JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, 2016
doi: 10.1111/1556-4029.13093
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Tasha P. Smith,[1] *B.S.; G. Andrew Smith,*[1] *M.S.; and Jeffrey B. Snipes,*[2] *J.D., Ph.D.*

A Validation Study of Bullet and Cartridge Case Comparisons Using Samples Representative of Actual Casework

ABSTRACT: The foundation of firearm and tool mark identification is that no two tools should produce the same microscopic marks on two separate objects that they would be inaccurately or wrongly identified. Studies addressing the validity of identification infrequently employ tests that mirror realistic casework scenarios. This study attempted to do so using a double-blind process, reducing test-taking bias. Test kits including bullets and cartridge cases but not the associated firearms were completed by 31 analysts from 22 agencies. Analysis of the results demonstrated an overall error rate of 0.303%, sensitivity of 85.2%, and specificity of 86.8%. Variability in performance across examiners is addressed, and the effect of examiners' years of experience on identification accuracy is explored. Finally, the article discusses the importance of studies using realistic case work scenarios when validating the field's performance and in providing courts with usable indicators of the accuracy of firearm and tool mark identification.

KEYWORDS: forensic science, firearm and tool mark identification, error rate studies, validation study, casework simulation, sensitivity and specificity

The foundation of the science of firearm and tool mark identification is that no two tools should produce the same microscopic marks on two separate objects that they would be inaccurately or wrongly identified. Firearm identification relies upon the human cognitive ability of pattern recognition that allows one to determine the individuality of a tool, through the physical comparison of microscopic marks. Years of research have proven that in the evaluation of consecutively manufactured tools – tools which show the greatest potential for leaving the same marks – these tools display sufficient individual differences that when subclass influence is excused, the origin of marks left by consecutively manufactured tools can be determined. The firearm and tool mark examiner often faces several common core questions, such as: Is it possible to identify or exclude a tool as having created a mark from all other possible tools? Can such exclusions and identifications be made with any degree of certainty? What is the range of certainty of this exclusion or identification? How do these findings translate to the everyday community or courts in way that is easy to understand by the layperson?

Much of the research to date has supported the theory of individualization and has been performed so through the microscopic comparison and observation of barrels, slides, knife blades, screwdrivers, and so forth (1–5). Some research has further been complemented by the use of statistical and mathematical models (6–12). Yet, often the validity of these measures is criticized (13–15).

While it is true that errors occur in all human endeavors, whether in computer programming or in an emergency room, the crucial benchmark for bases of comparison over time or across agencies/organizations is the frequency and likelihood of occurrence of these errors. In firearm and tool mark identification, the frequency with which errors occur is difficult to deduce because the outcome of the work is dependent on the presence of controls and quality checks. With mounting methodological criticisms and case decisions, the courts are not interested in a "theoretical error rate," which assumes that everything has been carried out properly and the correct answers have been reached. What they are interested in and what is of more value is what actually happens during routine casework. Additionally, courts want this data be reported with a level of understanding, certainty, and specificity of that commonly seen in DNA analyses (14). However, the level of understanding in firearm and tool mark identification that corresponds to that level of DNA analysis exists only on a subclass level, not on an individual level. The "human factor" in identification accounts for tremendous variability in analysis. Some of the most important questions that have arisen with validation studies include as follows:

- Can a validation study which is representative of actual casework in the field of firearm and tool mark examination be designed and implemented?
- Can this study be presented in a blind or double-blind format?
- Can such a test be designed that addresses the possibility of test-taking bias?
- Can the results be tabulated with a level of accuracy that is reasonably consistent across all examiners?

[1] San Francisco Police, Department Crime Lab, 1245 3rd Street, San Francisco, CA, 94158.

[2] Program in Criminal Justice Studies, San Francisco State University, San Francisco, CA.

Received 8 April 2015; and in revised form 18 Sept. 2015; accepted 4 Oct. 2015.

- To what extent is training and experience a factor in the examiner's decision-making process and outcome?
- Are results and error rate values consistent across studies and are they representative of actual casework values?
- Can these results be articulated in a way that is understandable and of value to the community in a precisely specified and scientifically justified way that leads to a well-characterized confidence limit?

The training of a firearms examiner is based on the understanding of the individualizing marks produced, where they come from and how they are made. This training involves a constant building and refining of what is called an examiner's criteria for identification. The criteria for identification are a subjective point refined through the experience and training of an examiner of what is sufficient and significant agreement in the individual microscopic marks of interest. Such a level of understanding cannot always be conveyed quantitatively; however, through methods such as QCMS the level of agreement that can be translated in a fashion understandable to the general public is approachable. Quantitative consecutive matching stria (QCMS) is a method of identification that provides a quantitative value to the evaluation of striated marks. Although QCMS is becoming more widely used in the field of firearm and tool mark identification, it is limited in that it only applies to striated marks. It is also limited in determining which lines in a pattern can be counted versus those that should not. When solely using pattern matching, it is the combination of the overall similarity of the pattern and the microscopic detail of the pattern of both striated and impressed marks that must meet an examiner's criteria for identification for an identification to be made. An examiner's knowledge base can only be developed and refined through the constant and consistent evaluation of known matches (KMs) to known nonmatches (KNMs) that allow for the assessment of individuality.

