Received: 27 July 2020 | Revised: 9 September 2020 | Accepted: 1 October 2020

DOI: 10.1111/1556-4029.14602

JOURNAL OF FORENSIC SCIENCES

PAPER

Criminalistics

Results of the 3D Virtual Comparison Microscopy Error Rate (VCMER) Study for firearm forensics

Chad Chapnick BScEE[1] | Todd J. Weller MS[1,2] | Pierre Duez MASc[1] | Eric Meschke MScCS[1] | John Marshall BS, MBA[3] | Ryan Lilien MD, PhD[1]

[1]Cadre Research Labs, Chicago, IL, USA

[2]Weller Forensics, Burlingame, CA, USA

[3]Royal Canadian Mounted Police (ret), Ottawa, Ontario, Canada

Correspondence
Ryan Lilien, MD, PhD, Cadre Research Labs, Chicago, IL 60091, USA.
Email: Ryan.Lilien@CadreResearch.com

Funding information
Funding support provided by the National Institute of Justice (NIJ) Grant No. 2017-IJ-CX-0024. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice.

Abstract

The digital examination of scanned or measured 3D surface topography is referred to as Virtual Comparison Microscopy (VCM). Within the discipline of firearm and toolmark examination, VCM enables review and comparison of microscopic toolmarks on fired ammunition components. In the coming years, this technique may supplement and potentially replace the light comparison microscope as the primary instrument used for firearm and toolmark examination. This paper describes a VCM error rate and validation study involving 107 participants. The study included 40 test sets of fired cartridge cases from firearms with a variety of makes, models, and calibers. Participants used commercially available VCM software which allowed digital data distribution, specimen visualization, and submission of conclusions. The software also allowed participants to annotate areas of similarity and dissimilarity to support their conclusions. The primary cohort of 76 qualified United States and Canadian examiners that completed the study had an overall false-positive error rate of 3 errors from 693 comparisons (0.43%) and a false-negative error rate of 0 errors from 491 comparisons (0.0%). This accuracy is supplemented by the participant's provided surface annotations which provide insight into the cause of errors and the overall consistency across the independent examinations conducted in the study. The ability to obtain highly accurate conclusions on test fires from a wide range of firearms supports the hypothesis that VCM is a useful tool within the crime laboratory.

KEYWORDS
3D imaging, cartridge cases, error rate, firearms identification, software, surface metrology, validation, virtual comparison microscopy, virtual examination

1 | INTRODUCTION

The comparison of fired ammunition components in pursuit of source attribution relies upon accurate visualization and comparison of microscopic features. This comparison has historically been accomplished using Light Comparison Microscopy (LCM). LCM uses a comparison microscope, consisting of two microscopes joined with an optical bridge. This instrument allows examiners to directly compare the microscopic marks on two specimens. LCM utilizing a comparison microscope has been used for over a hundred years and remains the primary method of comparison for microscopic toolmarks present on fired ammunition components [1]. In the past decade, a large body of research has been conducted in firearm examination using emerging technology of 3-dimensional (3D) surface topographies and computer comparison algorithms [2–11]. The use of 3D topographies is now termed Virtual Comparison Microscopy (VCM). These early research results suggest great potential for VCM to elevate microscopic toolmark comparison. Of course, incorporation of new technology into a laboratory requires validation and establishment of error rates. It is only by establishing well-founded error rates that the technology will truly benefit the criminal justice system. Technology with lower error rates gives confidence toward overall reliability of analysis.

In this paper, we describe a large research study involving over one hundred participants to validate and to establish error rates for one VCM platform. This study includes thirty-six different models of firearms spanning complexities representative of casework. A randomized study structure was employed to balance a large number of test sets with a low test burden on each participant. Participants utilized custom VCM software with digital data files. Results including both source conclusions and annotations of the specific toolmarks that led to those conclusions were submitted electronically. These results were summarized in numeric and visual form. The effectiveness of VCM can be quantified by the direct comparison of VCM error rates to those previously published for LCM.

The remainder of this paper is structured as follows: We start with a discussion of Virtual Comparison Microscopy. We then introduce the Virtual Comparison Microscopy Error Rate (VCMER) study. We next discuss methods and present subsections detailing scan acquisition, visualization software, test sets, pretest workshop, and study design. After presenting study results, we close with discussion.

