JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, 2018
doi: 10.1111/1556-4029.13916
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

CRIMINALISTICS

James E. Hamby,[1] Ph.D.; David J. Brundage,[2] M.S.; Nicholas D. K. Petraco,[3,4] Ph.D.; and James W. Thorpe,[5] Ph.D.

A Worldwide Study of Bullets Fired From 10 Consecutively Rifled 9MM RUGER Pistol Barrels—Analysis of Examiner Error Rate

ABSTRACT: This technical note is an update on a continuing study, first designed and initiated by Brundage et al. over twenty years ago (1–4), which seeks to test the community of forensic firearms examiners' ability to associate fired bullets with the barrels through which they passed. To date, 697 participants have utilized over 240 test sets consisting of bullets fired through 10 consecutively rifled RUGER P-85 pistol barrels. Here, we report on the results of the ongoing "10-barrel test" up until the point in time of writing this manuscript. To analyze the totality of data thus far collected, a Bayesian approach was selected. Posterior average examiner error rates are assigned assuming only vague prior information. Given the data found over the course of this diverse decades-long study, our most conservative value for average examiner error rate has a posterior mean of 0.053% with a 95% probability interval of [1.1×10^{-5}%, 0.16%].

KEYWORDS: forensic science, consecutively rifled barrels, criteria for identification, Daubert, firearms identification, fired bullets, ballistics imaging instrumentation, IBIS, SciClops®, scientific research, subclass characteristics, error rates, Bayesian statistics

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. It is recognized that striations are caused by imperfections in the rifling tools used to make gun barrels during the manufacturing process. The tools change during their use and potentially impart a continually changing set of striations. It would be expected therefore 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. We have extensively reviewed past studies that have been aimed at testing the veracity of this hypothesis and thus will not reproduce the discussion. Interested readers are directed to our previous publications (1–4).

Our work here is a large-scale expansion of the study originally presented by Brundage (1). In this update, we assess the rate at which examiners correctly associate fired bullets with the barrels through which they passed, given those barrels were consecutively manufactured. The statistical model we use, which was first proposed by Schuckers, 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"/"no-match" conclusions examiners' render for each bullet/barrel pair in the test. Using data from a worldwide sample of examiners, a posterior assignment of (average) examiner error rate is produced which helps 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?

Methodology

This study is a continuation of that first initiated by Brundage (1). As such, the examiner side of the "10-barrel" test procedure already appears in the literature and it will not be repeated here. The interested reader can refer to reference (4) for the complete design of the study.

Table 1 lists the total number of examiners who have ever taken the test along with counts of the inconclusive and incorrect identifications they rendered.

Statistical Model

We are interested in assigning the probability that on average, an examiner will call a match between a bullet and a barrel, when in fact that is not the case. We will refer to this probability as an average examiner error rate in that we consider the error rate, on average, across the set of examiners tested. When error

[1] International Forensic Science Laboratory & Training Centre, 410 Crosby Drive, Indianapolis, IN 46227.

[2] Independent Examiner, 2541 Belle Brook Drive, Franklin, TN 37067.

[3] Department of Sciences, John Jay College of Criminal Justice, City University of New York, 899 10th Avenue, New York, NY 10019.

[4] Faculty of Chemistry and Faculty of Criminal Justice, Graduate Center, City University of New York, 365 5th Avenue, New York, NY 10016.

[5] Department of Pure and Applied Chemistry, Forensic Science Division, University of Strathclyde, 16 Richmond Street, Glasgow, Scotland G1 1XQ, UK.

Corresponding author: Nicholas D. K. Petraco, Ph.D.
E-mail: npetraco@gmail.com

Received 11 June 2018; and in revised form 2 Aug. 2018; accepted 30 Aug. 2018.

