Getting Uncertainty Right: Simulation Study of Bootstrap Methods for Clustered Forensic Firearm Examination Data

***Abstract*—Forensic firearm examination data often involve a complex structure in which multiple examiners evaluate overlapping sets of cartridge case comparisons (csets). This creates dependence at both the examiner and cset levels, which standard bootstrap confidence intervals do not account for and may therefore underestimate uncertainty. In this study, we compare four bootstrap methods: naive row resampling, examiner-cluster resampling, cset-cluster resampling, and a two-way combined bootstrap that accounts for both clustering levels simultaneously. Using a plasmode simulation based on real forensic data, we run 200 repetitions across 48 scenarios that vary case difficulty, evidence quality, and the presence of a Hawthorne effect. For each method, we evaluate empirical 95% confidence interval coverage and average interval width. The results show that naive bootstrap methods consistently undercover the true rates by 25-35 percentage points. Examiner-cluster resampling performs best for estimating error rates among one-way methods, while the two-way combined bootstrap achieves near-nominal coverage for both error rates and inconclusive rates across all scenarios, making it the most reliable approach overall. These findings offer practical guidance for more accurate uncertainty reporting in forensic firearm examination studies.**

I. Introduction

Forensic firearm examination involves comparing markings on bullets or cartridge cases to determine whether they originate from the same source. These decisions are typically reported as identification, exclusion, or inconclusive outcomes, and are often used in legal settings where their reliability is critical.

Large-scale studies, such as the Bullet Black Box Study, have made it possible to estimate examiner performance empirically, including both error rates and inconclusive rates across a range of case difficulties and evidence qualities. However, these data have a complex structure: multiple examiners evaluate overlapping sets of comparison items, creating dependence at both the examiner and item levels. In addition, because examiners are aware they are being evaluated, their behavior may be influenced by research participation effects, such as increased caution [3], [4].

These features create challenges for statistical inference. In particular, confidence intervals for error and inconclusive rates are often constructed using methods that assume independence across observations [5]. When this assumption is violated, uncertainty is underestimated, and reported intervals may appear more precise than they truly are [7].

This paper evaluates bootstrap-based methods for constructing confidence intervals in this setting. We focus on two key estimands: the inconclusive rate and the conclusive error rate. Using a simulation study based on models fit to real firearm examination data, we compare four methods: a naive bootstrap, an examiner-clustered bootstrap, a comparison-set-clustered bootstrap, and a two-way combined bootstrap that accounts for both clustering levels simultaneously [7]. Our goal is to determine which approach produces intervals that accurately reflect uncertainty in the presence of two-level clustering.

II. Literature Review

Existing research on forensic firearm examination highlights several key challenges in interpreting examiner performance, particularly in how error rates, inconclusive decisions, and study design influence reported results.

A central issue in this literature is the treatment of inconclusive decisions. Dorfman and Valliant [1] demonstrate that different ways of handling inconclusives, such as treating them as errors, excluding them, or analyzing them separately, can lead to substantially different conclusions about examiner performance. They argue that inconclusives should be viewed as a distinct outcome reflecting uncertainty or insufficient evidence, rather than as a failure of the examiner. This perspective motivates the present study's focus on modeling inconclusive rates separately from error rates and treating them as a primary quantity of interest.

Another important theme is the role of study design in shaping observed performance. Koehler [2] emphasizes that error rates in forensic science are highly sensitive to the composition of test items, including their difficulty, representativeness, and the balance of mated and non-mated comparisons. In particular, studies with more challenging or lower-quality evidence tend to produce higher rates of both inconclusive and error decisions. This highlights that observed performance is not solely a function of examiner ability, but also of the testing conditions under which decisions are made. In this study, this insight motivates varying difficulty and quality distributions within the simulation design.

While these studies provide important insights into examiner behavior and test design, they focus primarily on point estimation rather than uncertainty quantification. In practice, many forensic studies report confidence intervals constructed using methods that assume independent observations [5]. However, firearm examination data are inherently clustered, as multiple decisions are made by the same examiner and multiple examiners evaluate the same items. Ignoring this dependence

structure can lead to underestimated variability and overly narrow confidence intervals.

Simulation studies provide a natural framework for evaluating statistical methods under such complex data structures. Morris, White, and Crowther [6] emphasize that well-designed simulation studies should use realistic data-generating mechanisms and assess performance through metrics such as empirical coverage and interval width. Their ADEMP framework (Aims, Data-generating mechanism, Estimands, Methods, Performance measures) provides a structured approach to designing and reporting simulation-based research, and it informs the design of the present study.

Despite recognition of clustering in forensic datasets, there has been limited work systematically evaluating how bootstrap methods perform under this type of two-way dependence structure. In particular, it remains unclear whether resampling at the examiner level, item level, or both is necessary to produce confidence intervals with correct coverage.

This study addresses that gap by comparing alternative bootstrap strategies in a realistic simulation setting and by linking their performance to the underlying variance structure of examiner and item effects. In doing so, it provides practical guidance for constructing confidence intervals that more accurately reflect uncertainty in forensic firearm examination studies.

III. Data and Exploratory Analysis

A. Data and Study Background

This study uses data from the Bullet Black Box Study, in which 49 firearm examiners each evaluated a subset of 1,240 cartridge comparison sets (csets) under black-box, casework-like conditions, producing 3,156 total responses. Each examiner-cset pair resulted in one of three outcomes: identification (ID), exclusion (Excl), or inconclusive (Insuff). Because participants were told the exercise was a research and proficiency test, their behavior may reflect test awareness. For pattern-comparison tasks, this is expected to primarily be evident as increased caution, appearing as a higher rate of inconclusive decisions on difficult or ambiguous comparisons rather than as a shift in outright error rates. This motivates including a Hawthorne parameter in the simulation.