The purpose of this study is to present the design and results of a study that has been developed to provide the discipline with a useable accurate error rate that is a clear and concise representation of the actual human work associated with firearms tool mark identification. It also addresses variability in sensitivity and specificity measures across multiple examiners. Finally, it attempts to determine whether there is any relationship between an examiner's years of experience and performance in identification.

Materials and Methods

Test Design

Each test was designed to have a similar feel to what an examiner typically encounters when working a case. It is routine within a criminalistics laboratory that a firearm examiner will receive evidence with little knowledge of the history of the evidence and such evidence is often presented without a firearm. Such situations limit what examiners have to make comparisons with, while also testing their knowledge of manufacturing processes, what is possible, and what is probable, in the operation of firearms. This study aimed to approximate everyday casework by providing examiners with a realistic, albeit simulated, case with no firearm. Such a design should provide a more realistic assessment of error rates in case work. This study is similar to a number of other studies; however, there are marked differences in the design to make it more realistic to what is seen on the bench on a daily basis. Like the studies by Smith (16) and others, the firearms used for test firing were obtained from crime-related cases and therefore were circulated in the general population and subjected to use, corrosion and abuse similar to that observed in a typical case. These tests were then circulated to active firearms examiners with varying years of experience and levels of training, working in laboratories which vary in their policies and procedures for making exclusions when the firearm is absent.

A primary criticism of many of the reported validation studies within the community is that many tests lack anonymity and some examiners are more conservative than others due to the fear of answering incorrectly. This may create a test-taking bias. The current test was as blind as possible except to the extent participants were aware that they were participating in a validation study. To provide as much separation as possible between researcher and participants, requests for participants were sent out by a third party via email or message board to maximize sampling randomness and eliminate any questions of bias between test administrator and the participants. All test takers and supervisors were unaware of the correct answers, and the test administrator was not privy to which individual in a particular laboratory was taking the test. Each test packet was different from the next, eliminating the likelihood of discussions between participants within the same laboratory resulting in any useful information being obtained. Although a number of the tests were sent out multiple times OR sent out on multiple occasions, they were never duplicated within the same laboratory. This not only provided us with a measure of reproducibility but also served as a quality check of the tests themselves. Each test was of similar difficulty. The number of identifications to exclusions varied from test to test, containing anywhere from 12 to 14 true identifications and 20–30 true eliminations as designed.

This study utilized both bullets and cartridge cases from eight different firearms that had been circulated in the general population and now reside in the San Francisco Police Department Crime Laboratory's Firearm Reference Collection. These firearms consisted of at least two with the same class characteristics; therefore, an evaluation of individual microscopic marks was necessary. A total of 406 true identifications and 760 true eliminations were possible within the 31 returned kits as they were designed. There were 1060 actual eliminations possible based on the "if-then" result of the actual conclusions within the test. The number of possible eliminations to identifications sought to challenge the examiners' criteria for identification using either pattern recognition or quantitative consecutive matching striae criteria while also challenging any testing preconceptions developed through the participation in other similar studies. In this study, there were no "knowns" with which to compare "unknowns." This feature is not usually found in traditional studies but is more reflective of the actual level of comparison work that an examiner may encounter. All test sets in this study consisted of at least one cartridge case and/or bullet (or bullet jacket) that did not identify to any other specimen within the test kit.

Materials

Six different types of ammunition consisting of 1104 cartridges were fired through eight different 40 caliber pistols. The various firearms were used because of their unique ability to mark ammunition in ways consistent with what is seen in everyday casework. Two different firearms of a similar make and model for each of the four firearm types were used. The make, model, general rifling characteristics, and serial numbers of the firearms used in this study are documented in Table 1.

TABLE 1—*Types of firearms used from SFPD reference collection.*

Make	Model	Caliber	GRC	Serial Number	Ammunition Type Fired per Firearm
Taurus	PT 101 AFS	0.40	6R	SLD18629D	92 UMC (CC and Bu); 92 WIN BEB (CC and Bu); 92 Hi-Shok/Hydra-shok (Bu); 92 American Eagle (CC)
	PT 101 AF	0.40	6R	SKJ01550/AFD	
Sig Sauer	P229	0.40	6R	AC19988	92 UMC (CC and Bu); 92 Speer GD (CC and Bu); 92 Hi-Shok/Hydra Shok (Bu); 92 American Eagle (CC)
	P229	0.40	6R	AC16713	
Smith and Wesson	4013	0.40	6L	THZ9553	92 UMC (Bu); 92 WIN BEB (Bu); 92 Hi-Shok/Hydra-shok (Bu)
	SW40C	0.40	6L	PAL5819	
Glock	22	0.40	6R	ARC775US	92 UMC (CC); 92 WIN BEB (CC); 92 American Eagle (CC)
	27	0.40	6R	CZR349US	

GRC, General Rifling Characteristics; CC, Cartridge cases; Bu, Bullets.