2 | MATERIALS AND METHODS

2.1 | Virtual comparison microscopy

Virtual comparison microscopy (VCM) involves the comparison of two specimens that have been measured using 3D scanning instrumentation. Once surfaces have been measured, only the digital files and software are necessary to compare two items. This concept was described in 2006 by Senin et al. [12] for use in firearm forensics. In the proposed system, the microscopic toolmarks present on fired ammunition are measured, digitally saved, and compared using custom computer software. The software replicates many of the visualization functions of the traditional light comparison microscope including the ability to position the sample and to adjust lighting. As a digital technology, VCM implements a number of features not possible with traditional microscopic methods. The work presented in this paper is only possible because of VCM's ability to access identical specimens and VCM's ability to annotate surfaces as means of documenting source conclusions. Additional uses of the VCM platform validated in this manuscript are described by Duez et al. [10].

Laboratory implementation of VCM requires validation and establishment of error rates. This can only be accomplished with VCM based studies. Duez et al. describe the first validation study of the Cadre TopMatch-3D VCM system (Cadre Forensics; Chicago, IL) [10]. They note that VCM has potential advantages over traditional light microscopy when used to archive and compare evidence, train examiners, proctor validation studies, and document comparisons and verifications. The Duez et al. study involved 56 participants (46 qualified examiners and 10 trainees) and two test sets. The examiners reported no false-positive or false-negative results, and one trainee reported two false identifications. Other internal validation studies have been conducted by the Royal Canadian Mounted Police (RCMP; Ottawa, Canada) (L. Knowles, conference presentation, AFTE Training Seminar, May 2019, Nashville TN) and Federal Bureau of Investigation Firearms/Toolmarks Unit (FBI FTU; Quantico, Virginia) (E. Smith, conference presentation, AFTE Training Seminar, May 2017, Denver CO) crime laboratories. The study described in this manuscript builds from the foundation established by these initial studies.

Highlights

- This paper studies Virtual Comparison Microscopy (VCM) for the examination of toolmarks.
- VCM technology is likely to become a central part of firearm and toolmark examination.
- The study's primary cohort involved United States and Canadian firearm and toolmark examiners.
- The low error rates of this study provide strong support for the validity and use of the studied VCM platform.
- The results lend support to the foundational principles of the firearm and toolmark examination discipline.

The study described below exploits several inherent advantages of VCM. First, virtual comparison microscopy ensures that all participants receive the same exact surfaces for comparison without the sample to sample variance that occurs in some LCM-based studies which use multiple test fire sets from the same firearms. Second, VCM allows efficient test sample distribution. For example, a LCM version of the VCMER study described below would have required the collection, labeling, and mailing of approximately 9600 samples (200 initial participants × 3 samples in each set × 16 sets per test) rather than the 120 samples (40 sets × 3 samples per set) used with VCM. Finally, VCM allows insight into the decision-making process for each submitted conclusion via surface annotations and annotation summary maps as described in the Test Implementation section below.

2.2 | 3D scanning instrument

Virtual comparison microscopy requires the use of measured 3D surface topographies. In this study, the 3D surface topographies of all study specimens were measured using the TopMatch-3D Scanner (Cadre Forensics—Chicago, IL) and analyzed using the Cadre VCM software. The TopMatch instrument uses enhanced photometric stereo for 3D scan acquisition [13,14]. To improve overall measurement quality, the system uses an elastomeric pad coated with a thin layer of pigment. The pigmented layer is pressed against the item to be measured and conforms to the surface (Figure 1). The pigmented pad temporarily changes the surface reflectivity of the target to allow accurate measurement of traditionally difficult geometries such as aperture shear, fine striations, and steep slopes. Information about this instrumentation and initial validation studies have previously been published [10,11,13,14]. In this study, the breech face impression and aperture shear marks of each expended cartridge case were measured at approximately 1.8 micron/pixel lateral resolution and sub-micron depth resolution. At the time of this study, the Cadre TopMatch instrument did not regularly capture certain features on fired cartridge cases, such as firing pin impressions. It should be noted that at the time of this publication, firing pins can now be measured using the Cadre system due to the use an updated elastomeric pad.