Two quantities are of central interest: the inconclusive rate, defined as the proportion of all decisions that are inconclusive, and the error rate, defined as the proportion of conclusive decisions that are incorrect. Accurate uncertainty quantification for both depends critically on the dependence structure of the data. The exploratory analysis below examines whether that structure is present and whether it operates at the examiner level, the cset level, or both.

B. Visualizing the Data

1) Case Difficulty Distribution: Figure 1 shows the distribution of case difficulty across all csets in the Bullet Black Box Study. The distribution is skewed toward Normal and Difficult cases, which indicates the observed aggregate rates are dominated by moderate-to-hard cases. To understand how bootstrap methods perform under a range of realistic test conditions, not just those mirroring this particular study, the simulation varies the case mix across four profiles: Easy, Medium, Difficult, and Very Difficult.

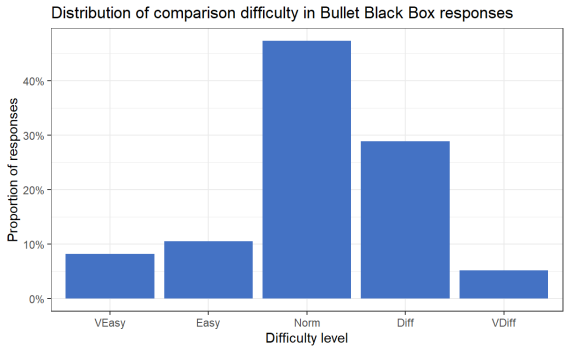

Fig. 1: Distribution of case difficulty in the Bullet Black Box Study. Most comparisons fall in the Normal or Difficult categories, motivating the simulation's four difficulty profiles.

2) Evidence Quality Distribution: Beyond difficulty, the quality of the evidence also shapes how often examiners call comparisons inconclusive or make errors. Figure 2 shows the joint distribution of difficulty and evidence quality in the Bullet Black Box Study. Within each difficulty band, items span the full range of quality grades, and the cells with the highest inconclusive and error rates are those that combine hard cases with low-quality evidence. Because quality and difficulty jointly determine the true parameter values the bootstrap is trying to cover, the simulation varies quality across four profiles, from predominantly high-quality (AA) to predominantly low-quality (CD), alongside the difficulty profiles.

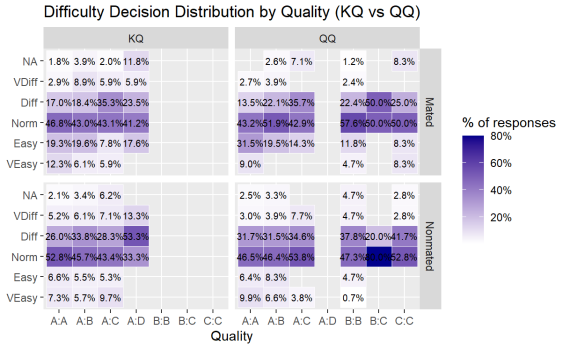

Fig. 2: Joint distribution of difficulty and evidence quality in the Bullet Black Box Study. The combination of high difficulty and low quality produces the highest rates of inconclusive and error decisions, motivating the simulation's four quality profiles.

3) Examiner-Level Clustering: Figure 3 shows the inconclusive rate for each of the 49 examiners, sorted in ascending order. The spread is wide: some examiners call fewer than 5% of their comparisons inconclusive, while others exceed 50%.

This dispersion is not noise, it shows persistent, examiner-specific tendencies that are consistent across items that persist across the different comparison items each examiner evaluates. A bootstrap that resamples individual decisions independently treats them as if they do not, which will systematically underestimate sampling variability.

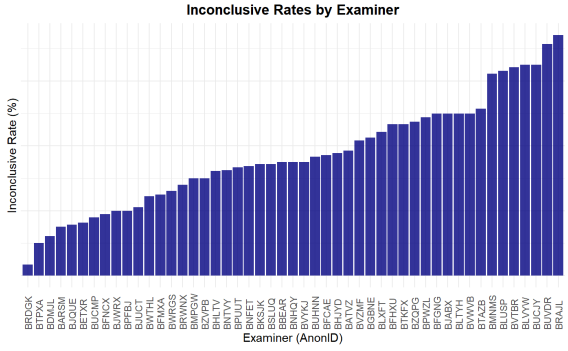

Fig. 3: Inconclusive rate by examiner, sorted in ascending order. The wide spread confirms strong examiner-level clustering in inconclusive decisions.

A parallel pattern holds for errors. Figure 4 shows each examiner's error rate among their conclusive decisions. Most examiners have error rates close to zero, indicating that errors are rare overall. However, a small but nontrivial subset of examiners have noticeably higher error rates, suggesting that error outcomes are not evenly distributed across examiners. Instead, the raw data show persistent between-examiner differences: some examiners contribute disproportionately to the observed errors, while most contribute few or none.

This pattern also differs from the pattern for inconclusive decisions. Whereas inconclusive rates vary meaningfully across both examiners and comparison sets, error rates appear to be driven more strongly by examiner-level differences. In other words, the main clustering structure for errors is examiner identity rather than item identity. This asymmetry between the two estimands becomes important later when evaluating which bootstrap method provides the most reliable uncertainty estimates for each outcome.

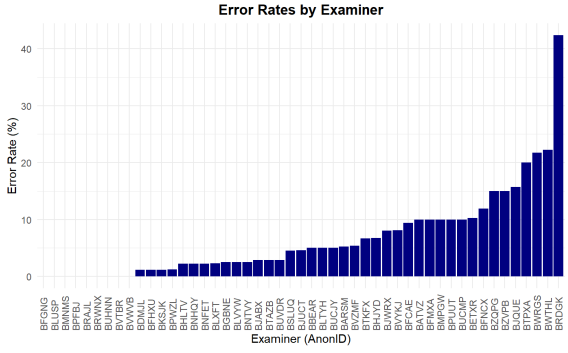

Fig. 4: Error rate by examiner among conclusive decisions. A subset of examiners have substantially higher error rates.