Six different types of ammunition were used in the execution of this study. A list of ammunition specifications is found in Table 2. Each of the fired bullets and cartridge cases was assigned a unique identifying number as a key. To decrease the chances of a recognizable pattern being observed by test takers, the identifying numbers were obtained using a random number generator program (17). The identifying number was inscribed on the ogive or base of the bullet and jackets; and on the side of the cartridge case using a Dremel model 290-01 engraver. The cartridges were fired into a horizontal water tank equipped with a "lab made" bullet retrieval trap, which was constructed using PVC pipe cut to the dimensions of the tank with durable mesh screen along the bottom. The design and use of this trap allowed for rapid collection of the multiple specimens fired in this study. Representative samples of some of the specimens from the test are provided in Figs 1–6.

Packet Preparation

A total of 50 study packets were prepared, each containing 12 randomly selected bullets/bullet jackets and 12 randomly selected cartridge cases, a supplementary comparison worksheet, an answer sheet, and directions for performing the study (Appendix S1). Each test packet was given its own unique identifier to maintain anonymity of the test participants. Participating laboratories were sent 1–3 packets at their request that were distributed by the supervisor, in most cases, to bench-level analysts. A total of 47 kits were distributed, with 34 returned, three of which were omitted because they violated the conditions of the study in one way or another.

For the purposes of collection, each firearm was fired individually, with all the specimens collected and placed into individually labeled containers. The container was labeled with the firearm make, model, and serial number information. The specimens were later engraved with a unique identifier supplied through the random number generator program. It should be noted that in some cases, only bullets/bullet jackets were collected, such as for the Sig Sauer and Smith and Wesson firearms. And in some cases, only the cartridge cases were collected, such as with the Glock firearms. The total evidence specimen count was 2208, of which 1200 were placed into 50 kits (containing 12 bullets/bullet jackets and 12 cartridge cases). The randomness of this study was maximized by thoroughly mixing all of the bullets/jackets after being scribed with their identifiers. Then, 12 were randomly selected and grouped from the container of bullets and cartridge cases by individuals from the laboratory. The scribed numbers were then recorded onto individual 2 ½" × 4 ¼" size envelopes and placed into the corresponding envelopes sealed with tape and then placed into individual test packets, labeled with a test number 1 thru 50. Over the next several days, examiners from the San Francisco Police Department Crime Lab Firearm and Tool Mark Unit evaluated the kits for their potential for identification, and to ensure that where identifications should be made, they could be made. The examiners had a range of training histories and levels of experience, as did members of the testing group. Following the kit evaluations, the test packets were sealed and shipped to the 47 participants representing approximately 30 different laboratories across the United States and abroad. Participants were given

TABLE 2—*Ammunition specifications.*

Ammunition Name/Brand	Cartridge	Grain	Primer	Case	Bullet Type/ Composition
Remington UMC	40 S&W	165/185	Nickel	Brass	FMJ/Copper
Federal Classic Hi-Shok	40 S&W	155/180	Brass	Brass	JHP/Copper
Federal Classic Hydra-Shok	40 S&W	155	Nickel	Nickel	JHP/Copper
Winchester WinClean BEB	40 S&W	165	Nickel	Brass	FMJ/Brass
Speer Gold Dot	40 S&W	180	Nickel	Nickel	JHP/Copper
American Eagle	40 S&W	180	Brass	Brass	FMJ/Copper

FMJ, Full Metal Jacket; JHP, Jacketed Hollow Point.

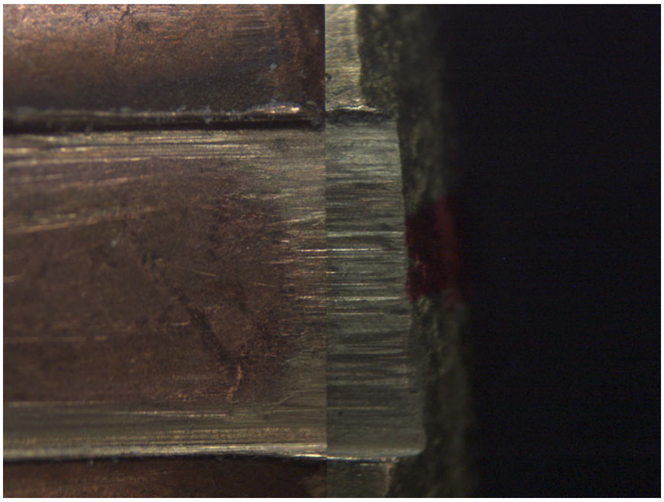

FIG. 1—*Kit # 22 Ex 1098 to Ex 1267 28X, LIMP 1.*[1]