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	630	4	7	0
Totals:	697	5	8	0

rates are small, as they should be for any forensic practice (and as we observe in this study), it turns out that it can be difficult to compute them precisely. Frequentist-based methods are known to perform poorly in this situation (5–7). Thus, we have opted to take a Bayesian approach from which we may infer a reasonable assignment of *average examiner error rate*, π_{aeer}, given the data we observe in this study (5,6). Below, we describe the model, due to Schuckers, which has been shown to render reasonable assignments for posterior error probabilities (average examiner error rate, $\pi_{\text{aeer},}$ in this case) even when they are very small. Schuckers method will produce a posterior distribution for the average examiner error rate that also allows for the expression of uncertainty in its value.

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

So the question now is, what are the data for this study? 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 ith unknown bullet ($i \in \{1, \ldots, n_j\}$), $x_i = 0$ if the examiner makes the correct “match” and $x_i = 1$ if the examiner makes an incorrect “match.” In symbols:

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

The actual data analyzed 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 $x_{i,j}$ represent the outcome for the jth 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 jth 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 1). Thus, if an examiner renders one or more inconclusive opinions, then $n_j < 15$.

Data for this study are the number of errors each examiner made, s_j, organized into a vector of length 697, **s**. Often sums of Bernoulli outcomes are modeled as arising from a binomial distribution. However, in our case, there can conceivably be some correlation between the 15 matching attempts (Bernoulli trials) each examiner undertakes; that is, an examiner's answer on one trial may affect their answer on another trial. Modeling the data with only the binomial would not take this 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 (6):

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

There were 697 examiners who contributed to the data set, and each examiner underwent n_j, possibly correlated, Bernoulli trials. Thus, the data for this study **s** are modeled as a product of beta binomial distributions:

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

The beta binomial distribution is (usually) parameterized in terms of two new parameters, α and β, instead of π_{aeer}. From them, we can recover the average examiner error rate, π_{aeer}, as follows:

$$\pi_{\text{aeer}} = \frac{\alpha}{\alpha+\beta}$$

Correlation between decisions for the jth examiner, C_j, is modeled as follows:

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

which is on a scale from 1 (no correlation between Bernoulli trials) to n_j (full correlation between Bernoulli trials). Note that under this model, the only dependence of C_j on examiner j is the number of opinions they render, n_j, as α and β are averaged over all examiners. Note also that we can reparameterize the correlation for this model to be on a more familiar 0–1 scale as follows:

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

For $\phi = 0$ ($C_j = 1$), we recover the binomial distribution from the beta binomial distribution. In this model, ϕ has no explicit dependence on the index j because any terms depending on j cancel out of the equation (to see, this substitutes the equation for C_j into the rightmost expression for ϕ). Hence, the index j has been dropped from ϕ. For the remainder of this paper, we will use the ϕ coefficient as our measure for correlation.

The prior knowledge concerning π_{aeer} can be updated with the likelihood, $p(\mathbf{s}\ |\pi_{\text{aeer}})$, via Bayes theorem:

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

This equation says that everything we currently "know" about the average examiner error rate is formed by what we believed about it before, combined with what we learned about it from the data. The quantity $p(\pi_{\text{aeer}}|\mathbf{s})$ is the posterior or "updated" probability distribution for π_{aeer} 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 π_{aeer} fairly evenly over the interval [0,1]. The prior for π_{aeer} 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, although 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_{aeer} 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 Fig. 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 render. We will call this the "low correlation prior." A priori, we do not really know that this is the case. In fact, we suspect that it is not. However, we will run the posterior analysis for π_{aeer} using this prior on ϕ for comparison with 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_{aeer} and ϕ using $\mu = 1$ and $\sigma = 3$. Note the prior for π_{aeer} is essentially unchanged from that shown in Fig. 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_{aeer} and ϕ using $\mu = 0.5$ and $\sigma = 0.5$. While the tails have thinned a bit, the prior for π_{aeer} still resembles those in Figs 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 appear below in the section Results and Discussion.