4) Comparison-Set-Level Clustering: Figure 5 shows the distribution of inconclusive and error rates across csets. For inconclusive decisions, the distribution is highly heterogeneous: many items are almost never called inconclusive, while others are consistently flagged as inconclusive by the majority of examiners who see them. This item-specific pattern is separate from examiner tendencies, it instead reflects the intrinsic differences in case difficulty and evidence quality that affect all examiners.

For error rates, variation across csets is more concentrated: most items have near-zero error rates, while a small subset accounts for a disproportionate share of observed errors. This pattern suggests that item-level differences exist, but they are sparse rather than broadly distributed across the test bank. In other words, error-rate variability is not primarily driven by consistent clustering at the cset level; instead, errors appear to be concentrated in a limited number of items and examiners.

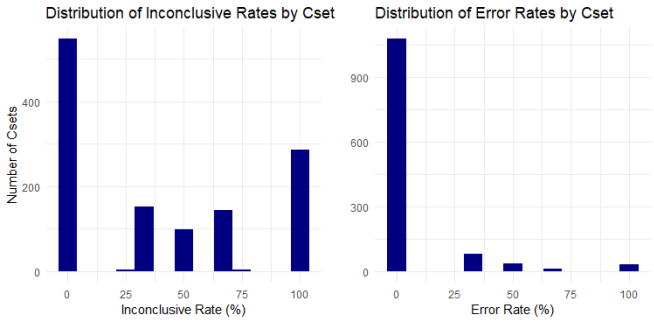

Fig. 5: Distribution of inconclusive rates (left) and error rates (right) across comparison sets. Wide item-level spread in inconclusive rates confirms meaningful cset-level clustering for that estimand.

Taken together, these four plots establish the key structural features of the data. Both examiners and comparison sets introduce clustering, but the dominant level differs by estimand: examiner-level clustering drives variability in the error rate, while both levels contribute to the inconclusive rate. A naive bootstrap that ignores this structure will produce intervals that are too narrow. Whether resampling at a single cluster level is sufficient, and which level to choose for each estimand, is precisely what the simulation study addresses.

IV. Methodology

A. Model

1) Overview: To simulate realistic forensic firearm examination data, we fit two mixed-effects logistic regression models to the Bullet Black Box Study data. The first model (Stage 1) predicts the probability that an examiner gives an inconclusive decision for a given comparison; the second model (Stage 2) predicts the probability of an error given that a conclusive decision was made. Together, these models define a two stage data-generating process that captures the key structural features of the original data.

Both models include fixed effects for case difficulty, examiner skill level, evidence quality (Quality1 and Quality2), item type (KQ or QQ), and a Hawthorne parameter H. The inconclusive model includes random intercepts for both examiners and comparison sets (csets). The error model includes only examiner-level random effects; item-level error variation could not be reliably estimated from the data ($\hat{\sigma}_{\text{cset,err}} \approx 0$), so no cset random effect was included.

2) Fitted Parameters and Variance Components: The estimated variance components from the two fitted models are:

- **Inconclusive model:** $\hat{\sigma}_{\text{exam,inc}} = 1.1677$, $\hat{\sigma}_{\text{cset,inc}} = 0.8605$
- **Error model:** $\hat{\sigma}_{\text{exam,err}} = 0.9655$, $\hat{\sigma}_{\text{cset,err}} \approx 0$

The large examiner-level standard deviation for inconclusive decisions (1.17 on the log-odds scale) indicates strong within-examiner dependence: decisions from the same examiner are highly correlated. The cset-level variance for inconclusive decisions is also substantial (0.861), indicating that both examiners and items are meaningful sources of clustering for this outcome. For error decisions, the dominant source of clustering is the examiner level; item-level variation is negligible. This asymmetry directly shapes which bootstrap methods perform best for each estimand, as discussed in Section V.

The strength of examiner-level dependence can be quantified through the intraclass correlation (ICC):

$$\text{ICC}_{\text{exam}} = \frac{\sigma^2_{\text{exam}}}{\sigma^2_{\text{exam}} + \sigma^2_{\text{cset}} + \pi^2/3},$$

where $\pi^2/3 \approx 3.29$ is the residual variance on the logistic scale. The large examiner-level variance implies that ignoring this structure systematically underestimates sampling variability, producing confidence intervals that are too narrow.

3) Two-Stage Decision Model: Each examiner-item pair (i, j) produces a decision through two sequential stages.

Stage 1 (Inconclusive or conclusive): The inconclusive indicator $Y_{ij,\text{inc}}$ is drawn as:

$$Y_{ij,\text{inc}} \sim \text{Bernoulli}(p_{ij,\text{inc}}),$$

where

$$\text{logit}(p_{ij,\text{inc}}) = X_{ij}\hat{\beta}_{\text{inc}} + u_{i,\text{inc}} + v_{j,\text{inc}} + 0.5 \cdot H.$$

Here X_{ij} encodes case difficulty, evidence quality, examiner skill, and item type; $\hat{\beta}_{\text{inc}}$ are the fixed-effect coefficients estimated from the real data; and the random intercepts are drawn as:

$$u_{i,\text{inc}} \overset{\text{iid}}{\sim} \mathcal{N}(0,\, \hat{\sigma}^2_{\text{exam,inc}}), \qquad v_{j,\text{inc}} \overset{\text{iid}}{\sim} \mathcal{N}(0,\, \hat{\sigma}^2_{\text{cset,inc}}).$$

The Hawthorne term $0.5 \cdot H$ adds a constant shift to the log-odds of an inconclusive decision. It is applied only in Stage 1, showing the expectation that heightened test awareness increases examiner caution-appearing as a higher rate of inconclusive decisions rather than a change in outright error rates [3], [4].

