جامعة نايف العربية للعلوم الأمنية
NAIF ARAB UNIVERSITY FOR SECURITY SCIENCES
تأسست ١٩٧٨ Est. 1978

Naif Arab University for Security Sciences
Arab Journal of Forensic Sciences and Forensic Medicine
المجلة العربية لعلوم الأدلة الجنائية والطب الشرعي
https://journals.nauss.edu.sa/index.php/AJFSFM

الجمعية العربية لعلوم الأدلة الجنائية والطب الشرعي
Arab Society for Forensic Sciences and Forensic Medicine

An Update on a World Wide Study of Bullets Fired From 10 Consecutively Rifled 9mm Ruger Pistol Barrels: Analysis of Examiner Error Rate Using Bayesian Statistics

تحديث على دراسة عالمية للطلقات النارية التي تم إطلاقها من عشر سبطانات محززة لمسدس روجر عيار 9 مم بصورة متتالية: تحليل لمعدل خطأ الفاحص باستخدام الإحصاءات الاحتمالية التحديثية

James E. Hamby[1*], Eric M. Warren[2], David J. Brundage[3], Nicholas D. K. Petraco[4,5], James W. Thorpe[6].

[1]International Forensic Science Laboratory and Training Centre, New Port Richey, FL, USA.

[2]SEP Forensic Consultants, Memphis, TN, USA.

[3]Independent Examiner, Franklin, TN, USA.

[4]Department of Sciences, John Jay College of Criminal Justice, City University of New York, New York, NY, USA.

[5]Faculty of Chemistry and Faculty of Criminal Justice, Graduate Center, City University of New York, New York, NY, USA.

[6]University of Strathclyde, Department of Pure and Applied Chemistry, Forensic Science Division, Glasgow, Scotland, UK..

Received 25 Aug. 2023; Accepted 07 June 2024; Available Online 25 July 2024.

Abstract

This technical note is the latest update on a continuing study, first designed and initiated by Brundage et al. over twenty years ago. This study was borne out of increased scrutiny of firearms identification in response to Daubert v. Merrell Dow Pharmaceuticals, Inc. in 1993. The purpose was to determine whether forensic firearms examiners were able to associate fired bullets with the barrels through which they had been fired. To date 792 participants from 36 countries have utilized over 240 test sets consisting of bullets fired through 10 consecutively rifled Ruger P-85 pistol barrels. Here we provide an update on the results of the ongoing "10 barrel test" up until the point in time of writing. To analyze the data thus far collected, a Bayesian approach was again selected. Posterior examiner error rates are estimated assuming only vague prior information. Given the data found over the course of this diverse decades long study, our most conservative estimate for examiner error rate has a posterior median of 0.03% with a 95% probability interval of [2×10^{-6} %, 0.1%].

Keywords: Forensic sciences, consecutively rifled barrels, criteria for identification, Daubert, firearms identification, fired bullets, ballistics imaging instrumentation.

المستخلص

هذه المذكرة الفنية هي آخر تحديث لدراسة مستمرة، صممها وأطلقها لأول مرة بروندج وآخرون منذ أكثر من عشرين عامًا. نشأت هذه الدراسة من التدقيق المتزايد على تحديد الأسلحة النارية استجابة لقضية داوبرت ضد ميريل داو فارماسوتيكالز، Inc. عام 1993. كان الهدف هو تحديد ما إذا كان فاحصو الأسلحة النارية الجنائيون قادرين على ربط الرصاصات التي تم إطلاقها بالبراميل التي تم إطلاقها من خلالها. حتى الآن، شارك 792 مشاركًا من 36 دولة في أكثر من 240 مجموعة اختبار تتكون من رصاصات تم إطلاقها عبر 10 براميل متتالية لبنادق روجر بي 85. نقدم هنا تحديثًا على نتائج «اختبار البراميل العشرة» المستمر حتى وقت كتابة هذا التقرير. لتحليل البيانات التي تم جمعها حتى الآن، تم اختيار النهج البايزي مرة أخرى. يتم تقدير معدلات خطأ الفاحص الخلفي بافتراض وجود معلومات أولية غامضة فقط. بالنظر إلى البيانات التي تم العثور عليها على مدار هذه الدراسة المتنوعة التي استمرت عقودًا، فإن تقديرنا الأكثر تحفظًا لمعدل خطأ الفاحص له وسط احتمالي يبلغ 0.03٪ مع فاصل احتمال 95٪ يبلغ [2 × 10^{-6}٪، 0.1٪].

الكلمات المفتاحية: علوم الأدلة الجنائية، سبطانات متتالية الحزوزات، معايير التعريف، داوبر، تحديد الأسلحة النارية، الرصاصات التي تم إطلاقها، أدوات تصوير المقذوفات.

* Corresponding Author: James E. Hamby

Email: jimhamby14@aol.com

doi: 10.26735/SMIK6534

1. Introduction

Current practices in firearm and toolmark identification training and actual laboratory casework are based on the hypothesis that fired bullets can be positively associated with the gun that fired them [1]. It is recognized that striations are caused by microscopic imperfections in the rifling tools used to make gun barrels during the manufacturing process. The tools used to manufacture firearms change during their use and therefore impart a continually changing set of striations on the items manufactured [2]. It would therefore be expected, that the greatest amount of similarity (and thus the greatest chance for identification error) would be encountered with firearms that are consecutively rifled using the same rifling tool [3]. Previous studies addressing this fundamental topic [4-7] will not be discussed further here.

The current work is an update of the previous expansion [6] of the Brundage study [4]. In this update we estimate the rate at which examiners correctly associate fired bullets with the barrels through which they have been fired, given those barrels were consecutively manufactured. The statistical model used, which was first proposed by Schuckers [8], takes into account our prior ignorance about the rate at which examiners commit identification errors and combines it with a sample of examiner test results. The model also takes into account possible correlations between the "match" and "no-match" conclusions examiners render for each bullet/barrel pair in the test. A posterior estimate of examiner error rate was produced which helped to quantify an answer to the question: Can projectiles fired from consecutively manufactured gun barrels be correctly associated with the barrel through which they passed most of the time?