[1] Land Impression

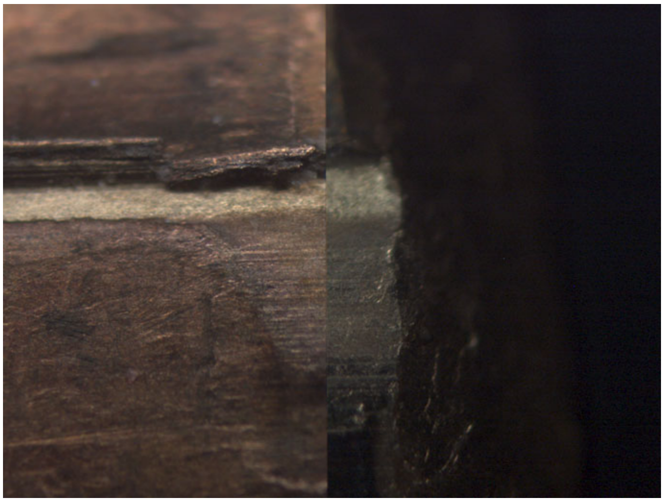

FIG. 2—*Kit # 22 Ex 1288 to Ex 1124 55X, LIMP 2.*

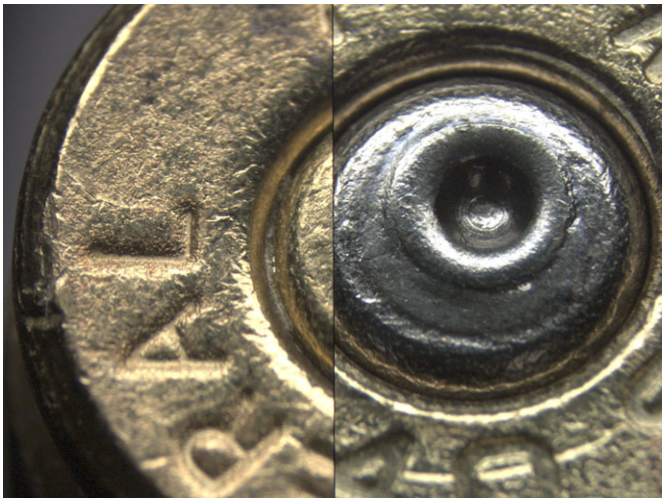

FIG. 4—*Kit # 27 Ex 1191 to Ex 1834 14X, BFM 2.*[2]

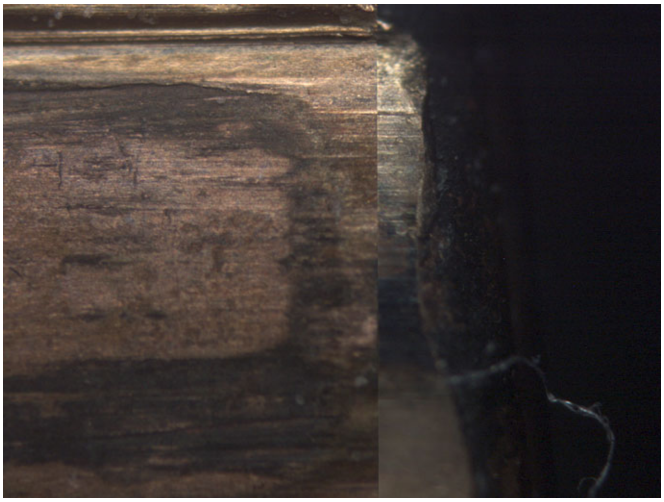

FIG. 3—*Kit # 22 Ex 1288 to Ex 1844 55X, LIMP 1.*

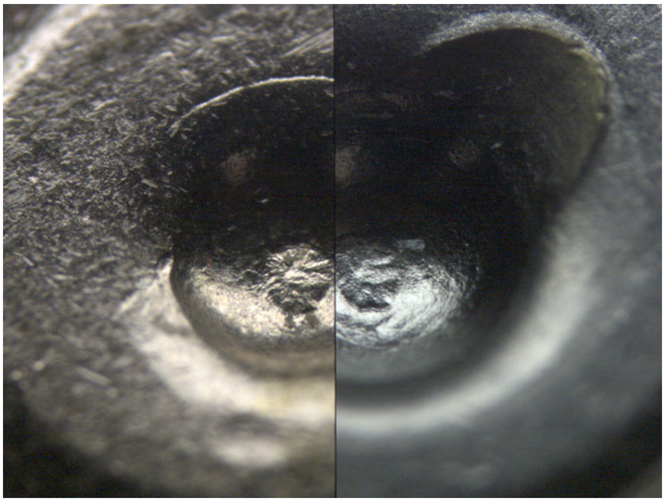

FIG. 5—*Kit #27 Ex 1238 to Ex 1760 35X, FPIM.*[3]

varying amounts of time to complete the test, based on phases of this research project, and it was requested that both answer sheets and kits be returned upon completion. Time duration was estimated to be between 2 and 12 months. Twenty-two different laboratories/laboratory systems across the country (and one abroad) were represented in the results received.