Stage 2 (Error or correct, if conclusive): If $Y_{ij,\text{inc}} = 1$ the process stops. If $Y_{ij,\text{inc}} = 0$, an error indicator is drawn as:

$$Y_{ij,\text{err}} \mid Y_{ij,\text{inc}} = 0 \sim \text{Bernoulli}(p_{ij,\text{err}}),$$

where

$$\text{logit}(p_{ij,\text{err}}) = Z_{ij}\hat{\beta}_{\text{err}} + u_{i,\text{err}}, \qquad u_{i,\text{err}} \overset{\text{iid}}{\sim} \mathcal{N}(0,\, \hat{\sigma}^2_{\text{exam,err}}).$$

No cset random effect is included, consistent with the near-zero item-level variance estimated from the data. The estimated coefficients and variance components serve as the fixed "true" parameters across all simulation scenarios.

B. Simulation Methodology

1) Overview and Motivation: This study uses a Monte Carlo simulation to evaluate four bootstrap methods for constructing confidence intervals for the inconclusive rate and the error rate. The simulation is structured using the ADEMP framework (Aims, Data-generating mechanism, Estimands, Methods, Performance measures) [6], allowing systematic assessment of empirical coverage and interval width across a range of realistic conditions.

The core challenge is that forensic firearm examination data have a two-way clustered structure, each examiner evaluates multiple items, and each item is evaluated by multiple examiners, which standard confidence intervals ignore. Ignoring this structure overstates the effective sample size and produces intervals that are systematically too narrow. We compare four bootstrap strategies: a naive bootstrap that treats observations as independent, an examiner-cluster bootstrap, a comparison-set (cset) cluster bootstrap, and a two-way combined bootstrap that accounts for both clustering levels simultaneously.

2) Objectives:

a) Primary: The primary objective is to identify which bootstrap method produces the most accurate empirical coverage for the inconclusive rate and the error rate. By "accurate coverage," we mean that a nominal 95% confidence interval should contain the true parameter value approximately 95% of the time across repeated studies. Coverage consistently below 90% indicates anti-conservative intervals that overstate precision; coverage consistently above 97-98% indicates unnecessarily wide intervals.

We expect the naive bootstrap to perform poorly because it ignores both levels of clustering. The examiner-cluster and cset-cluster bootstraps each account for one level of dependence, and their relative performance is expected to reflect which clustering level dominates for each estimand. The two-way combined bootstrap, which accounts for both levels simultaneously, is expected to provide the most reliable coverage overall.

b) Secondary:

To assess how mean confidence interval width varies across methods and scenarios, since the preferred method minimizes width while maintaining coverage near 95%.

1) To evaluate whether bootstrap performance is sensitive to scenario factors: case difficulty, evidence quality, and Hawthorne effect magnitude.
2) To compare method performance separately for the inconclusive rate and the error rate, since the two estimands have different underlying dependence structures.

3) Data-Generating Mechanism:

a) Synthetic Study Structure: Each simulated dataset represents one run of a proficiency study and mirrors the structure of the Bullet Black Box Study:

- 49 examiners ($i = 1, \ldots, 49$).
- 100 comparison items ($j = 1, \ldots, 100$): 50 known-questioned (KQ) and 50 questioned-questioned (QQ).
- Within each type, 50% of items are mated (same firearm) and 50% are non-mated, assigned randomly in each dataset.
- Not every examiner evaluates every item. The number of KQ items assigned to examiner i is drawn from a truncated normal distribution:

$$n_{i,\text{KQ}} \sim \text{TruncNormal}(\mu = 29.4,\ \sigma = 13.2,\ \min = 9,\ \max = 50)$$

  with the same distribution applied independently for QQ items. The parameters $\mu = 29.4$ and $\sigma = 13.2$ were estimated directly from the Bullet Black Box Study data to preserve the realistic variation in workload across examiners. A truncated normal was chosen because the observed assignment counts are approximately symmetric and bounded within a feasible range determined by the study design.
- Given $n_{i,k}$ assigned items of type $k \in \{\text{KQ}, \text{QQ}\}$, those items are sampled without replacement from the corresponding pool. This creates an incomplete crossed design in which most examiner-item pairs are observed, but examiners evaluate different numbers of items, an important feature for the bootstrap, since resampling examiners also resamples blocks of varying size.

Examiner and cset random effects are drawn from the distributions specified in Section IV at the start of each simulated dataset and held fixed for all decisions within that dataset, reflecting the assumption that individual examiner and item tendencies are consistent within a given study.

b) Scenario Factors: The simulation varies three factors, producing $4 \times 3 \times 4 = 48$ scenarios in total.

c) Factor 1: Case difficulty mix.: Each item is assigned a difficulty level from five ordered categories: Very Easy, Easy, Normal, Difficult, and Very Difficult. Four difficulty profiles were defined to span the full range from predominantly easy to predominantly very difficult testing conditions, allowing us to assess how bootstrap performance changes as the true rates shift across realistic study designs:

d) Factor 2: Hawthorne effect magnitude.: The Hawthorne effect is modeled as an additive shift on the log-odds of an inconclusive decision, as described in Section

Profile	VEasy	Easy	Normal	Difficult	VDiff
Easy	0.35	0.35	0.20	0.07	0.03
Medium	0.10	0.20	0.40	0.20	0.10
Difficult	0.03	0.07	0.15	0.55	0.20
VeryDifficult	0.02	0.05	0.13	0.25	0.55

TABLE I: Case difficulty distributions across profiles. Distributions were chosen to span the range from easy to very difficult testing conditions and to bracket the actual distribution observed in the Bullet Black Box Study (Figure 1).