2. Methods

The procedure and test design are briefly discussed here, for additional information on the design of the study in the context of the statistical analysis presented, see previous literature [5]. One Ruger P-85 9mm caliber semiautomatic pistol (manufactured in 1990), serial number: 302-06291 with one 15-cartridge capacity magazine was used to generate all samples to be examined. Ten consecutively rifled 9mm caliber barrels manufactured by Ruger for their P-85 pistol were collected. Twenty thousand Winchester 9mm caliber NATO, 124 grain FMJ cartridges, lot number: Q4312, Head stamp: WCC96 were used to generate test samples, as well as a selection of vintage 9mm Luger cartridges manufactured in Canada during WWII. The pistol with numbered slide was test fired into a vented 800 gallon water recovery tank, located in the firearms section of the Indianapolis-Marion County Forensic Services Agency (IMCFSA), Indianapolis, Indiana and the samples were engraved with a unique identification code. The test samples were then placed into envelopes and mailed out in protective packaging to prevent handling damage.

2.1. Construction of Test Sets

The test was set up as a "closed set" test where all of the 15 unknown bullets were fired in one of the 10 consecutively rifled barrels, with at least one bullet from each barrel and no more than three from any one barrel. These unknown bullets were packaged with a control set consisting of two bullets fired from each of the 10 barrels. A total of 240 such test sets were prepared.

Test firing commenced on July 8, 1999 and concluded on August 10, 2000. Production of the test ultimately involved shooting 16,800 cartridges; 1,680 from each of the 10 consecutively

Table 1- Combined results of the previous Brundage study and this study.

Test Series	# Examiners Participating in Test	# Examiners Reporting Inconclusives	#Inconclusively Identified Bullets	#Incorrectly Identified Bullets
Brundage	67	1	1	0
Hamby	725	4	7	0
Totals	792	5	8	0

manufactured barrels. All 16,800 cartridge cases were test fired using the same slide installed on the Ruger P-85 semiautomatic pistol.

Seven cartridges were test fired for each test sequence and combined into 'groups' by barrel and firing sequence number in order to allow for the same relative amount of barrel wear on both the control and the unknown bullets.

The test sets were individually packaged according to the sequence of the test set being fired and continued until all 240 test sets were completed. A 10% random sampling of the 240 prepared sets was conducted before the sets were shipped to participants. The random sample was examined with an optical comparison microscope, ensuring that there were enough surface features such that it was feasible to identify the 15 'unknowns' to the 'known' bullets. Next, the samples were mailed out in padded envelopes with instructions and a blank answer form. Table 1 lists the number of examiners who took the test in this study along with counts of the inconclusive and incorrect identifications they rendered.

2.2. Distribution of Test Sets

Although all of the original 240 test sets have been distributed to forensic laboratories, universities and researchers around the world, twenty additional test sets were recently constructed, composed of the Canadian manufactured WWII ammunition as described above. Ten of these sets were sent to European forensic laboratories, and 10 sets were distributed throughout the United States. Additionally several polymer clone sets, using the double cast method previously described [9,10], have been distributed.

3. Statistical Model

We are interested in estimating the probability that an examiner will reach the conclusion that there is a match between a bullet and a barrel, when in fact that is not the case. Under the design of this study, this is a false discovery rate (FDR). To illustrate the FDR concept (and understand how inconclusive opinions can be handled) consider the standard 2-by-2 contingency table known from hypothesis testing:

A "discovery" for FDR is a rejection of the null hypothesis. A "discovery" in our case is the event of an examiner rendering an opinion of a "match" between a bullet and a barrel with the null hypothesis of "no match" implied. A false discovery [11,12] is an opinion of match rendered in a comparison when in fact a bullet was fired through a different barrel than that concluded by the examiner. The false discovery rate is, on average, how often we can expect this error to happen [11]:

$$\mathbb{E}\left[\frac{V}{R \vee 1}\right].$$

False discovery rate in the context of this study is an expression of *examiner error rate*, which

is how we will refer to this statistic throughout the paper. When error rates are small it turns out that it can be difficult to estimate them precisely. Frequentist based methods are known to perform poorly in this situation [8,13,14]. Thus we have opted to take a Bayesian approach from which we may infer a reasonable estimate of *examiner error rate*, π_{eer}, given the data observed in this study [8,13]. Below we describe the model framework, due to Schuckers, which has been shown to render reasonable estimates for π_{eer}, even when they are very small.

A Bayesian technique takes what is "known" or "believed" about an unknown parameter (examiner error rate, π_{eer}, in our case) and represents it as a prior probability distribution $p(\pi_{eer})$. When the data (s) is measured, all the information it contains about π_{eer}'s value is contained in its likelihood function, or "probability model" for the data, $p(s|\pi_{eer})$.

Each time an examiner renders an opinion of "match" they can be correct or incorrect. We treat the outcome as a Bernoulli random variable, x_i, which can take on the value of 0 or 1. That is, for the i^{th} unknown bullet ($i \in \{1, \ldots, nj\}$), $\chi_i = 0$ if the examiner makes the correct "match" and $\chi_i = 1$ if the examiner makes an incorrect "match". In symbols:

$$x_i = \begin{cases} 1, & \text{if examiner renders incorrect ID} \\ 0, & \text{if examiner renders correct ID} \end{cases} \qquad x_i \sim \text{Bernoulli}(\pi_{\text{eer}})$$

The actual data analysed will be the sum of the n_j Bernoulli random variables constituting the outcome of the test for each examiner. To make this more explicit, let $\chi_{i,j}$ represents the outcome for the j^{th} examiner rendering an opinion on the ith unknown bullet. Then s_j is a random variable representing the number of wrong IDs rendered by the j^{th} examiner:

$$s_j = \sum_{i=1}^{n_j} x_{i,j}$$

If an examiner renders an opinion of match or no-match for each bullet to a barrel, then $n_j = 15$. For this study however, examiners were not barred from rendering an opinion of inconclusive. Because inconclusive is neither correct nor incorrect, this outcome affects the total number of possible positive match opinions an examiner *could* render on the test. That is, inconclusive opinions affect max(R) (cf. Table 2). Thus if an examiner renders one or more inconclusive opinions $n_j < 15$.