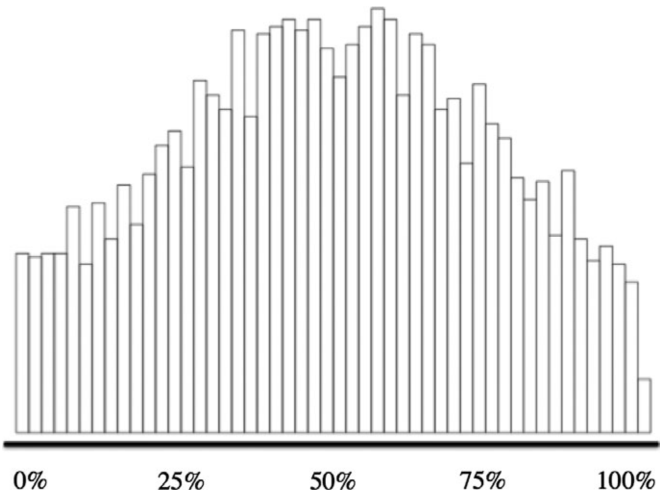

FIG. 1—*Simulation of the low-correlation prior for average examiner error rate: π_{aeer}, with $\mu = 1$ and $\sigma = 15$. Prior mean and median are both approximately 50%.*

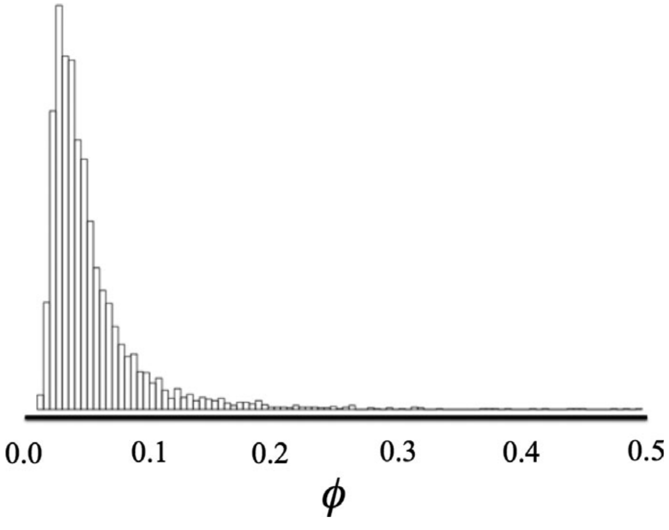

FIG. 2—*"Low-correlation prior": Simulation of the prior for correlation between Bernoulli trials: ϕ, with $\mu = 1$ and $\sigma = 15$. Prior mean = 0.05, prior median = 0.04.*

Posterior analysis for π_{aeer} was carried out using these three priors: "low," "moderate," and "high" correlation. The joint probability density for the Schuckers model may be very compactly represented as the directed acyclic graph (DAG) shown in Fig. 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 697), we can use it to update our prior assumptions about α and β and hence our knowledge about the average examiner error rate π_{aeer}.

The posteriors for α and β were determined by sampling the joint probability density with the statistical modeling software Stan (8). 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) (9). A total of approximately 7500 (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 average examiner error rate given the data was computed as described above.

Results and Discussion

A total of six hundred and ninety-seven (697) responses have been received from a total of 32 countries. A laboratory noted an inconclusive result in that they could not associate a test bullet with any of the known bullets because of reported damage (1). Furthermore, two of the examiners taking the test reported insufficient individual characteristics for two of the bullets in their particular test sets. They noted that they arrived at their decision because of tank rash on the bullets (2–4). Finally, two trainee examiners reported that they could not make associations for 5 of the unknown bullets across their test sets. The first