Model. We consider three values: $H \in \{0.0, 0.2, 0.5\}$. No Hawthorne term is included in the error model, consistent with the expectation that test awareness primarily affects whether examiners commit to a conclusion, not whether that conclusion is correct once made [3], [4].

e) Factor 3: Evidence quality mix.: Each item is assigned two quality ratings (Quality1 and Quality2) drawn from categorical distributions. Four quality profiles were defined to span the range from high- to very-low-quality evidence observed in the Bullet Black Box Study. Since in the real study, there is a very few of D quality items presented, we choose to represent CC, and CS as two poor performing scenarios. Quality1 and Quality2 are paired within each scenario (AA, BB, CC, CD), yielding four quality combinations:

	Quality1			Quality2			
Profile	A	B	C	A	B	C	D
A (High)	0.80	0.15	0.05	0.75	0.15	0.07	0.03
B (Medium)	0.15	0.70	0.15	0.10	0.70	0.15	0.05
C (Low)	0.05	0.25	0.70	0.07	0.15	0.68	0.10
D (Very low)	0.05	0.20	0.75	0.05	0.10	0.15	0.70

TABLE II: Evidence quality distributions for Quality1 and Quality2. Each row sums to 1 within each quality measure.

f) Fixed factor: Examiner skill mix.: The distribution of examiner skill is fixed at $\mathbf{p}_{\text{skill}} = (0.25, 0.50, 0.25)$ for low, medium, and high skill levels across all scenarios, reflecting the approximate skill distribution observed in the Bullet Black Box Study.

The full scenario grid is shown in Table III.

4) Estimands:

a) Inconclusive Rate: Let $\mathcal{O}$ be the set of all observed examiner-cset pairs, with $N = |\mathcal{O}|$. The inconclusive rate is:

$$\hat{\theta}_{\text{inc}} = \frac{1}{N} \sum_{(i,j) \in \mathcal{O}} \mathbf{1}[Y_{ij,\text{inc}} = 1]\,.$$

Higher values reflect more difficult cases, lower-quality evidence, or greater caution due to the Hawthorne effect.

b) Error Rate: Let $\mathcal{C} = \{(i,j) \in \mathcal{O} : Y_{ij,\text{inc}} = 0\}$ be the set of conclusive decisions. The error rate is:

$$\hat{\theta}_{\text{err}} = \frac{\sum_{(i,j) \in \mathcal{C}} \mathbf{1}[Y_{ij,\text{err}} = 1]}{|\mathcal{C}|}.$$

Inconclusive decisions are excluded, as they are not errors. Because errors are relatively rare, this estimate is based on

Scenarios	Difficulty	Hawthorne H	Quality
01-12	Easy	0, 0.2, 0.5	A:A, B:B, C:C, C:D
13-24	Medium	0, 0.2, 0.5	A:A, B:B, C:C, C:D
25-36	Difficult	0, 0.2, 0.5	A:A, B:B, C:C, C:D
37-48	VDiff	0, 0.2, 0.5	A:A, B:B, C:C, C:D

TABLE III: Scenario structure for the simulation study.

fewer observations and is more variable than the inconclusive rate.

c) True Parameter Values via Monte Carlo Integration: For each scenario, the true values θ_{inc} and θ_{err} are approximated using $K = 50{,}000$ independent Monte Carlo draws:

$$\theta_{\text{inc}} \approx \frac{1}{K}\sum_{k=1}^{K} p_{\text{inc}}^{(k)}, \qquad \theta_{\text{err}} \approx \frac{\sum_{k=1}^{K} p_{\text{err}}^{(k)}\,(1-p_{\text{inc}}^{(k)})}{\sum_{k=1}^{K}(1-p_{\text{inc}}^{(k)})}.$$

With $K = 50{,}000$, Monte Carlo error is negligible, providing reliable benchmarks for evaluating bootstrap coverage.

5) Bootstrap Methods: We evaluate four bootstrap methods. In all cases, we draw $B = 300$ resamples, compute the statistic of interest on each, and form a 95% percentile interval:

$$\text{CI}_{0.95} = \left[\hat{\theta}^*_{(0.025)},\ \hat{\theta}^*_{(0.975)}\right].$$

The methods differ only in how resamples are constructed.

a) Naive Bootstrap: The naive bootstrap resamples individual observations (examiner-cset pairs) with replacement, keeping the total sample size fixed at N. This treats all observations as independent and ignores clustering, producing confidence intervals that tend to be too narrow. It is included as a baseline.

b) Examiner-Cluster Bootstrap: The examiner-cluster bootstrap resamples examiners rather than individual observations. At each iteration, 49 examiners are drawn with replacement, and all of their associated observations are included as a block. Duplicated examiners receive new identifiers. This captures between-examiner variability (σ^2_{exam}) but does not account for item-level dependence.

c) Comparison-Set-Cluster Bootstrap: The cset-cluster bootstrap resamples comparison items. At each iteration, 100 csets are drawn with replacement, and all observations for those items are included as a block. Duplicated csets receive new identifiers. This captures item-level variability (σ^2_{cset}) but not examiner-level dependence.

d) Two-Way Combined Bootstrap: Because the data are cross-classified by both examiners and comparison sets, neither one-way cluster bootstrap fully captures the dependence structure. To account for both clustering levels simultaneously, we implement the two-way combined bootstrap proposed by Cameron, Gelbach, and Miller (2011) [7]. This method combines the bootstrap distributions from the three methods above as:

$$\hat{\theta}^*_{\text{combined}} = \hat{\theta}^*_{\text{exam}} + \hat{\theta}^*_{\text{cset}} - \hat{\theta}^*_{\text{naive}},$$

where $\hat{\theta}^*_{\text{exam}}$, $\hat{\theta}^*_{\text{cset}}$, and $\hat{\theta}^*_{\text{naive}}$ denote bootstrap statistics from the examiner-cluster, cset-cluster, and naive bootstraps respectively. Intuitively, this adds the variation captured by each one-way bootstrap and subtracts the naive component to avoid double-counting the overlap. The 95% percentile interval is formed from the 2.5th and 97.5th quantiles of the combined distribution. This approach requires no additional data or modeling assumptions beyond those already used by the one-way methods.