Data for this study are the number of errors each examiner made, s_j, organized into a vector of length 792, s. Often sums of Bernoulli outcomes are modelled as arising from a Binomial distribution. However in our case, there can be correlation between the 15 matching attempts (Bernoulli trials) each examiner undertakes; i.e. an examiner's answer on one trial may affect their answer on another trial. Modelling the data with only the binomial distribution would not take the correlation into account.

The Beta-binomial distribution is a generalization of the binomial distribution that naturally accounts for correlation between Bernoulli trials, and is the likelihood we will use to model the number of errors in identification each examiner commits [8]:

$$p(s_j|\alpha,\beta,n_j) = \binom{n_j}{s_j} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+s_j)\Gamma(\beta+n_j+s_j)}{\Gamma(\alpha+\beta+n_j)}.$$

The first term on the right of the equality sign is the binomial coefficient and the G's are gamma functions. There were 792 examiners who contributed to the data set, and each examiner underwent n_j, possibly correlated, Bernoulli trials. Thus the data for this study s, is modelled as a product of Beta-binomial distributions:

$$\mathbf{s} \sim \prod_{j=1}^{697} \text{Beta-binomial}(\alpha,\beta,n_j) = \prod_{j=1}^{697} \binom{n_j}{s_j} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+s_j)\Gamma(\beta+n_j+s_j)}{\Gamma(\alpha+\beta+n_j)}$$

The Beta-binomial distribution, is (usually) parameterized in terms of two new parameters, $\alpha\beta$, instead of π_{eer}. From α and β we can recover the examiner error rate, π_{eer}, as:

$$\pi_{\text{eer}} = \frac{\alpha}{\alpha + \beta}.$$

Correlation between the n_j decisions made by the j^{th} examiner, C_j, is modeled as:

$$C_j = \frac{\alpha + \beta + n_j}{\alpha + \beta + 1}.$$

The correlation coefficient C_j is on a scale from 1 (no correlation between Bernoulli trials) to n_j (full correlation between Bernoulli trials). We can also write the correlation for this model on a more familiar 0-1 scale as:

$$\phi = \frac{1}{\alpha + \beta + 1} = \frac{C_j}{\alpha + \beta + n_j}$$

For ϕ = 0 (C_j = 1) we recover the Binomial distribution from the Beta-Binomial distribution. Full correlation between Bernoulli trials corresponds to ϕ = 1. For the remainder of this paper, we will use the ϕ coefficient as our measure for correlation.

The prior knowledge concerning peer can be updated with the likelihood, p(s lpeer), via Bayes' theorem:

$$p(\pi_{\text{eer}}|\mathbf{s}) = \frac{p(\mathbf{s}|\pi_{\text{eer}})\ p(\pi_{\text{eer}})}{p(\mathbf{s})}$$

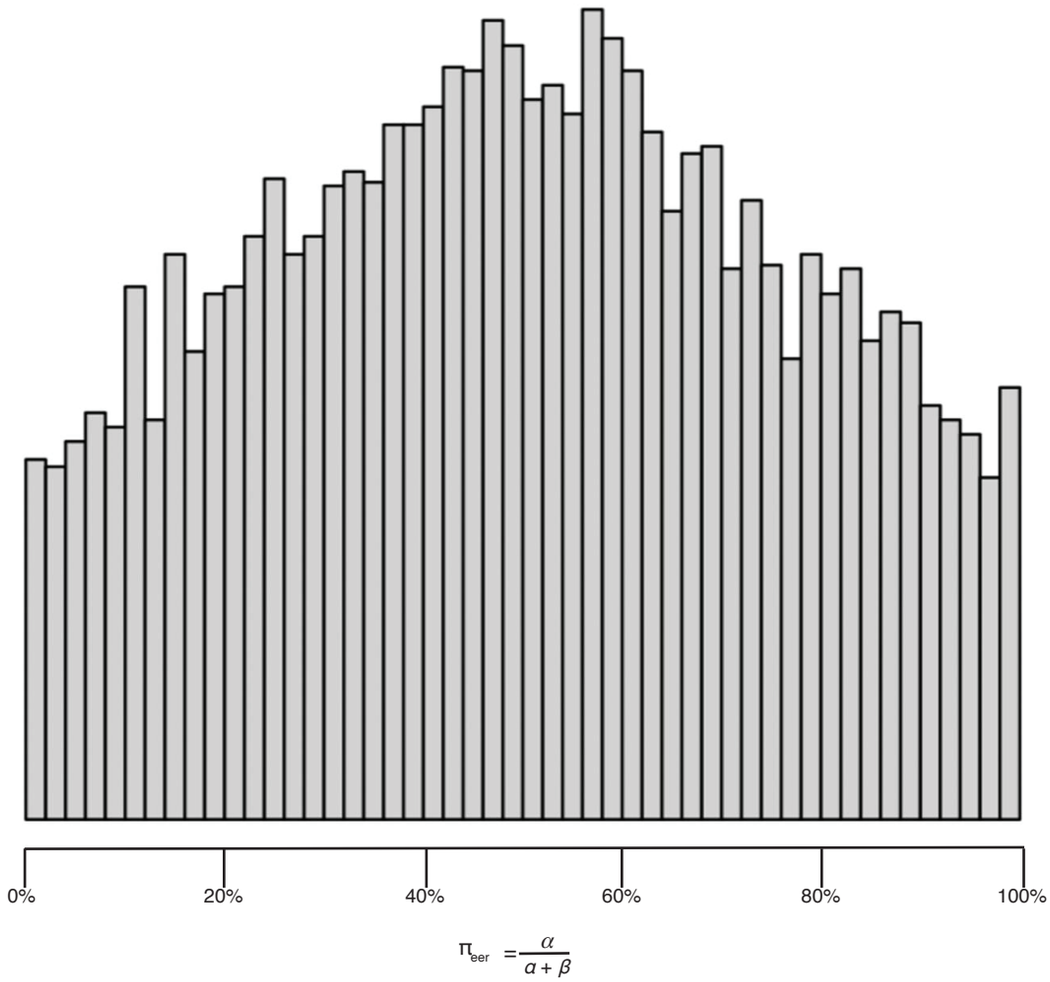

Figure 1- *Simulation of the low correlation prior for examiner error rate: π_{eer}, with μ = 1 and σ = 15. Prior mean and median are both approximately 50%.*

This equation says that everything we currently "know" about the examiner error rate is formed by what we believed about it before, combined with what we learned about it from the data. The quantity $p(\pi_{eer} \mid s)$ is the posterior or "updated" probability distribution for peer in light of the data we observe.