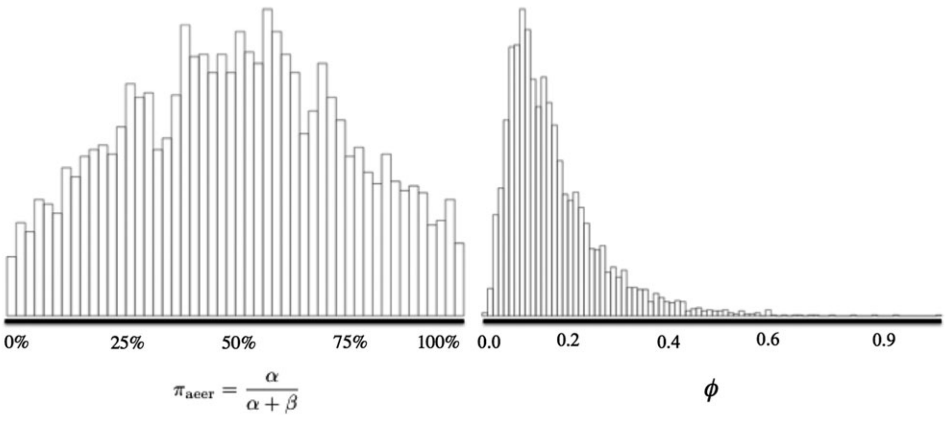

FIG. 3—*"Moderate-correlation prior": Simulation of the prior for π_{aeer} and φ, with $\mu = 1$ and $\sigma = 3$. Prior mean/median on π_{aeer} is 50%/50%. Prior mean/median on ϕ is 0.19/0.16.*

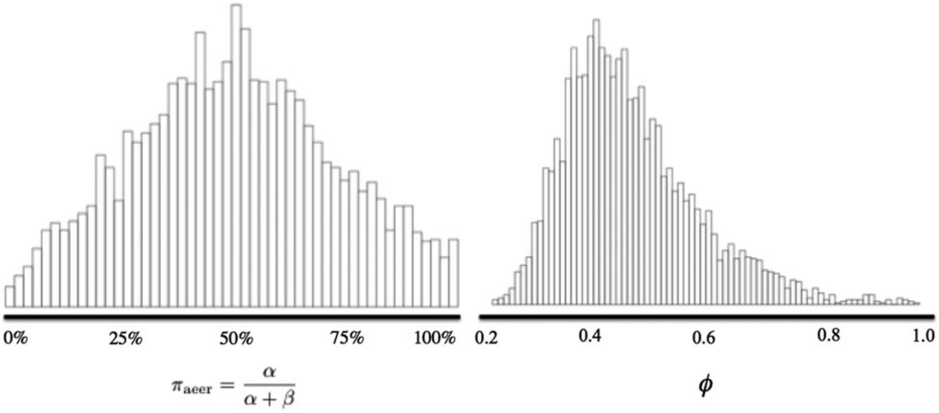

FIG. 4—*"High-correlation" prior: Simulation of the prior for π_{aeer} and φ, with $\mu = 0.5$ and $\sigma = 0.5$. Prior mean/median on π_{aeer} is 50%/50%. Prior mean/median on ϕ is 0.47/0.45.*

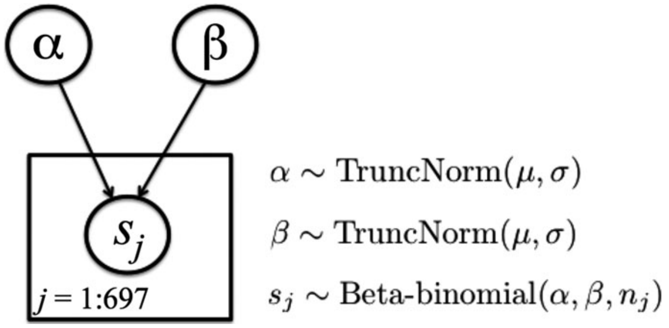
α
β
s_j
$j = 1{:}697$
$\alpha \sim \text{TruncNorm}(\mu, \sigma)$
$\beta \sim \text{TruncNorm}(\mu, \sigma)$
$s_j \sim \text{Beta-binomial}(\alpha, \beta, n_j)$