Results

We report two types of analyses in this section. First, we examine the overall error rates, sensitivity and specificity levels, in an aggregate fashion with no attention given to differences in examiners. Second, we provide additional analysis that looks at sensitivity and specificity levels as they are distributed across the 31 examiners, as well as the effect of years of experience on identification performance.

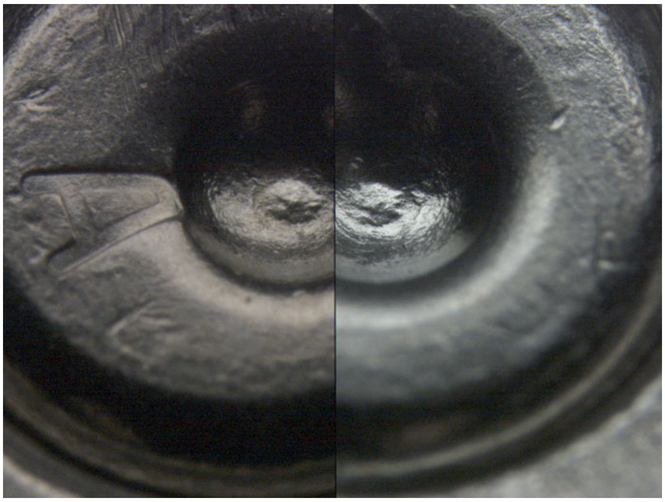

FIG. 6—*Kit #27 Ex 1341 to Ex 1760 28X, FPIM.*

[2] Breechface Mark

[3] Firing Pin Impression Mark

TABLE 3—*Compiled overall study report.*

Total # Kits Distributed	Total # of completed kits returned*	% Participation	# of Laboratories Represented[‡]	Avg. years of Experience	Min Years of Experience	Max Years of Experience
47	34[†]	31/47 = 0.659 = 65.9%	22	12.1 years	3 years	46 years

Specimen Population	Total # Identifications Reported	Total # of True Identifications of Kits returned	Total # False Identifications	Total # Comparisons	Total # Inc Reported
Cartridge Cases	191	199	1	693	39
Bullets	156	207	0	955	165

Specimen Population	Total # Eliminations Reported	Total # of True Eliminations of Kits returned	Total # True Eliminations Adjusted	Total # False Eliminations
Cartridge Cases	406	400	441	3
Bullets	519	360	619	1

Specimen Population	Sensitivity	Specificity	Error Rate: False Identification	Error Rate: False Elimination	Overall Error Rate
Cartridge Cases	190/199 = 0.955	403/441 = 0.914	1/693 = 0.144%	3/693 = 0.433%	5/1648 = 0.303%
Bullets	156/207 = 0.754	518/619 = 0.837	0	1/955 = 0.105%	

Overall Sensitivity	Overall Specificity
346/406 = 85.2%	921/1060 = 86.8%

*This refers to answers that have been submitted not necessarily physical kit.

[†] The data from three kits were not used in the calculations for noted reasons (see report notes page).

[‡] Of returned kits.

Aggregate Analysis

Table 3 summarizes the analysis of data. In addition to the overall error rate, we also measured sensitivity and specificity. Sensitivity was defined as the number of positive conclusions (identifications) actually obtained from the test divided by the number of true positives (true identifications). The sensitivity of a study is important because it relates to the test's ability to identify positive results – in this case positive associations of like origin when they exist. It measures the proportion of actual positives that are correctly identified. Specificity was also measured in this study. The specificity is the number of negative conclusions (eliminations) actually obtained from a test divided by the number of true negatives possible (true eliminations). The specificity measures the proportion of negatives which are correctly identified. This relates to a test taker's ability to properly identify negative results.

Table 3 shows the sensitivity and specificity of the cartridge cases and bullets. The sensitivity and specificity for cartridge cases was 95.5% and 91.4%; for bullets, 75.4% and 83.7%. The false-positive and false-negative error rate for cartridge case evaluation was calculated by taking the number of false identifications or false eliminations over the total number of cartridge case comparisons made using the most conservative approach. A false-positive result is one in which an association is made which is incorrect. Likewise, a false-negative result is when an association is not made, when it should be. The false-positive error rate recorded for the evaluation of cartridge cases in this study was 0.144%, and the false-negative error rate was 0.433%. For bullets, the false-positive error rate was 0.0% and false-negative error rate was 0.105%. The overall error rate was 0.303%, overall sensitivity 85.2%, and overall specificity 86.8% (see Table 4).