For the prior, we would like to assume little (there is no such thing as a completely uninformative prior) which for us amounts to spreading possible values for peer fairly evenly over the interval [0,1]. The prior for peer must be specified in terms of priors for α and β. For these we take fairly diffuse truncated normal distributions:

$\alpha, \beta \sim \text{TruncNorm}(\mu, \sigma)$

Gaussians are proper probability densities (i.e. normalizable, though this is not strictly necessary), and we truncate them because $\alpha > 0$ and $\beta > 0$ for the Beta-binomial distribution. To maintain α and β above 0 we take as a practical truncation point 1×10^{-8}. Figure 1 shows a simulation of the prior for π_{eer} with $\mu = 1$ and $\sigma = 15$. It is fairly uninformative and has a (prior) mean of about a 50% error rate.

These values for μ and σ imply a distribution for correlation that is shown in Figure 2.

This is a fairly informative prior on ϕ and indicates that we initially believe that there is not much correlation between the ID opinions an examiner will

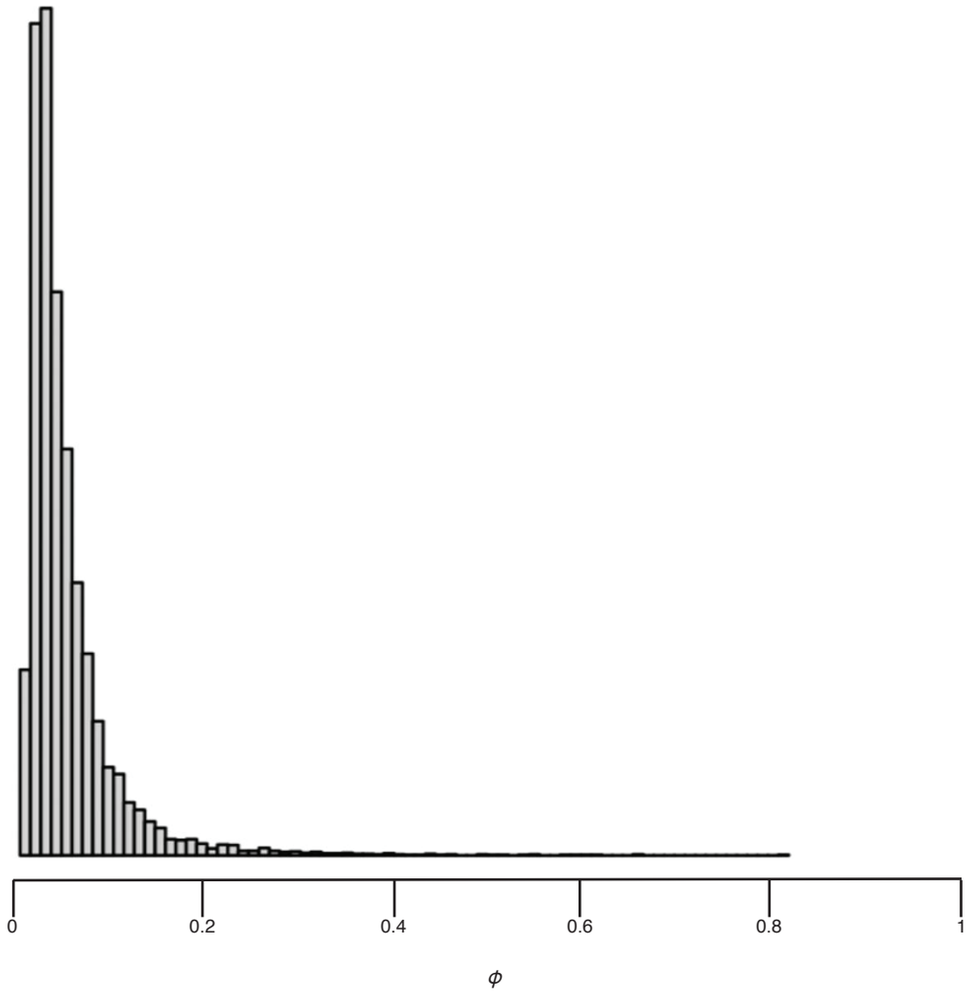

Figure 2- *"Low correlation prior": Simulation of the prior for correlation between Bernoulli trials: ϕ, with $\mu = 1$ and $\sigma = 15$. Prior mean = 0.05, prior median = 0.04.*