FIG. 5—*DAG for the Schuckers' 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."*

trainee examiner could not associate one bullet, and the second trainee could not associate four bullets. In each instance, the examiners reported their findings as 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 semiautomated systems:

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

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

Evaluation

Background information was provided on approximately 630 of the questionnaires. Responses were obtained from 32 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, the Netherlands, New Zealand, Norway, Pakistan, Palestine, Panama, the Philippines, Saudi Arabia, Singapore, Switzerland, South Africa, Thailand, Trinidad and Tobago, the United Arab Emirates, the 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 have not significantly changed from those of previously reported iterations. We refer the interested reader to (4) for the complete information.

Analysis of Average 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 average error rate π_{aeer}.

Note for any high-performance "classifier," the count of errors made will be low. However, in this situation, that is, 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 (5). For this reason, we have opted for the Schuckers model presented in the section Methods (6).

Inconclusive opinions were not forbidden as responses for test participants. The problem with inconclusives in a binary decision paradigm is that they do not neatly fit into either the "correct" or "incorrect" categories. They are not wrong decisions in that it was felt by the examiner that no decision could be made. One could count them as technically "correct," but this is not without criticism, not least of which because it can lead to an underestimation of average examiner error rate. In this study, we compromise between the two extremes and do not consider them in the analysis. This affects the number of decisions made, n_j, by examiners who rendered them. It has the penalizing affect of decreasing the decision sample size for the given examiner and therefore contributes to increasing the uncertainty on the average examiner error rate by widening its posterior distribution. It is not as penalizing, however, as counting inconclusive responses as incorrect, which they are not. 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. second paragraph of Results and Discussion section).

Table 2 summarizes the posterior average examiner error rate probabilities $\pi_{\text{aeer}}|\mathbf{s}$ under assumptions of "low," "moderate," and "high" correlation between responses for each examiner.

The intervals presented in Table 2, and throughout the paper, all represent the highest (posterior or prior) density set with 95% probability. Note these probability intervals are also commonly referred to as credibility intervals. That is, they are the narrowest regions that encompass $\pi_{\text{aeer}}|\mathbf{s}$ with 95% posterior probability. A graphical summary of these results appears in Fig. 6. The whiskers of the plots range over the support for $\pi_{\text{aeer}}|\mathbf{s}$ 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 average examiner error rate lies with 95% probability. The first thing to note is that while the posterior mean/median average assignments are all low, they do increase with increasing correlation. We can see though from Fig. 6 that this effect is relatively small. The most conservative assignment is that which results from a "high-correlation" prior assumption. Those posterior quantities are a posterior mean average examiner error rate of 0.053% with a 95% probability interval of [$1.1 \times 10^{-5\%}$, 0.16%].

It is also interesting to examine what the data have to say about our prior beliefs on correlation between responses for each examiner, ϕ. As a matter of note, we did attempt to parameterize this model for average examiner error rate directly in terms of π_{aeer} and ϕ, instead of α and β. This would allow putting explicit priors on ϕ, for example a "uninformative" wide distribution such as Uniform (0,1). Unfortunately, convergence of the MCMC chains was extremely slow and the resulting posterior samples were highly correlated leading to very low effective posterior sample sizes. For this reason, we used the (typical) α and β parameterization. Still though under this parameterization, wide "uninformative" priors could be placed on hyperparameters μ and σ. This also leads to extremely slow convergence and highly correlated posterior samples. Thus, we chose to examine the level of correlation (ϕ) coarsely in terms of "low correlation," "moderate correlation," and "high correlation" by choosing appropriate values for fixed hyperparameters μ and σ. Table 3 lists the prior summaries for ϕ, to reiterate what we mean by "low," "moderate," and "high" correlation.