2.3 | Study design

The power of error rate studies and validation studies is related to the breadth and complexity of specimens included. In theory, a validation study should represent all possible items encountered in casework. In practice, a validation study must balance the need for a large number and variety of specimens with the time required for completion. Most studies rely on volunteer examiners who participate while continuing normal laboratory responsibilities. Studies which are extremely large are likely to have low completion rates complicating the statistical interpretation of the results. The VCMER study uses a Balanced Incomplete Block Design (BIBD) structure to optimize its efficiency [15]. BIBD tests are often used to evaluate a large number of experimental variants when not all variants can be tested by all participants. In the context of the VCMER study, each participant examines a block (or group) of multiple test sets where each test set is a group of two knowns and one unknown. The term *incomplete* means that not all test sets are evaluated by each participant (e.g., the blocks are not complete). This incompleteness results in each participant not having to analyze all test sets. The term *balanced* means that every pair of test sets are seen by the same number of participants. Balancing the pairs allows better comparison of test set performance. In most studies, the BIBD balance properties are not perfectly achieved. That is, some test sets will be evaluated by slightly more participants than others. One cause of this imbalance is that some individuals who sign up for a study (and are therefore assigned test sets to analyze) never complete the analysis. This slight incompleteness does not impair the computation of performance statistics. The VCMER study BIBD structure was randomized in both test set order and test set inclusion. This randomization created stronger test integrity because each participant completed a different test.

A total of forty test sets were included in the VCMER study (note: forty-one test sets were created; however, one set was withdrawn because the items were inadvertently labeled with the wrong caliber. No false identifications or false eliminations were reported from

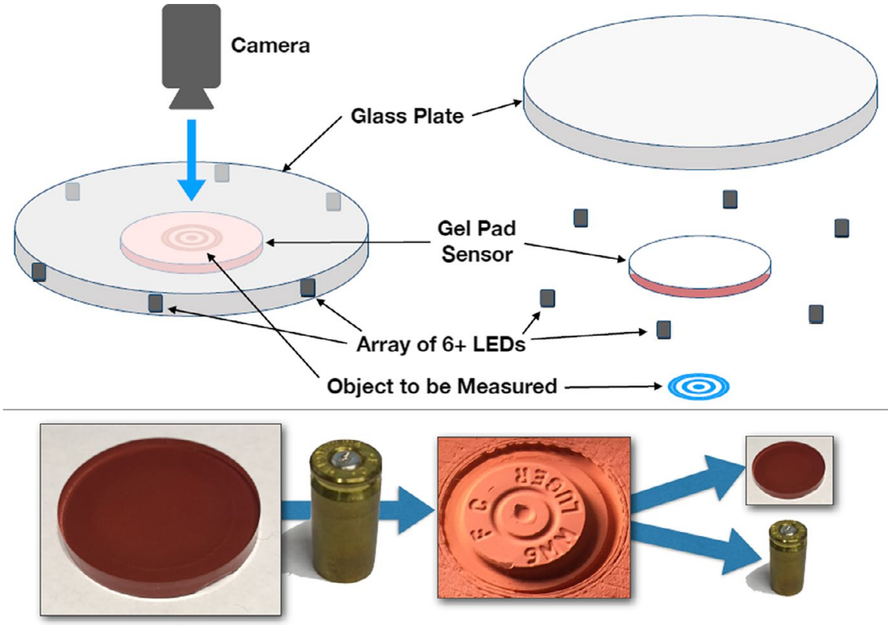
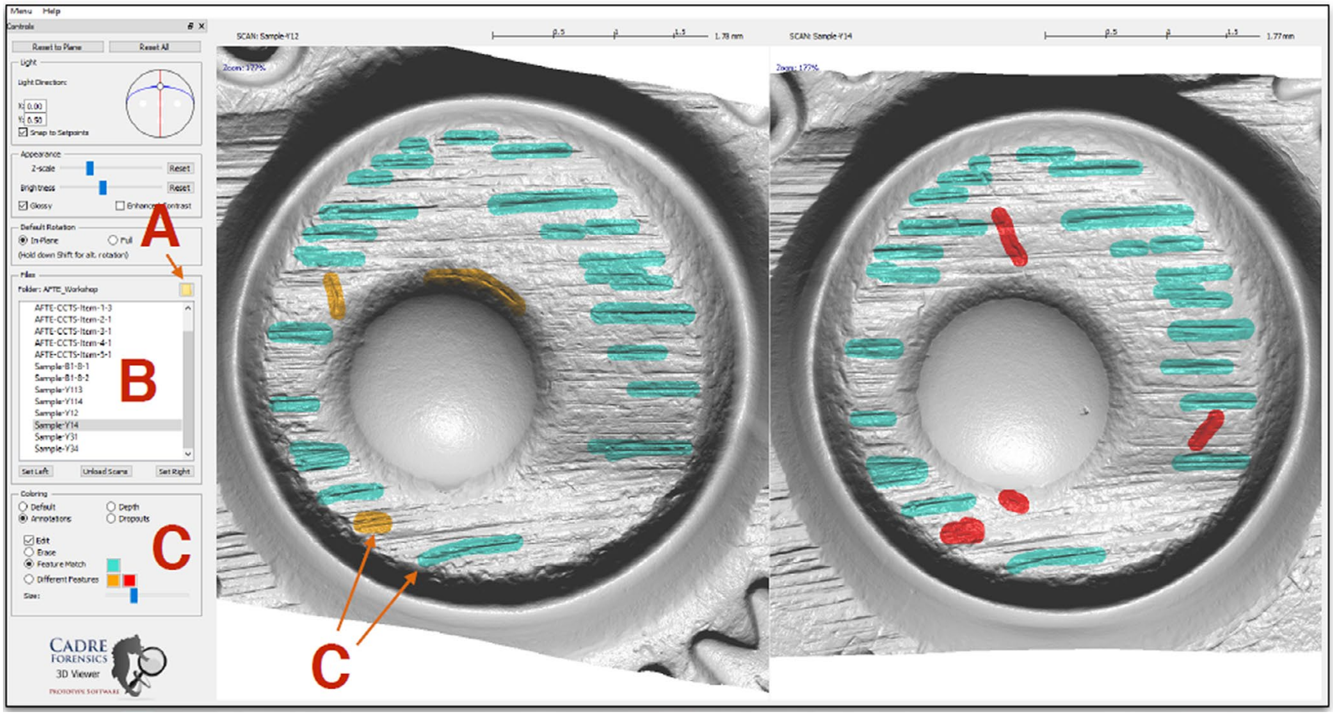