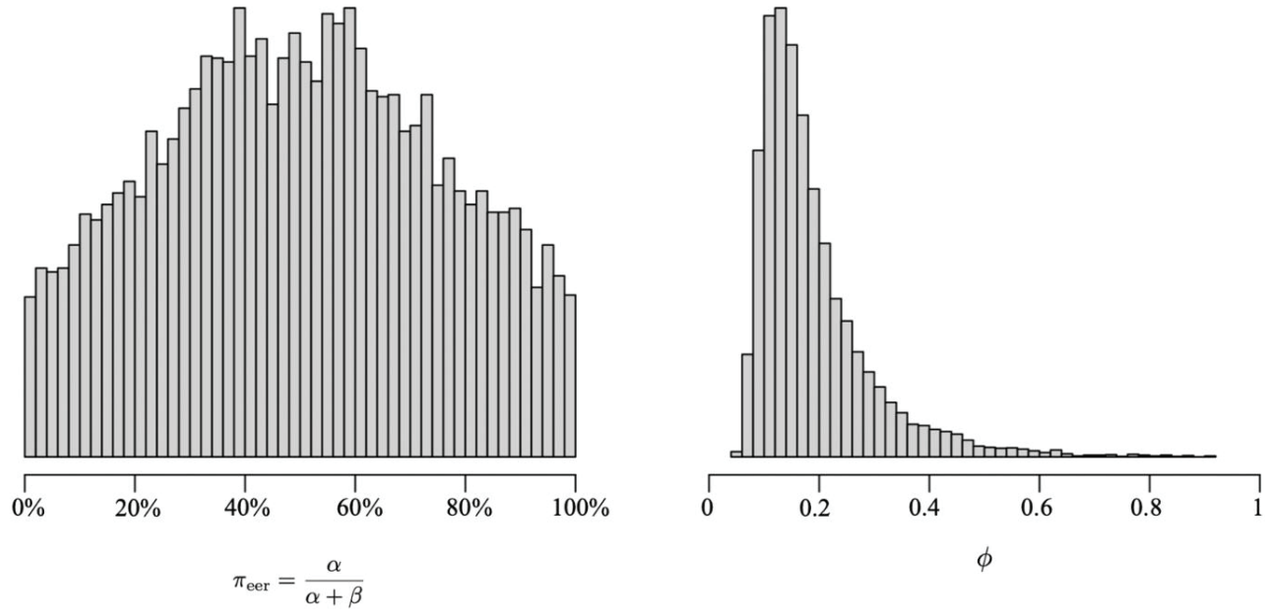

Figure 3- *"Moderate correlation prior": Simulation of the prior for π_{eer}, and ϕ, with $\mu = 1$ and $\sigma = 3$. Prior mean/median on πeer, are 50%/50%, . Prior mean/median on ϕ are 0.19/0.16.*

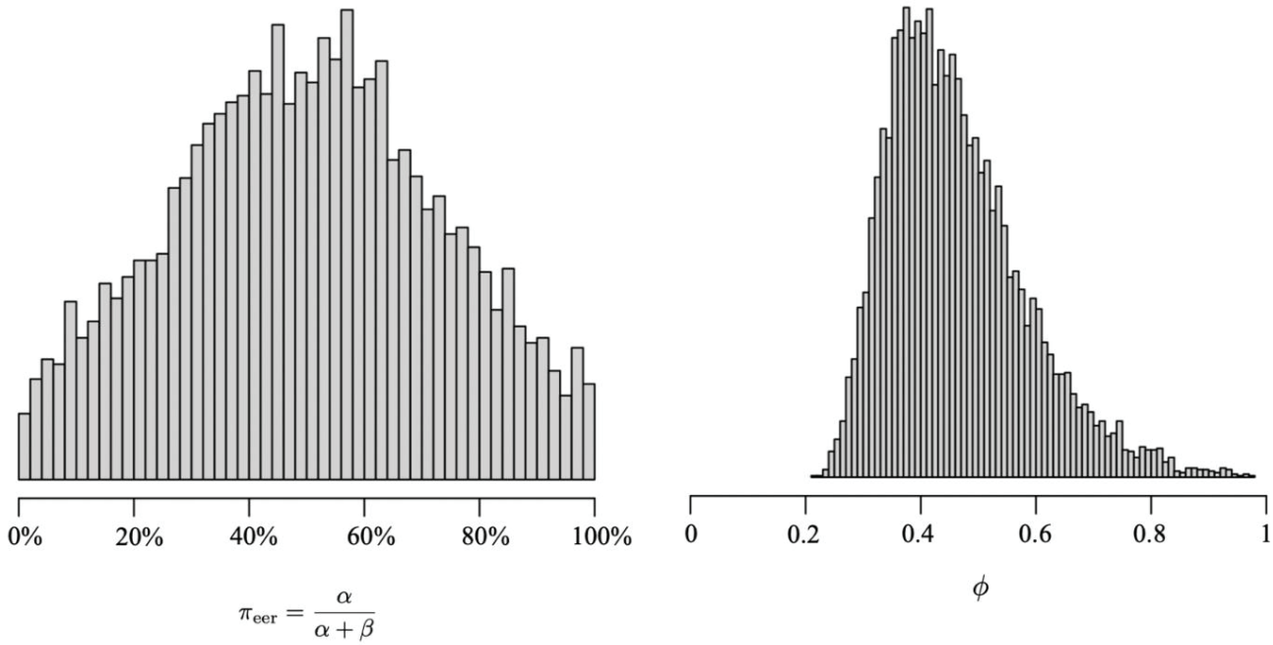

Figure 4- *"High correlation" prior: Simulation of the prior for π_{eer}, and ϕ, with $\mu = 0.5$ and $\sigma = 0.5$. Prior mean/median on πeer, are 50%/50%, . Prior mean/median on ϕ are 0.47/0.45.*

render. We will call this the "low correlation prior". A priori we don't really know that this is the case. In fact we suspect that it is not. However, we will run the posterior analysis for π_{eer} using this prior on ϕ for comparison to other choices for a prior on ϕ.

To change the prior on ϕ, for this study we simply changed the values μ and σ. Figure 3 shows the implied priors on π_{eer} and ϕ using $\mu = 1$ and $\sigma = 3$.

Note the prior for on π_{eer} is essentially unchanged from that shown in Figure 1 where $\mu = 1$ and $\sigma = 15$.

However the prior on ϕ has significantly spread out, now with mean 0.19. We will call this the "moderate correlation prior". Figure 4 shows the implied priors on π_{eer} and ϕ using $\mu = 0.5$ and $\sigma = 0.5$.

While the tails have thinned a bit, the prior for π_{eer} still resembles those in figures 2 and 3. The prior mean and median are both still also 50%. The prior for ϕ however now has a much fatter right tail than the previous priors with significant mass from 0.6 to 0.9 (prior mean is 0.47). We will call this the "high correlation prior". Further discussion and justification for the chosen parameterization of this model appears below in the results and discussion section.

Posterior analysis for π_{eer} was carried out using these three priors; "low", "moderate" and "high" correlation. The joint probability density for the Schuckers' model framework may be very compactly represented as the directed acyclic graph (DAG) shown in Figure 5.