6) Performance Measures:

a) Empirical Coverage Rate: For each scenario and method, we run $M = 200$ simulation repetitions. In each repetition we generate one dataset, construct a bootstrap confidence interval, and record whether it covers the true value:

$$C^{(m)} = \mathbf{1}\left[\hat{\theta}^{*(m)}_{(0.025)} \le \theta \le \hat{\theta}^{*(m)}_{(0.975)}\right].$$

The empirical coverage rate is:

$$\widehat{\text{cov}} = \frac{1}{M}\sum_{m=1}^{M} C^{(m)}.$$

Ideally this should be close to 0.95. With $M = 200$, deviations of approximately ± 0.02-0.03 are expected due to simulation noise.

b) Mean Confidence Interval Width: We also record the mean interval width:

$$\overline{W} = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}^{*(m)}_{(0.975)} - \hat{\theta}^{*(m)}_{(0.025)}\right).$$

Narrower intervals are preferred when coverage is maintained near 95%. Both measures are computed separately for the inconclusive rate and the error rate, across all methods and scenarios.

7) Computational Execution and Reproducibility: Due to the scale of the simulation (48 scenarios $\times$ 200 repetitions $\times$ 300 bootstrap resamples per repetition), all scenarios were run in parallel on a high-performance computing cluster using SLURM job arrays. Scenarios were divided into three groups of 16 and executed through separate scripts (`scenarios_1_16.R`, `scenarios_17_32.R`, `scenarios_33_48.R`). The two-way combined bootstrap was run through a corresponding set of scripts (`scenarios_1_16_combined.R`, etc.). Each job handled a single scenario and saved results independently to an `.rds` file.

Reproducibility was ensured by setting the random seed at the start of each repetition using `set.seed(42+m)`.

All model parameters and simulation settings are defined in `simulation_functions.R`, so the full study can be reproduced without external data. Code is available at https://github.com/panditk455/StatsCOMPSGroup1.

V. Results

We evaluated the empirical coverage and average confidence interval (CI) width of four bootstrap methods across 48 simulated scenarios varying difficulty, evidence quality, and Hawthorne-effect magnitude. The four methods were the naive bootstrap, examiner-clustered bootstrap, cset-clustered bootstrap, and two-way combined bootstrap. The nominal target coverage was 95%.

Figure 6 summarizes the key tradeoff across all methods and scenarios: it plots empirical coverage against mean CI width, where the upper-left corner represents the ideal-narrow intervals with correct coverage. The naive bootstrap clusters in the lower-left, confirming that its narrow intervals come at the cost of severe undercoverage. The combined bootstrap (purple) is the only method that consistently reaches the 95% target for both estimands.

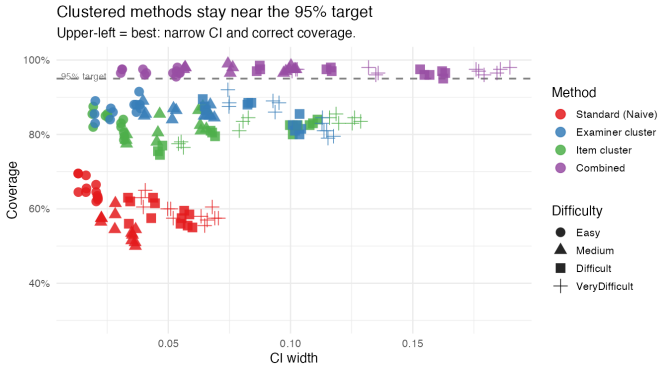

Fig. 6: Coverage vs. CI width for all four bootstrap methods across 48 scenarios. Upper-left is best: narrow CI with correct coverage. The combined method (purple) is the only one that reaches the 95% target consistently for both estimands.

Across all scenarios, the naive bootstrap showed serious undercoverage for both estimands. For error rates, the average empirical coverage was 59.6%, with a range from 50.0% to 69.5%. For inconclusive rates, performance was even worse, with an average coverage of 38.0% and a range from 29.5% to 44.0%. Although the naive bootstrap produced the narrowest confidence intervals (average width 0.038 for error rate and 0.032 for inconclusive rate), these intervals were far too narrow to provide reliable coverage. This confirms that treating examiner, cset responses as independent observations greatly underestimates uncertainty, a problem directly relevant to existing black-box studies that have adopted naive resampling as their default [5].

For error-rate estimation, the examiner-clustered bootstrap performed best among the one-way methods, with an average coverage of 85.7% (range: 79.0%-92.0%), compared with 81.9% (range: 74.5%-87.5%) for the cset-clustered bootstrap. Even so, the examiner-clustered method never reached the 95% target, with a maximum observed coverage of 92.0%. For inconclusive-rate estimation, the pattern reversed: the cset-clustered bootstrap performed best among one-way methods, achieving an average coverage of 86.1% (range: 77.5%-94.0%), compared with 70.2% for the examiner-clustered bootstrap. This outcome-specific difference shows the underlying variance structure, examiner-level clustering dominates error-rate variability, while both levels contribute to inconclusive-rate variability. The detailed breakdown across all 48 scenarios is shown in Figures 7 and 8.