A total of 204 inconclusive results (neither identification nor elimination) were reported for the evaluation of the cartridge cases and bullets/bullet jackets in this study, for which a true identification or a true elimination could have been made. Such a response is scientifically valid and acceptable, indicating an insufficient agreement or disagreement of individual microscopic marks of value. It was observed that there were 68 inconclusive responses that should have been identifications and 136 inconclusive responses that should have been eliminations. Of the 68 inconclusive responses that should have been identifications, 62 (91.2%) were for bullets and six (8.8%) for cartridge cases. Of the 136 inconclusive responses that should have been elimina-

TABLE 4—*Descriptive statistics for years of experience, sensitivity, and specificity (*N = *31).*

		Cartridge		Bullet		Overall	
	Years Exp.	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Mean	12.19	0.96 (0.08)	0.93 (0.14)	0.75 (0.23)	0.85 (0.09)	0.85 (0.14)	0.88 (0.10)
Min	3.00	0.71	0.50	0.17	0.67	0.50	0.64
Max	46.00	1.00	1.00	1.00	1.00	1.00	1.00
25%	5.00	0.86	0.93	0.57	0.79	0.67	0.82
50%	11.00	1.00	1.00	0.83	0.80	0.92	0.90
75%	16.00	1.00	1.00	1.00	0.93	1.00	0.96

Standard deviations are shown in parentheses. 95% Confidence Interval for bullet sensitivity = 0.68–0.84. 95% Confidence Interval for bullet specificity = 0.81–0.88. Confidence intervals not reported for cartridge cases due to non-normal distributions.

tion responses, 103 (75.7%) were for the evaluation of bullets and 33 (24.3%) for cartridge cases.

There are several variables that can affect how a particular tool marks an object. In this study, these variables include pressure differences between test fires, wear in the microscopic marks, differences in cartridge materials, and use, abuse, and debris, which can create a level of ambiguity in the individual microscopic marks from consecutive test fires within a single firearm. Also, there are internal variables such as the policies and procedures that laboratories use which dictate when an examiner can declare an elimination when the firearm is absent.

Additional Analysis

In this section, we engage in further analysis of sensitivity and specificity, this time looking at results across the 31 examiners. No analysis of error rates is appropriate in this fashion, as there were so few errors that no meaningful variability exists across examiner kits. The questions we address here are as follows: what does the variation in sensitivity and specificity look like across the examiners; and to what extent is there a relationship between years of experience of the examiners and their sensitivity and specificity levels.

Table 4 reports descriptive information on these variables, including their mean, standard deviation, minimum and maximum, and quartiles. Note that for sensitivity and specificity levels, the means will be slightly different than the overall levels reported in the first section of our findings. This is because the kits varied in their denominators, and thus, averaging 31 kits with different denominators will result in means that vary from overall levels that are calculated without taking into account differences across examiners/kits.

Distributions of sensitivity and specificity for bullets and cartridge cases were different. For bullets, both measures were on a normal distribution, according to a one-sample Kolmogorov–Smirnov test. For cartridge cases, however, the null hypothesis of normality was rejected for both measures. This is primarily because more examiners were likely to have a perfect (1.0) sensitivity and specificity for cartridge cases than for bullets. For bullets, only nine of 31 kits were associated with a perfect sensitivity, whereas for cartridge cases, 23 kits were perfect in sensitivity. Similarly, for specificity, 5 of 31 kits were perfect for bullets, compared to 20 for cartridge cases. Thus, we could only calculate confidence intervals across examiners for bullets, not for cartridge cases. A 95% confidence interval for bullet sensitivity levels ranges from 0.68 to 0.84. A 95% confidence interval for bullet specificity levels ranges from 0.81 to 0.88. Sensitivity ratings, then, varied much more dramatically across examines than did specificity.

Years of experience varied from 3 to 46. Although the sample size was too small to make general conclusions about the relationship between years of experience (YOE) and sensitivity and specificity, we still performed some limited analysis. The correlation between YOE and both sensitivity levels was near zero. However, the correlation between YOE and both specificity levels was approximately 0.25, with a p-level of 0.08 (not significant at 0.05, but close). To see whether there may be a more complex (rather than linear) relationship between YOE and sensitivity and specificity, we broke the levels down by four categories of YOE, which are consistent with the quartiles in Table 4. Table 5 reports this analysis (for bullets only; no meaningful patterns emerge with cartridge cases). For sensitivity, levels jump up markedly from those at the beginning of their career (0.63 for 1–5 YOE to 0.84 for those with 6–11 YOE), and then tails off back to 0.71 for those with more than 17 YOE. The pattern is quite different for specificity, with a general gradual increase from an average of 0.82 for those with 1–5 YOE to a 0.90 for those with 17+ YOE, with a little movement in the middle categories.