The DAG shows visually how the data's likelihood depends on the parameters α and β. Since we have examiner error data (s_i, i = 1 through 792 we can use it to update our prior assumptions about α and β, and hence our knowledge about the mean examiner error rate π_{eer}.

The posteriors for α and β were determined by sampling the joint probability density with the statistical modelling software Stan [15]. Eight chains were used with 10,000 warm-up and 10,000 sampling iterations each. After warm-up, the chains were thinned by keeping only every 10th sample. R-hat convergence diagnostics were all 1.0 (the chains are effectively converged) [16]. A total of approximately 7,500 (marginal) samples for α and β were drawn from the posterior using each prior. With posterior samples of α and β in hand, the overall examiner error rate given the data was computed as described above.

4. Results and Discussion

A total of seven hundred and ninety-two (792) responses have been received from a total of 36 countries. In the original Brundage study, one laboratory reported an inconclusive result in that they were unable to associate an unknown bullet with the known bullets due to damage to the projectile [4]. In the expanded studies by Hamby et al. two examiners felt that there were insufficient individual characteristics on two of the bullets due

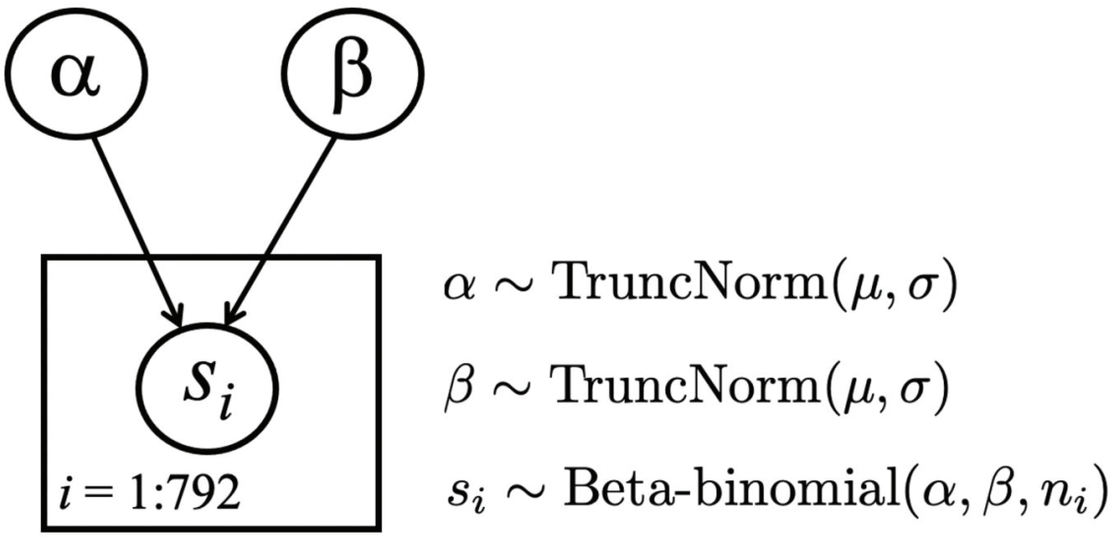

Figure 5- *DAG for the probabilistic model of error rate. Parameters μ and σ are fixed and taken to be (1, 15) "low correlation prior", (1, 3) "moderate correlation prior" and (0.5, 0.5) "high correlation prior".*

to tank rash (tank rash is an unofficial term used by some firearms examiners to denote the damage caused when fired bullets strike the bottom of the water recovery tank.) [5-7]. In another instance, two trainee examiners were unable to correctly associate 5 of the unknown bullets (1 for one examiner, 4 for the second examiner). In each instance, the examiners reported their findings as an inconclusive. No misidentifications were found for any of the above iterations of the "10-barrel test".

Eight test sets were also examined using 'ballistics' imaging equipment. The sets were examined using the following semi-automated systems:

- Intelligent Automation's SciClops™ - Maryland, United States (1 set);
- Automated Land Identification System (ALIS) - Tokyo, Japan (1 set);
- Integrated Ballistics Identification System (IBIS)™ - Georgia, United States (1 set);
- BulletTRAX-3D™ - Forensic Technology - Montreal, Canada (2 sets);
- National Institute of Standards and Technology (NIST) - Maryland, United States (2 sets);
- Plu-neox Sensofar 3D™ – Alabama Department of Forensic Sciences (1 set);
- EVOfinder Scan Bi™, Forensic Institute, Zurich, Switzerland (1 set);
- BalScan™, Forensic Institute, Czech Republic (1 set),
- BulletTRAX-HD3D™ – National Forensic Science Services, Ladyville, Belize (1 set).

The operators of each system reported correct answers. As a side note, this subset of data provided by the semi-automated systems indicates that they can be helpful to the forensic examiner and effective when properly used by an experienced operator.

5. Evaluation

Background information was provided on approximately 630 of the questionnaires. Responses were obtained from 36 countries on four continents. Participants from the following countries contributed to this worldwide research project: Algeria, Australia, Barbados, Belgium, Belize, Botswana, Canada, China, Czechoslovakia, Germany, Greece, Israel, Jamaica, Japan, Jordan, , Mexico, Netherlands, New Zealand, Norway, Pakistan, Palestine, Panama, Philippines, Saudi Arabia, Singapore, Switzerland, South Africa, Thailand, Trinidad & Tobago, United Arab Emirates, United Kingdom and the United States. In the United States, responses were received from examiners in 49 states and the territories of Guam and Puerto Rico. Several states and/or provinces from Australia and Canada submitted responses as well. Demographic data of this continued work has not significantly changed from that of previously reported iterations. We refer the interested reader to the study of Hamby et al. for the complete information [6].

6. Analysis of Examiner Error Rate

Empirically, no errors were made in this aggregate 10-barrel study. A total of five examiners called eight inconclusives between them. The goal of this study is now to take a principled probabilistic approach to infer what the data say about the overall examiner error rate π_{eer}.

Note for any high performance "classifier" the count of errors made will be low. However in this situation, i.e. when examiners make few to no errors, the theoretical mean error rate becomes difficult to determine because it is so small. In fact, classic frequentist based interval estimates completely fail in this situation without ad-hoc corrections [13]. For this reason we have opted for the Schuckers' model framework presented in the methods section [8].