FIGURE 1 GelSight Scanning Setup. The 3D scanning technique (GelSight) is based on the use of a silicone elastomeric pad with embedded micron-scale thick layer of pigment. (Top Row) The Gel Pad sensor is placed between a glass plate and the item being imaged. When the object to be measured is raised into the gel, the gel and pigment conform to the object (Bottom Row). The gel's pigment removes all unwanted surface reflectance properties (e.g., metal specularity). LED lights are sequentially illuminated and a set of captured images is combined into an accurate 3D surface. In the TopMatch scanner, this is an automated process with the camera, lens, glass plate, and LEDs all being fixed and automated. (Bottom Row) A cartridge case is pressed into a gel pad (5 mm thick, 38 mm diameter) allowing the pigment to conform to the cartridge surface. After scanning, the cartridge is removed and the gel can be used again [Color figure can be viewed at wileyonlinelibrary.com]



Light Direction:
X: 0.00
Y: 0.50
☑ Snap to Setpoints
☑ Glossy
□ Enhance Contrast
Default Rotation
● In-Plane ○ Full
(Hold down Shift for alt. rotation)
Folder: APTE_Workshop
Coloring
○ Default ○ Depth
● Annotations ○ Dropouts
☑ Edit
○ Erase
● Feature Match
○ Different Features

FIGURE 2 Virtual Comparison Microscopy (VCM) Software. The VCM software provides a virtual comparison microscope. Examiners can adjust the virtual light position, manipulate the cartridge case orientation, position, and zoom (locked or unlocked). In a typical workflow, the user first selects a folder of scans (A) then sends individual scans to the left or right view panel (B). Pairs of cartridge cases can be annotated (C) to indicate regions of similarity or difference. Annotations and high-resolution screenshots can be saved for use in presentations. The VCM software was modified to incorporate a testing mode on which the VCMER study was conducted [Color figure can be viewed at wileyonlinelibrary.com]

this withdrawn test set). A total of seventeen different manufacturers, consisting of thirty-six different models and three different calibers were selected. 55% of the firearms were 9 mm Luger caliber, 32.5% were 40 S&W caliber, and 12.5% were 45 Auto caliber. Each test set consisted of three samples: two known expended cartridge cases and one unknown expended cartridge case. Seventeen of the test sets were known matches (KM), and twenty-three of the test sets were known non-matches (KNM). The items within each KNM set had the same class characteristics. Test fires were manually attributed a level of "complexity". In this study, complexity refers to the quality and quantity of individual marks present on the scan. High complexity samples may be expected to have a higher inconclusive rate than low complexity samples. Of the forty test sets, 30% were rated "low" complexity, 38% were rated "medium" complexity, and 32% were rated "high" complexity. For breech face class characteristics, 48% had linear features, 23% had granular features, and 8% had arched features. Aperture shear marks were present in 45% of the test sets. The complexity metric, in combination with different class characteristics, makes, models, and caliber ensured that the study covers a variety of samples that examiners would expect to see in routine casework.