Figure 7 shows several violin plots, which display the effect on the probability density for ϕ as we introduce the data and move from prior to posterior beliefs. The first thing to note is that the posterior for ϕ is fairly similar to the prior whether or not we initially believe there is "low," "moderate," or "high" levels of correlation. This essentially means that the data are not able to inform our opinions about the correlation between examiner responses in this study. In such a case, it is probably best then to be conservative and a priori assume there is a moderate or high amount of correlation. Thus, $\pi_{\text{aeer}}|\mathbf{s}$ with an a priori "high correlation" ϕ represents our best assessment for average examiner error rate given the data thus far obtained.

TABLE 2—*Summary of posterior results for* $\pi_{\text{aeer}}|\mathbf{s}$*: inferred average examiner error rate-based responses data for the test shown in Table 1.*

	"Low Correlation" Prior	"Moderate Correlation" Prior	"High Correlation" Prior
Mean $\pi_{\text{aeer}}\|\mathbf{s}$	0.015%	0.026%	0.053%
Median $\pi_{\text{aeer}}\|\mathbf{s}$	0.010%	0.017%	0.036%
Probability interval*,† $\pi_{\text{aeer}}\|\mathbf{s}$	[2.5×10^{-6}%, 0.043%]	[3.5×10^{-6}%, 0.080%]	[1.1×10^{-5}%, 0.16%]

*These are the 95% highest posterior density intervals.

†Also commonly called a credibility interval.

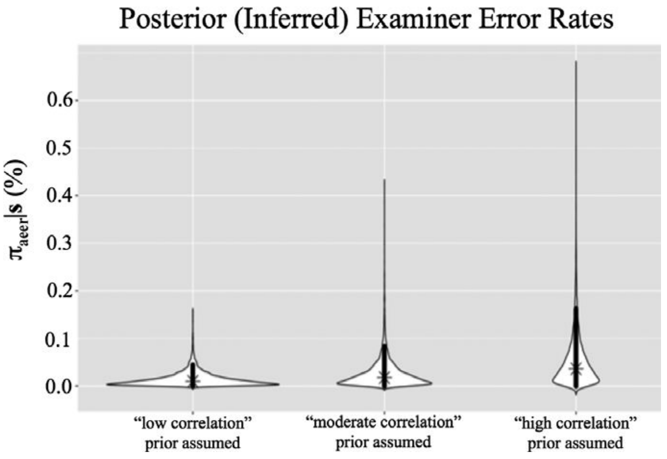

FIG. 6—*Violin plot graphical summaries of posterior results for $\pi_{aeer}|$**s**, average examiner error rate based on their responses to the test. The black stars are the posterior medians. The thick black vertical lines are the 95% highest posterior density intervals.*

TABLE 3—*Summary of priors on ϕ, the examiner decision correlation parameter.*

	"Low Correlation" Prior[†]	"Moderate Correlation" Prior[‡]	"High Correlation" Prior[§]
Mean ϕ	0.05	0.2	0.5
Median ϕ	0.04	0.2	0.4
95% Probability interval*ϕ	[0.01, 0.1]	[0.06, 0.4]	[0.3, 0.7]

*These are the 95% highest prior density intervals.
[†] $\mu = 1$, $\sigma = 15$ fixed hyperparameters.
[‡] $\mu = 1$, $\sigma = 3$ fixed hyperparameters.
[§] $\mu = 0.5$, $\sigma = 0.5$ fixed hyperparameters.

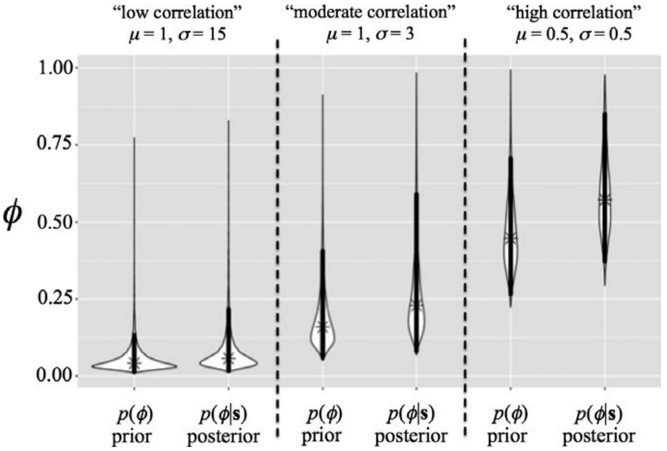

FIG. 7—*Violin plot graphical summaries of posterior results for $\phi|s$, response correlation.*