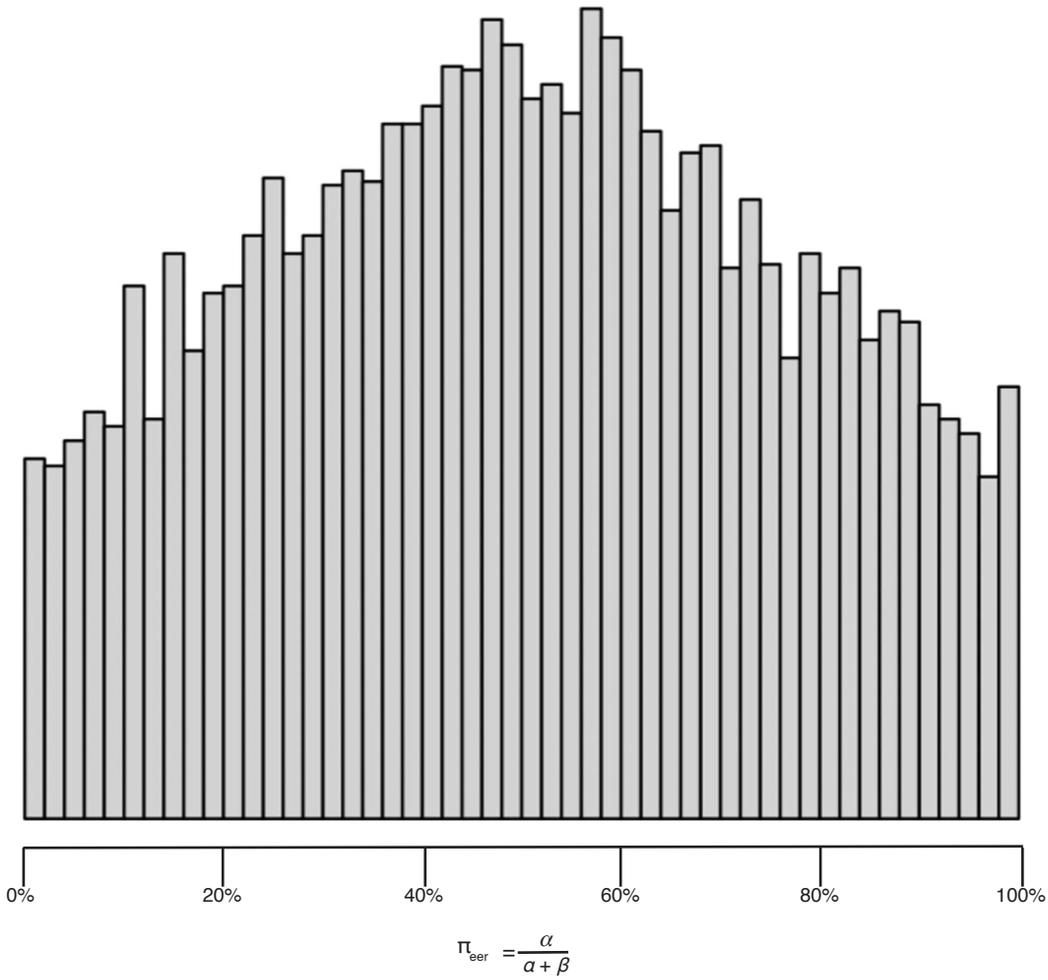

FIG 6- *Box-and-whiskers graphical summary of posterior results for π_{eer},|s, examiner error rate based on their responses to the test. The thin horizontal black lines are the posterior medians. The thick black vertical lines are the 95% highest posterior density intervals. Low, medium and high correlation refers to the modeled intra-response correlation between each of the examiners' 15 matching tasks.*

Inconclusive opinions were not forbidden as responses for test participants. Their presence factors into the statistical analysis by affecting n_j. For respondents who rendered an ID on each test exemplar (correct or incorrect) n_j = 15. For the five participants who rendered inconclusive opinions the n_j's were equal to 14, 13, 13, 14 and 11 respectively (cf. first paragraph of the Results and Discussion section).

Table 3 summarizes the posterior examiner error rate probabilities π_{eer}|s under assumptions of "low", "moderate" and "high" correlation between responses for each examiner.

The intervals presented in Table 3, and through out the paper, all represent a highest (posterior or prior) density set with 95% probability. That is, they are the narrowest regions that encompass π_{eer}|s with 95% posterior probability. A graphical summary of these results appears in Figure 6.

The whiskers of the plots range over the support for π_{eer}|s resulting from the MCMC calculation. The thick black vertical lines represent the 95% highest posterior density sets indicating the narrowest posterior region where we believe the examiner error rate lies with 95% probability given

the data observed. The first thing to note is that while the posterior mean/median examiner error rate estimates are all low, they do increase with increasing intra-response correlation. As can be seen in Figure 6 however, this effect is relatively small. The most conservative estimate is that which results from a "high correlation" prior assumption. Those posterior quantities are a posterior median examiner error rate of 0.03% with a 95% probability interval of [2×10^{-6} %, 0.1%].

7. Conclusion

The design of this multi-decade study was intended to explore if examiners and researchers in forensic firearms analysis could accurately identify 15 'unknown' bullets; obtained by test firing 10 consecutively rifled semiautomatic pistol barrels. A total of 792 completed tests have been received up until this point in time, which includes sixty-seven responses from examiners who participated in the original study [4]. Of the 11,880 unknown bullets examined, three examiners felt that there were insufficient individual characteristics on two of the bullets (due to tank rash), two trainee examiners were unable to correctly associate 5 of the unknown bullets, reporting their findings as an inconclusive. The remaining 11,872 'unknown' bullets were correctly identified by participants to the provided 'known' bullets. The fact that there no actual misidentifications have been reported up until this point empirically demonstrates the efficacy of the training and procedures used to ascribe bullets fired from consecutively rifled barrels.