The BIBD study structure involved each participant receiving a random selection of sixteen of the forty test sets. We anticipated the sixteen test sets to require no more than a five to eight-hour time commitment from each participant. It is possible through a random BIBD design that a participant could end up randomly with zero known matches (KMs) or zero known non-matches (KNMs). It was important each participant examined a mixture of KM and KNM and so we enforced that all participants had between four and eleven known matches in their test. This balancing maintains the overall BIBD design criteria (Figure 2).

2.4 | Participants

A total of 107 participants completed the test. Participants were asked to self-report demographic information via an electronic worksheet. The USA had 56 qualified examiners and 7 non-qualified examiners, Canada had 20 qualified examiners and 1 non-qualified examiner, and the rest of the world (herein referred to as "International") had 21 qualified examiners and 2 non-qualified examiners. We defined "Qualified examiner" to mean that the individual is qualified by their laboratory to perform independent casework. Due to the low number of participants from countries other than the USA and Canada, the identity of these countries and number of individuals from each are not provided to avoid unintentionally unmasking the anonymity of test takers. Participants were asked about their years of experience. 25 (23%) had three or fewer years of experience in firearm and toolmark examination, 47 (44%) had between three and ten years of experience, and 35 (33%) had more than ten years of experience. Participants were asked if their laboratory was ISO accredited. 89% (56 of 63) of those from the USA, and 95% (20 of 21) of those from Canada reported affiliation with accredited laboratories. International participants had a lower rate of accreditation with 39% (9 of 23) reporting they were from accredited laboratories. Participants were asked their experience with VCM: participants from the USA reported 3% routine VCM use, 73% had used VCM a few times, and 24% had no prior experience. Canadian participants self-reported 5% routine VCM use, 62% had used VCM a few times, and 33% had no prior experience. International participants self-reported 9% routine VCM use, 43% had used VCM a few times, and 48% reported no experience. Finally, 71% of USA participants used a desktop computer (29% used a laptop). Among Canadian participants, desktop use was 100%.

2.5 | Test implementation

Each participant was provided VCM testing software (Figure 2). When accessing the software, each participant was required to log in with their assigned participant ID and webcode (i.e., their credentials). The first time a user enters this information the software requests permission to access the network to download the training and testing data sets assigned to that ID (users without network access were provided an alternative means of download). Participants were also provided a training manual. The manual guided participants through all essential software functionality using the training data. The last step of the training required successful completion of a proficiency-style test (three knowns compared to four unknowns). As described above, test sets were randomized, so while each participant was presented with test sets numbered one through sixteen the numbering was not consistent between participants. The software kept track of each test set and each participant. Participants were asked to complete their studies independently marking source conclusions, comments, and surface annotations (see below). Test takers were required to use the 5-point AFTE Range of Conclusions when expressing their conclusions for each test set comparison [16]. The definitions for each category of conclusion were provided within the comparison software (see Table 1). Upon completion of all sixteen test sets, the user uploaded their results, which was comprised of the conclusions, annotations, and questionnaire.

Virtual comparison microscopy facilitates participants' ability to support their conclusions with similarity and dissimilarity surface annotations. These annotations document the geometry used in reaching a source conclusion. Examiners were instructed to "mark (annotate) all individual marks (similar and different) that you used to reach the conclusion you marked on your Worksheet." In this manner each participant creates an individual annotation map for each test set. For each test set, it is possible to combine the individual annotation maps into a summary annotation map indicating the percentage of participants which annotated each section of the scan surface. Individual and summary annotation maps can be created for both similarity and difference annotations. Individual annotation maps are colored blue (similar) and red (dissimilar). Summary annotation maps use a "jet" colormap where hot colors (closer to red) indicate that a high percentage of examiners marked the indicated regions of the surface and cool colors (closer to blue) indicate a low percentage (Figure 3 – Top Row). As discussed below, the annotations provide insight into examiner decisions, both individually and collectively, highlighting areas used when reaching conclusions.

