdf, 2018.

[118] M. Saks, J. Koehler, The individualization fallacy in forensic science, Vand Law Rev 61 (2008) 199–219.

[119] C. Monturo, The effect of the machining process as it relates to toolmarks on surfaces, AFTE J 42 (3) (2010) 264–266.

[120] O. Ohar, T. Lizotte, Extracting ballistic forensic intelligence: microstamped firearms deliver data for illegal firearm traffic mapping – technology, in: F. Dickey, R. Beyer (Eds.), Optical Technology for Arming, Safing, Fuzing, and Firing V, SPIE, 2009, pp. 2–46.

[121] FDA, Guidance for industry: E9 statistical principles for clinical trials U.S. Food and Drug Administration. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9-statistical-principles-clinical-trials, 1998.

[122] T.A. Lang, D.F. Stroup, Who knew? The misleading specificity of "double-blind" and what to do about it, Trials 21 (1) (2020) 697, https://doi.org/10.1186/s13063-020-04607-5.

[123] J.J. Koehler, S. Liu, Fingerprint error rate on close non-matches, J. Forensic Sci. 66 (1) (2021) 129–134, https://doi.org/10.1111/1556-4029.14580.

[124] AFTE, Criteria for Identification Committee report, AFTE J 24 (3) (1992) 337–338.

[125] OSAC, Standard Scale of Source Conclusions and Criteria for Toolmark Examinations (Proposed), Firearms and Toolmarks Subcommittee, Organization of Scientific Area Committees for Forensic Science, 2020. https://www.nist.gov/document/range-source-conclusions-and-criteria-toolmark-examinations.

[126] I.E. Dror, G. Langenburg, Cannot decide": the fine line between appropriate inconclusive determinations versus unjustifiably deciding not to decide, J. Forensic Sci. 64 (1) (2019) 10–15, https://doi.org/10.1111/1556-4029.13854.

[127] S. Bunch, G. Wevers, Application of likelihood ratios for firearm and tool mark analysis, Sci. Justice 53 (2) (2013) 223–229, https://doi.org/10.1016/j.scijus.2012.12.005.

[128] K.E. Carter, M.D. Vogelsang, J. Vanderkolk, T. Busey, The utility of expanded conclusion scales during latent print examinations, J. Forensic Sci. 65 (4) (2020) 1141–1154, https://doi.org/10.1111/1556-4029.14298.

[129] C. Champod, C.J. Lennard, P. Margot, M. Stoilovic, Fingerprints and Other Ridge Skin Impressions, CRC Press, Boca Raton, 2016.

[130] D.H. Kaye, The Miami Dade bullet-matching study surfaces in United States v. Romero-Lobato, Forensic Science, Statistics & the Law. http://for-sci-law.blogspot.com/, 2019.

[131] AFTE, Range of conclusions, Association of Firearm and Tool Mark Examiners. https://afte.org/about-us/what-is-afte/afte-range-of-conclusions, 2020.

[132] General Requirements for Informed Consent, 45 CFR § 46.116, Department of Health and Human Services, United States, 2007.

[133] Protection of Human Subjects, 28 CFR § 46, Department of Justice, United States, 2016.

[134] R.B. Miller, C.S. Hollist, Attrition bias, in: N. Slalkind (Ed.), Encyclopedia of Measurement and Statistics, Sage Publications, Inc, Thousand Oaks, CA, 2007, pp. 57–60.

[135] K. Ahern, R. Le Brocque, Methodological issues in the effects of attrition: simple solutions for social scientists, Field Methods 17 (1) (2005) 53–69, https://doi.org/10.1177/1525822X04271006.

[136] J.C. Dumville, D.J. Torgerson, C.E. Hewitt, Reporting attrition in randomised controlled trials, Br. Med. J. 332 (7547) (2006) 969–971, https://doi.org/10.1136/bmj.332.7547.969.

[137] E.F. Law, K.B. Morris, C.M. Jelsema, Determining the number of test fires needed to represent the variability present within 9mm Luger firearms, Forensic Sci. Int. 276 (2017) 126–133, https://doi.org/10.1016/j.forsciint.2017.04.019.

[138] SWGGUN, Elimination Factors Related to FA/TM Examinations, 2011. https://www.nist.gov/document/guidelinesforeliminationfactorsrelatedtofa-tmexaminationspdf.