The international nature of the study demonstrates that the results are not produced by some localized effect or sampling bias. However the lack of actual errors makes it difficult to calculate the true error rate. For purposes of discussion – and considering that the Daubert legal ruling in the United States discusses an 'error' rate; we decided to exploit a Bayesian framework described by Schuckers to estimate a matching systems performance when no or few errors are observed. By using this statistical framework a reasonable estimate of examiner error rates given our observations. Recently, critics have decried this test design, comparing it to a matching problem on an exam or a Sudoku puzzle [17]. The suggestion is that examiners narrow down their choices as they identify unknown bullets to specific barrels, thus reducing the sampling space for the remaining unknown bullets. In practice though, the situation is more complex. This comparison relies on the assumption that no mistakes are made at the beginning of the task; if a mistake is made and an unknown bullet is identified to the wrong barrel, then that error would propagate throughout the test leading to a higher error rate. A valid shortcoming of this study is that it is a closed set test design and each test exemplar has a match.

This study shows that there are identifiable features on the surface of bullets that may link them to the barrel that fired them. Errors due to subclass characteristics, which one could conjecture would be a significant issue when consecutively rifled barrels are involved, has not been a problem for the examiners who participated in the "10-barrel test". Overall, the study as reported up until this point in time, has continued to demonstrate that the identification process has an extremely low error rate if the fired bullets are in good condition and the examiners have been trained under currently accepted regimes [18]. In fact this error rate is too low to empirically be found and must be inferred with Bayesian statistical methods. This study also shows that various statements made about the inability of examiners to associate fired bullets to consecutively rifled barrels are clearly incorrect. It should be noted that 781 participants conducted their examinations

using conventional optical comparison microscopy while 11 participants used some type of ballistics imaging to conduct their examinations.

Using the Schuckers statistical model, posterior mean/median examiner error rates were determined to be 0.01%/0.008% assuming "low" examiner intra-response correlation (denoted ϕ in this study). These estimates increased slightly to 0.02%/0.02% and 0.05%/0.03% under "moderate" and "high" correlation. Inconclusive opinions factored into the analysis by affecting the total number of matches that could be called. This effectively decreases the sample size for the examiner calling the inconclusive(s).

Though the data did not strongly change prior assumptions of correlation, increasing correlation did increase the posterior examiner error rate estimates and widened the uncertainty (highest posterior density intervals) around the error estimates. Our most conservative posterior estimate for examiner error rate assumes correlation is high within an examiner's responses. Given the data collected for this study, we believe the error rate to be in the range of [2×10-6 %, 0.1%] with 95% probability. Note that all of our computations started a priori assuming the examiner error rate was about 50% and overall it was fairly uncertain.

In circumstances where bullets are deformed or fragmented, the comparison process may be more difficult. Another limitation of this study is that bullets that were not fired through one of the ten consecutively manufactured barrels were not included in the test sets. There inclusion could conceivably increase the inferred examiner error rate. These criticisms are appropriate. To accommodate them we are currently conducting a redesigned "10-barrel test" which does not suffer from these issues. The data gathered will ultimately be compared with that found here.

Conflict of interest

The authors declare no conflicts of interest.

Source of funding

The authors received no financial support for the research, authorship or publication of this paper.

Acknowledgments

The authors declare that they have no known conflicts or competing financial interests. The authors gratefully acknowledge the participation of everyone that submitted their data for this research project.

References

1. Goddard CH. Scientific Identification of Firearms and Bullets. J Crim Law Criminol. 1926;17(2):254–63.
2. Gunther JD, Gunther CO. The Identification of Firearms, from Ammunition Fired Therein with an Analysis of Legal Authorities. New York, NY: John Wiley & Sons; 1935. 342 p.
3. Churchman JA. The Reproduction of Characteristics in Signatures of Cooey Rifles. RCMP Gaz. 1949;11(5):133–40.
4. Brundage DJ. The Identification of Consecutively Rifled Gun Barrels. AFTE J. 1998;30(3):438–44.
5. Hamby JE. The Examination, Evaluation and Identification of 9mm Cartridge Cases Fired from 617 Different GLOCK Model 17 & 19 Semiautomatic Pistols. AFTE J. 2009;41(4):310–24.
6. Hamby JE, Brundage DJ, Petraco NDK, Thorpe JW. A Worldwide Study of Bullets Fired From 10 Consecutively Rifled 9MM RUGER Pistol Barrels—Analysis of Examiner Error Rate. J Forensic Sci. 2019;64(2):551–7.
7. Hamby JE. Forensic Firearms Examination [Ph.D. Dissertation]. [Glasgow, Scotland]: University of Strathclyde; 2001.
8. Schuckers ME. Interval estimates when no failures are observed. In: AutoID'02 Proceedings: Workshop

on Automatic Identification Advanced Technologies [Internet]. Tarrytown, NY: IEEE; 2002. p. 37–41. Available from: http://myslu.stlawu.edu/~msch/biometrics/papers/autoidpaper.pdf

9. Biasotti AA. Plastic Replicas in Firearms and Tool Mark Identifications. J Crim Law Criminol. 1956;47(1):110.
10. Law EF, Morris KB. Three-Dimensional Analysis of Cartridge Case Double-Casts. J Forensic Sci. 2020;65(6):1945–53.
11. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.
12. Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat. 2003;31(6):2013–35.
13. Agresti A, Coull BA. Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. Am Stat. 1998 May 1;52(2):119–26.
14. Schuckers ME. Computational Methods in Biometric Authentication: Statistical Methods for Performance Evaluation. 1st ed. New York, NY: Springer Science & Business Media; 2010. 330 p.
15. Stan Development Team. Stan: A C++ Library for Probability and Sampling [Internet]. 2018. Available from: https://mc-stan.org/
16. Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Stat Sci. 1992 Nov;7(4):457–72.
17. President's Council of Advisors on Science and Technology. Report on Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. Washington, D.C.: Executive Office of the President; 2016 Sep p. 147.
18. National Institute of Justice. Firearms Examiner Training. 2010 [cited 2023 Feb 9]. Online Course: Firearms Examiner Training. Available from: https://firearms-examiner.training.nij.gov/usermanagement/login_form?came_from=https%3A//firearms-examiner.training.nij.gov/&retry=&disable_cookie_login__=1&page_n=0&page_nm=.