TABLE 5—*Sensitivity and specificity by years of experience.*

Years of Experience	Bullet Sensitivity	Bullet Specificity
1–5 (N = 8)	0.63	0.82
6–11 (N = 9)	0.84	0.85
12–16 (N = 7)	0.81	0.82
17–46 (N = 7)	0.71	0.90

Discussion and Conclusions

The number of true eliminations and true identifications varied from test to test. This design provided a realistic study approximating how examiners perform their actual case work. In Giroux's study of consecutively manufactured screwdrivers, he took 80 questioned tool marks and eight known tool marks which were produced using three consecutively manufactured screwdrivers (18). Ten questioned tool marks were randomly numbered, and the eight known test marks were sent to eight different examiners. Examiners were asked to compare the known mark to the unknown marks and render a conclusion. Within this test, there were 29 true identifications and 51 true eliminations. The false-positive error rate was 0% and false-negative error rate 3.4%. The sensitivity was reported as 75.9% and specificity 15.7%, suggesting that examiners are far less likely to eliminate based on the individual characteristics than to make identification when the latter is possible. However, the decision of inconclusive (or no-conclusions) is not accounted for in this test. Such a result is common because in the way that this test is constructed a response of no-conclusion does not have a direct impact on how the results are tabulated. Such a limitation to the test can produce results that are unrealistic to the nature of typical firearm/tool mark examinations, the prediction of error rate, and the number of actual comparisons made.

In this study, by contrast, it was observed in the tabulation of the results that a cause-and-effect exists within the scope of the examination when an inconclusive (neither identification nor elimination) response is reported. During the evaluation of the data, it was observed that for the examination of cartridge cases in this study, which is similar to casework, of the 400 true eliminations that existed (within the 31 tests) as the test was originally designed a total of 406 were reported, three of which were false, however, that leaves three above what was theoretically possible. Yet, when the six inconclusive responses that should have been identifications are factored in, there is an adjustment of 41 additional elimination responses that are now possible based on the inconclusive response. As an examiner renders an opinion of inconclusive, they are now obligated to compare additional items within a group that otherwise would not necessarily need to be compared if an identification or elimination had been made. By default, this creates an independent group in the process requiring its own set of comparisons. This is the case in a number of the comparisons made within this study.

For example: Group A consists of items 1, 2, 3, 4; and Group B consists of items 5, 6, 7, 8. Traditionally, within group A, there are three comparisons, and within group B, there are three comparisons and one comparison between groups A and B.

Because the elimination of any one item in Group A to any one item in Group B separates the two groups, therefore theoretically there is only one true elimination possible as designed. However, the reality is that it is possible and also a correct response to evaluate group A and be inconclusive in items 1 and 2 to 3 and 4. Group A (1, 2) and Group C (3, 4) and Group B (5, 6, 7, 8), and now Group A and B eliminate and Group C and B eliminate, while Group A and Group C are inconclusive. So although as designed, there was only one elimination possible, based on the response of neither identification nor elimination, which is not incorrect, there are now two true eliminations possible. This same cause-and-effect occurs for each time an inconclusive response exists that should have been an identification.

A re-evaluation of the data, taking this information into consideration, is what produced a higher aggregate specificity measurement of 91.4% for cartridge cases and 83.7% for bullets, than what would be typically expected, based on previous studies. There were 406 cartridge case eliminations reported, three of which were false, leaving 403 reported eliminations (406-3 = 403). There were 41 eliminations created from the six inconclusive responses that should have been identifications, leaving 362 actual elimination responses reported (403-41) ignoring the inconclusive responses. This creates a specificity measurement of 90.5% (362/400). In the case of the bullet evaluation, there were 519 bullet eliminations, one of which was false, leaving 518 reported eliminations. There were 259 eliminations created from the 62 inconclusive responses that should have been identifications, leaving 259 actual eliminations responses reported (518-259), ignoring the inconclusive responses. This creates a specificity measurement of 71.9% (259/360) for the evaluation of bullets. Such a measurement is consistent with what is expected based on past studies; however, it is not an accurate assessment of the level of comparisons actually made in casework. The realistic evaluation shows that as comparisons are made, an examiner becomes more and more specific in his assessment of the information. Although it has been argued that examiners are less likely overall to make eliminations, the results of this study indicate that in actual casework, overall they are 86.8% likely to make elimination when elimination can be made, and that examiners are 85.2% likely to make identification when identification can be made.

While the error rate is the most important measure of the quality for forensic comparison examinations, sensitivity and specificity are also indicators of a test's quality and should be given fair consideration. The overall results from this study were different from previous study results. They provide a more accurate data point indicative of the capabilities of the discipline of firearm and tool mark identification to make conclusive identifications and exclusions with regard to the origin of a mark. This study assessed the overall scientific validity and quality of the examination of ammunition components. Although definitely useful in court and of value scientifically, caution should be used when applying these results to estimate error rates in a generalized sense. A number of factors, such as a laboratory's quality assurance program (which includes verifications and peer review), would influence error rates in casework.

The participant pool for this study (N = 31) was fairly impressive when considering how much time and effort each examiner volunteered to the study. A number of the participants had some type of formal CMS training, although pattern matching was primarily used within the test, with only two participants noting the use of CMS during their examination. With such variability in mind, it only adds weight to the results that indicate that firearm and tool mark identification does follow valid methodology and that proper training provides each examiner with the skills necessary to make the correct associations.

An official questionnaire responded to by examiners once the test was completed indicated that the general feeling was that this test did take considerably longer to complete than other tests they had taken or had anticipated. The level of difficulty of the test was also commented on as being more difficult than other studies and more representative of actual casework type distribution, which was the goal of this study.