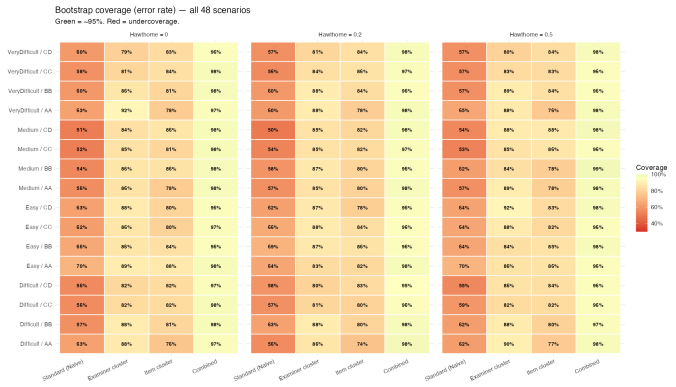

Fig. 7: Bootstrap coverage for error rate across all 48 scenarios and four methods. Green = 95% nominal. Red = undercoverage. The combined method is the only one that consistently reaches the target.

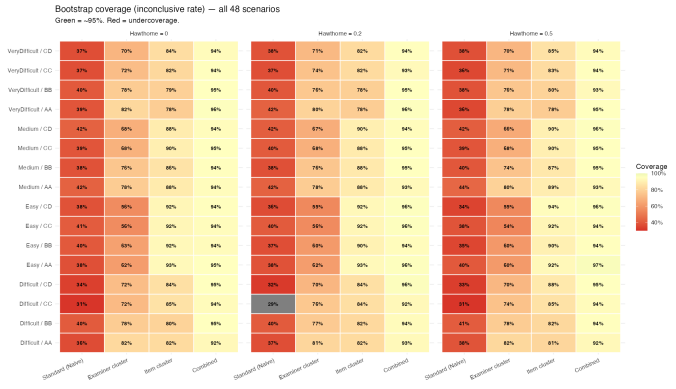

Fig. 8: Bootstrap coverage for inconclusive rate across all 48 scenarios and four methods. The combined method stays near 95% across all conditions; the examiner-clustered bootstrap is particularly weak for this estimand.

The two-way combined bootstrap substantially outperformed all one-way methods on both estimands. For error rates, it achieved an average coverage of 97.1% (range: 95.0%-99.0%) and had the highest coverage in all 48 scenarios. For inconclusive rates, it achieved an average coverage of 94.4% (range: 92.5%-97.0%), again the highest across all scenarios. In this way, the combined bootstrap was the only method to consistently achieve near-nominal coverage for both estimands

simultaneously, confirming that accounting for both clustering levels is necessary to produce reliable confidence intervals in this setting.

The Hawthorne effect had little impact on coverage across all methods. For the combined bootstrap, average error-rate coverage was 97.3%, 96.9%, and 97.2% for $H = 0$, 0.2, and 0.5 respectively, and inconclusive-rate coverage was 94.3%, 94.4%, and 94.4%. Similar stability was observed for the one-way methods. These results show that the choice of bootstrap method matters far more than the size of the Hawthorne effect.

The combined bootstrap produced the widest confidence intervals, with average widths of 0.102 for error rates and 0.138 for inconclusive rates, compared with 0.069 and 0.075 for examiner-clustered intervals, and 0.064 and 0.109 for cset-clustered intervals. This reflects the cost of correctly accounting for both clustering levels. The narrow intervals from the naive bootstrap (0.038 and 0.032) came with severe undercoverage and should not be interpreted as greater precision.

Overall, the simulation results show that bootstrap confidence intervals must account for the two-level clustering structure in forensic comparison data. The naive bootstrap consistently underestimates uncertainty and should not be used for either estimand. Among one-way methods, the examiner-clustered bootstrap works better for error rates and the cset-clustered bootstrap for inconclusive rates, but neither reaches the 95% target. The two-way combined bootstrap provides near-nominal coverage for both outcomes across all 48 scenarios and is the most reliable method for this type of clustered forensic data.

VI. Discussion and Conclusion

A. Discussion

1) Interpreting the Results: The simulation results clearly show that ignoring clustering leads to severely anti-conservative confidence intervals [7], [8]. However, they also reveal a more subtle pattern: the most important source of clustering depends on the estimand.

Examiner-level resampling works best for error-rate inference because, once inconclusive decisions are removed, the remaining variation in correct versus incorrect decisions is mainly driven by consistent examiner behavior, such as skill, risk tolerance, and decision thresholds [1]. In contrast, cset-level resampling works best for inconclusive-rate inference because the decision to call a comparison inconclusive is strongly influenced by item-level properties such as difficulty and evidence quality, which are shared across all examiners who evaluate the same item [2].

Even so, the best-performing single-level methods did not reach the nominal 95% coverage target. The examiner bootstrap achieved a maximum of 92.0% coverage for error rates, while the cset bootstrap reached a maximum of 94.0% for inconclusive rates. This gap reflects the fact that both clustering levels operate simultaneously [7]. A method that resamples only examiners captures between-examiner variability but ignores remaining correlation across items, and a method that resamples only csets has the opposite limitation.

To address this, we implemented the two-way combined bootstrap proposed by Cameron, Gelbach, and Miller [7], which combines bootstrap distributions as $\hat{\theta}^*_{\text{combined}} = \hat{\theta}^*_{\text{exam}} + \hat{\theta}^*_{\text{cset}} - \hat{\theta}^*_{\text{naive}}$. This method achieved near-nominal coverage for both estimands across all 48 scenarios, confirming that accounting for both clustering levels simultaneously is necessary and sufficient for reliable inference in this setting.

Another important finding is that coverage remained consistent across Hawthorne levels ($H \in \{0, 0.2, 0.5\}$) for all four methods. This indicates that the Hawthorne effect, as modeled here, changes the true inconclusive rate but does not affect the relative performance of the bootstrap methods [3], [4]. In other words, the choice of resampling unit, not the magnitude of test awareness, is the key factor driving coverage performance.