[139] E.J.A.T. Mattijssen, C.L.M. Witteman, C.E.H. Berger, N.W. Brand, R.D. Stoel, Validity and reliability of forensic firearm examiners, Forensic Sci. Int. 307 (2020) 110112, https://doi.org/10.1016/j.forsciint.2019.110112.

[140] A.K. Welch, History and manufacturing process of the Jennings/Bryco/Jimenez Arms pistols, AFTE J 45 (3) (2013) 260–266.

[141] A.K. Welch, Breech face subclass characteristics of the Jimenez JA Nine pistol, AFTE J 45 (4) (2013) 336–349.

[142] C. Monturo, Breech face marks of the Bryco Arms model Jennings Nine, AFTE J 31 (1) (1999) 67–69.

[143] L. Chumbley, J. Kreiser, T. Lizotte, O. Ohar, T. Grieve, B. King, D. Eisenmann, Clarity of microstamped identifiers as a function of primer hardness and type of firearm action, AFTE J 44 (2) (2012) 145–155.

[144] J. Davis, Primer cup properties and how they affect identification, AFTE J 42 (1) (2010) 3–22.

[145] E.D. Smith, J.L. Stephenson, Identification of Bullets Fired from Consecutively Manufactured Double-Broached Ruger® SR9c Barrels Utilizing Comparison Microscopy and Confocal Microscopy, AFTE Annual Training Seminar, New Orleans, 2016.

[146] S. Norris, Subclass Characteristics in Recent Ruger Handguns, AFTE Annual Training Seminar, Dallas, 2015.

[147] N. Jenzen-Jones, Chambering the next round, Small Arms Survey (2016). http://www.smallarmssurvey.org/fileadmin/docs/F-Working-papers/SAS-WP23-cartridge-technologies.pdf.

[148] Wolf Ammunition, Wolf Polyformance, Steel Cased Ammo, Wolf Performance Ammunition. http://wolfammo.com/steel-casing.aspx, 2020.

[149] S. Jacobs, History of Wolf Ammo, Ammo.Com, 2020. https://ammo.com/brands/wolf-ammo?ammo_casing=94#brand-history.

[150] T.B. Renegar, R.M. Thompson, A. Zheng, T. Vorburger, J. Song, J. Soons, J. Yen, An Improved Vacuum Casting Method for the Replication of Reference Bullet, National Institute of Standards and Technology, 2014. http://www.nist.gov/forensics/upload/Renegar.pdf.

[151] Interpol, Validated process for creating double casts, handbook on the collection and sharing of ballistics data. https://www.interpol.int/content/download/8151/file/14Y0277-INTERPOL-BALLISTICS-INFORMATION-EN.pdf, 2014, 61-81.

[152] A. Koch, H. Katterwe, Castings of complex stereometric samples for proficiency tests in firearm and tool mark examinations, AFTE J 39 (4) (2007) 299–306.

[153] E.F. Law, K.B. Morris, Three-dimensional analysis of cartridge case double-casts, J. Forensic Sci. 65 (6) (2020) 1945–1953, https://doi.org/10.1111/1556-4029.14549.

[154] Z. Geradts, K. Jan, C. Van Brakel, The production of replicas of bullets and cartridges, AFTE J 28 (1) (1996) 41–44.

[155] S. Bell, S. Sah, T.D. Albright, S.J. Gates, M.B. Denton, A. Casadevall, A call for more science in forensic science, Proc. Natl. Acad. Sci. Unit. States Am. 115 (18) (2018) 4541–4544, https://doi.org/10.1073/pnas.1712161115.

[156] F. Dong, Y. Zhao, Y. Luo, W. Zhang, K. Zhang, Specificity of characteristic marks on cartridge cases from 3070 consecutive firings of a Chinese Norinco QSZ-92 9 mm pistol, J Forensic Sci Med 5 (2) (2019) 87–94, https://doi.org/10.4103/jfsm.jfsm_6_19.

[157] G. Wevers, M. Neel, J. Buckleton, Comprehensive statistical analysis of striated toolmark examinations part 2: comparing known matches and known non-matches using likelihood ratios, AFTE J 43 (2) (2011) 137–145.

[158] E.J.A.T. Mattijssen, C.L.M. Witteman, C.E.H. Berger, X.A. Zheng, J.A. Soons, R.D. Stoel, Firearm examination: examiner judgments and computer-based comparisons, J. Forensic Sci. 66 (1) (2021) 96–111, https://doi.org/10.1111/1556-4029.14557.