During the past several years, significant research has been published in the evaluation of fired ammunition components. This research has included the test fire of firearms numerous times to evaluate the changes in microscopic characteristics observed on the fired bullets and cartridge cases, as well as the test firing of consecutively rifled firearms to determine whether the projectiles could be identified to the barrel from which they were fired. It has been found in every research project involving such examinations that a properly trained firearm and tool mark examiner has the ability to identify a surface marked by a tool back to the particular tool that made the mark, and likewise eliminate a particular tool on the same basis. However, many of these studies have not included the impact of the inconclusive response when evaluating their data. As indicated through this study, a conclusion of neither identification nor elimination adds weight and value to the clear response of identification or elimination. Examiners are trained to be more conservative when making their evaluations and a response of inconclusive means that a particular examiner has not seen enough information to say that two items have been marked by the same tool or that they have not been marked by the same tool. Courts should be more inclined to take validation studies into greater consideration when evaluating the probative value of testimony and evidence, when the studies are conducted in a fashion that resembles actual casework.

Acknowledgments

The authors would like to take this opportunity to express their warmest and sincerest thanks to the San Francisco Police Department Criminalistics Laboratory Firearm and Tool Mark Unit for the resources, time, and personnel to complete this research project. An expression of thanks to Ron Nichols for his assistance in the compilation and interpretation of the data acquired during this study. Thanks to Kelsey Ellison, SFPD Firearm Unit intern, for her assistance in constructing the "laboratory-made" bullet retrieval trap, firing of test specimens, and creating test packets. Special thanks to the National Firearms Examiner Academy (NFEA) for the foundational training and work on this project. Additionally, special thanks go out to all of the firearm and tool mark examiners who participated in this study. Participating in this study took a great deal of time and commitment to complete for the advancement of the science.

References

1. Matty W. A comparison of three individual barrels produced from one button rifled barrel blank. AFTE J 1985;17(3):64–9.
2. Brundage D. The identification of consecutively rifled gun barrels. AFTE J 1998;30(3):438–44.
3. Thompson E, Wyant R. Knife identification project. AFTE J 2003;35 (4):366–70.
4. Bachrach B, Jain A, Jung S, Koons RD. A statistical validation of the individuality and repeatability of striated tool marks: screwdrivers and tongue and groove pliers. J Forensic Sci 2010;55(2):348–57.

5. Weller T, Zheng A, Thompson R, Tulleners F. Confocal microscopy analysis of breech face marks of fired cartridge cases from 10 consecutively manufactured pistol slides. J Forensic Sci 2012;57(4):912–7.
6. Biasotti AA. A statistical study of the individual characteristics of fired bullets. J Forensic Sci 1959;4(1):34–50.
7. Biasotti AA, Murdock JE. Criteria for the identification or state of the art of firearms and toolmark identification. AFTE J 1984;16(4):16–24.
8. Faden D, Kidd J, Craft J, Chumbley LS, Morris M, Genalo L, et al. Statistical confirmation of empirical observations concerning toolmark striae. AFTE J 2007;39(3):211–20.
9. Chumbley LS, Morris MD, Kreiser MJ, Fisher C, Craft J, Genalo LJ, et al. Validation of toolmark comparisons obtained using a quantitative comparative, statistical algorithm. J Forensic Sci 2010;55(4):953–61.
10. Neel M, Wells M. A comprehensive statistical analysis of striated tool mark examinations part 1: comparison known matches and known non-matches. AFTE J 2007;39(3):176–98.
11. Weavers G, Neel M, Buckleton J. A comprehensive statistical analysis of striated tool mark examinations part 2: comparing known matches and non-known matches using likelihood ratios. AFTE J 2011;43(2):137–45.
12. Buckleton J, Nichols R, Triggs C, Weavers G. An exploratory Bayesian model for firearms and tool mark interpretation. AFTE J Fall 2005;37 (4):352–61.
13. Gutowski S. Error rates in the identification sciences. Forensic Bull, 2005;23–9. ISSN 1447-6673.
14. National Academy of Sciences. Strengthening forensic science in the United States: a path forward. Washington, DC: The National Academies Press, 2009.
15. Budlowe B, Bottrell M, Bunch S, Fram R, Harrison D, Meagher S, et al. A Perspective on errors, bias, and interpretation in forensic sciences and direction for continuing advancement. J Forensic Sci 2009;54(4):798–809.
16. Smith ED. Cartridge case and bullet comparison validation study with firearms submitted in casework. AFTE J 2005;37(2):130–5.
17. http://www.random.org/sequences
18. Giroux B. Empirical and validation study: consecutively manufactured screwdrivers. AFTE J 2009;41(2):153–155.

Additional information and reprint requests:
Tasha P. Smith, B.S.
San Francisco Police Department
Criminalistics Laboratory
Firearm and Tool Mark Unit
1245 3rd Street, Building 606
San Francisco, CA 94158
E-mail: tbaham06@gmail.com

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Appendix A.