2) Real-World Implications: These findings have direct implications for how uncertainty is reported in forensic firearm examination studies. Many large-scale black-box studies have either reported only point estimates or constructed confidence intervals using methods that treat observations as independent [5]. Our results show that such naive approaches consistently underestimate uncertainty, sometimes by 30-40 percentage points [1], [5].

As a result, reported intervals such as "the error rate is 1% (95% CI: 0.6%-1.4%)" may give a false impression of precision if clustering is not accounted for [2]. In courtroom and policy settings, this can be misleading: a narrow confidence interval may suggest that the error rate is known very precisely, when in reality the correct interval would be much wider once examiner and cset dependence are properly included.

The practical recommendation from this study is to use the two-way combined bootstrap as the default for both estimands. It is straightforward to implement, requires no additional data or modeling beyond what is already used in standard studies [7], and reliably achieves near-nominal coverage regardless of which clustering level dominates. Where computational constraints require a simpler approach, the examiner-clustered bootstrap is the best fallback for error-rate estimation and the cset-clustered bootstrap for inconclusive-rate estimation, but both should be understood as underestimates of the true uncertainty.

3) Limitations: Several limitations should be acknowledged in this study. First, this is a plasmode simulation study [6], meaning that the parameters are based on a single dataset (the Bullet Black Box Study). As a result, the findings may not generalize to settings with very different examiner populations, item characteristics, or laboratory conditions. In particular, the near-zero cset-level variance for errors ($\hat{\sigma}_{\text{cset,err}} \approx 0$) is specific to this dataset; studies with more item-level variation in error rates may show different patterns.

Second, the simulation assumes that the two-stage logistic model is correctly specified. In practice, the true data-generating process may include non-logistic behavior, interactions between examiner skill and item difficulty, or learning effects over time [1]. While this model provides a realistic

and tractable approximation, deviations from it could affect coverage results.

Third, the study size is fixed at 49 examiners and 100 csets. Bootstrap methods can become less stable with fewer clusters [7], so performance in smaller studies may differ.

Finally, we evaluated only the percentile bootstrap [9]. Other approaches, such as bias-corrected and accelerated (BCa) [10] or studentized bootstrap intervals, may provide improved performance in some settings and should be explored in future work.

B. Conclusion

This study evaluated four nonparametric bootstrap methods for constructing confidence intervals for error rates and inconclusive rates in forensic firearm examination data. Using a plasmode simulation [6] based on the Bullet Black Box Study across 48 scenarios, we found that the naive bootstrap severely underestimates uncertainty in all scenarios and should not be used for either estimand [5].

Among one-way methods, the examiner-clustered bootstrap is preferable for error-rate inference and the cset-clustered bootstrap for inconclusive-rate inference, consistent with the underlying variance structure [1]. However, neither achieves the nominal 95% coverage target, reflecting the fact that both clustering levels are present simultaneously [7].

The two-way combined bootstrap [7] achieves near-nominal coverage for both estimands across all 48 scenarios and is the recommended approach for forensic firearm examination data with a two-level clustered structure. These findings have direct implications for practice: current naive-bootstrap intervals systematically overstate precision, and replacing them with the combined bootstrap would yield more accurate and honest uncertainty estimates for both scientific reporting and legal decision-making [2].

Future work should explore the performance of the combined bootstrap under different study sizes, particularly with fewer examiners, extend the approach to other forensic disciplines such as fingerprint analysis [11], investigate alternative interval types such as BCa [10] or studentized bootstraps, and validate bootstrap performance against observed data rather than simulated replications.

VII. Generative AI Statement

Generative AI tools were used during the coding part and the writing part of this manuscript as follows:

- **Google Gemini** was used to assist in editing and improving conciseness in the Abstract, Introduction, and Limitations sections.
- **OpenAI** was used for code debugging, advice on configuring SLURM cluster scripts for the parallel computing, identifying suitable data visualization styles, and initial literature search.

All statistical models, simulation implementations, final literature selections, empirical analyses, and manuscript conclusions were independently conducted, authored, and verified by the authors, who take full responsibility for the content of this manuscript.

References

[1] A. H. Dorfman and R. Valliant, "Inconclusives, errors, and error rates in forensic firearms analysis: Three statistical perspectives," *Forensic Science International: Synergy*, vol. 4, p. 100239, 2022.

[2] J. J. Koehler, "Proficiency tests to estimate error rates in the forensic sciences," *Law, Probability and Risk*, vol. 12, no. 1, pp. 89–98, 2013.

[3] N. Scurich *et al.*, "The Hawthorne effect in studies of firearm and toolmark examiners," *Journal of Forensic Sciences*, vol. 70, no. 4, pp. 1329–1337, 2025.

[4] J. McCambridge, J. Witton, and D. R. Elbourne, "Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects," *Journal of Clinical Epidemiology*, vol. 67, no. 3, pp. 267–277, 2014.

[5] R. A. Hicklin *et al.*, "Accuracy and reproducibility of bullet comparison decisions by forensic examiners," *Forensic Science International*, 2024.

[6] T. P. Morris, I. R. White, and M. J. Crowther, "Using simulation studies to evaluate statistical methods," *Statistics in Medicine*, vol. 38, no. 11, pp. 2074–2102, 2019.

[7] A. C. Cameron, J. B. Gelbach, and D. L. Miller, "Robust inference with multiway clustering," *Journal of Business & Economic Statistics*, vol. 29, no. 2, pp. 238–249, 2011.

[8] K.-Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986.

[9] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

[10] B. Efron, "Better bootstrap confidence intervals," *Journal of the American Statistical Association*, vol. 82, no. 397, pp. 171–185, 1987.

[11] S. A. Cole, "More than zero: Accounting for error in latent fingerprint identification," *Journal of Criminal Law and Criminology*, vol. 95, no. 3, pp. 985–1078, 2005.

VIII. GITHUB LINK FOR PROJECT:

[redacted]