Conclusion

The design of this multidecade study was intended to explore whether 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 697 completed tests have been received up until this point in time, which includes sixty-seven responses from examiners who participated in the original study, by Brundage. Of the 10,455 unknown bullets examined, three takers of the test reported insufficient individual characteristics for two of the test bullets and two trainees could not associate five of the test bullets to their known counterpart bullets. The trainee examiners reported their findings as inconclusive. The remaining 10,447 "unknown" bullets were correctly identified by participants to the provided "known" bullets. No misidentifications have been reported up until this point. Based on these empirical results, we infer this is due to the efficacy of the training and procedures used to ascribe bullets fired from consecutively rifled barrels.

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 an advanced Bayesian technique due to Schuckers and determine a reasonable assignment of average examiner error rates given our observations.

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, have 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 finds 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 (10). 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 686 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 average examiner error rates were determined to be 0.015%/0.010% assuming "low" intra-examiner opinion correlation (denoted ϕ in this study). These values increased slightly to 0.026%/0.017% and 0.053%/0.036% 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).

Although the data did not strongly change prior assumptions of correlation, increasing correlation did increase the posterior average examiner error rate assignments and widened the uncertainty (highest posterior density intervals) around the error assessment. Our most conservative posterior value for average 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 [$1.1 \times 10^{-5\%}$, 0.16%] with 95% probability. Note that all of our computations started a priori assuming the average examiner error rate was about 50%, and overall, it was fairly uncertain (i.e., the prior for π_{aeer} was fairly flat between 0% and 100%, cf. Figs 1, 3, and 4).

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. Their inclusion could conceivably increase the inferred average 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.

Acknowledgments

The authors gratefully acknowledge the participation of everyone that submitted their data for this research project.

References

1. Brundage J. The identification of consecutively rifled gun barrels. AFTE J 1998;30(1):438–44.
2. Hamby J. Forensic firearms examination, chapter 3 [dissertation]. Glasgow, Scotland: University of Strathclyde, 2001.
3. Hamby J. The identification of consecutively rifled 9 mm pistol barrels: a pre-publication update. Proceedings of the 38th AFTE Training Seminar; 2007 May 1-5; San Francisco, CA. San Francisco, CA: AFTE, 2007.
4. Hamby J, Brundage DJ, Thorpe JW. The identification of bullets fired from 10 consecutively rifled 9 mm Ruger pistol barrels: a research project involving 507 participants from 20 countries. AFTE J 2009;41 (2):99–110.
5. Agresti A, Coull B. Approximate is better than exact for interval estimation of binomial proportions. Am Stat 1998;52(2):119–26.
6. Schuckers ME. Interval estimates when no failures are observed. In: Ratha NK, Bolle RM, editors. AutoID'02 Proceedings: Workshop on Automatic Identification Advanced Technologies; 2002 March 14-15; Tarrytown, NY. New York, NY: IEEE Robotics and Automation Society and The Association for Automatic Identification and Data Capture Technologies (AIM), 2002;37–41.
7. Schuckers ME. Computational methods in biometric authentication: statistical methods for performance evaluation. New York, NY: Springer, 2010.
8. Stan Development Team [computer program]. Stan: A C++ Library for Probability and Sampling, Version 1.3, 2013; http://mc-stan.org/(accessed March 7, 2018).
9. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statist Sci 1992;7(4):457–72.
10. National Institute of Justice. Training: firearms examiner training (account sign-up required); http://firearms-examiner.training.nij.gov/(accessed August 29, 2018).