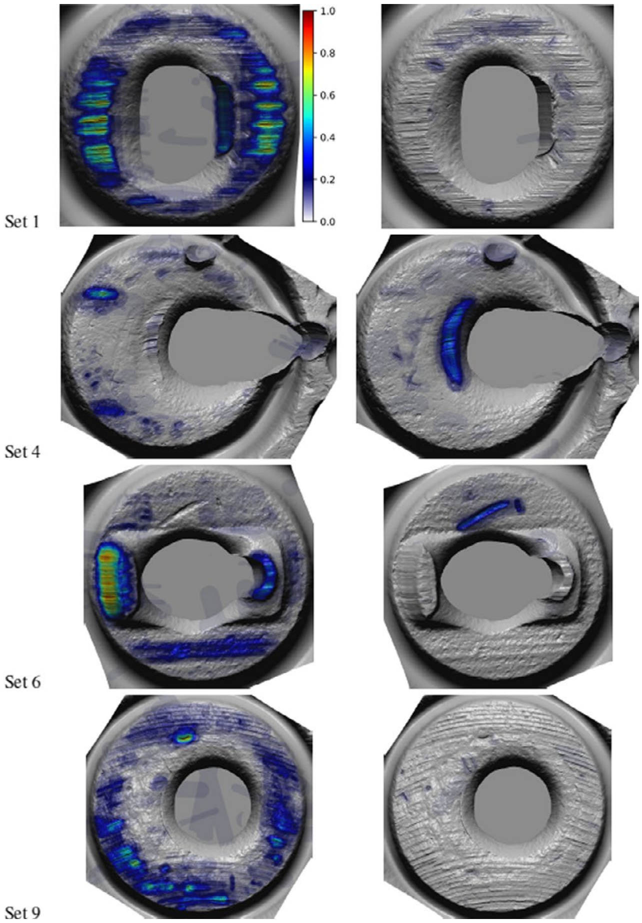


FIGURE 3 Summary annotation maps for four KM test sets for USCAN Qualified Examiners. Left is the similarity map, and right is the dissimilarity annotation map. Surface color indicates the portion of participants that marked the region as similar (left) or different (right). Color scale is shown next to Set 1 [Color figure can be viewed at wileyonlinelibrary.com]

3 | RESULTS

3.1 | USCAN results

For error rate performance metrics, our primary focus is the 76 qualified examiners from the USA and Canada (herein referred to as "USCAN"). This group of test participants represents the current and primary users of 3D/VCM technology for reaching source conclusions. Error rates metrics for other participant groups are also reported in the Additional Statistics section

TABLE 1 AFTE Range of Conclusions

AFTE Range of Conclusions	
Conclusion Category	Definition of Conclusion
Identification	Agreement of all discernible class characteristics and sufficient agreement of a combination of individual characteristics where the extent of agreement exceeds that which can occur in the comparison of toolmarks made by different tools and is consistent with the agreement demonstrated by toolmarks known to have been produced by the same tool.
Inconclusive-A	Agreement of all discernible class characteristics and some agreement of individual characteristics, but insufficient for an identification.
Inconclusive-B	Agreement of all discernible class characteristics without agreement or disagreement of individual characteristics due to an absence, insufficiency, or lack of reproducibility.
Inconclusive-C	Agreement of all discernible class characteristics and disagreement of individual characteristics, but insufficient for an elimination.
Elimination	Significant disagreement of discernible class characteristics and/or individual characteristics.

below. United States and Canadian qualified examiners reported three false-positive errors and no false-negative errors from a total of 1184 comparisons. The USCAN false-positive error rate for all reported conclusions was 3/693 = 0.43% (95% confidence interval: 0.09%–1.26%). All 95% confidence intervals were calculated using Clopper–Pearson Exact Binomial method and appear in parentheses after each statistic. The USCAN false-negative error rate for all reported conclusions was 0/491 = 0.0% (0.0%–0.75%). These results provide strong support of a relatively low error rate for qualified USCAN examiners using VCM. Table 2 shows the summary of reported source conclusions on the AFTE Range of Conclusions.

The source conclusions for individual KM test sets are shown in Table 3 (top). Similarity and difference annotation maps for the KM are presented in Figures 3 through 7. Almost all KM test sets have 100% identification. The annotation maps for the KMs demonstrate strong agreement in the specific areas of surface geometry used among independent examinations. For example, test set 1 (Figure 3) involved the comparison of specimens from a Springfield Armory XD45 with linear breech face marks. The annotation maps show orange, yellow, and green similarity coloring around certain irregularly shaped marks. This indicates the approximately 60%–70% of the individuals who examined and correctly identified these samples used these specific areas. Another example of strong agreement from the independent examinations is found in test sets 6, 14, 19, 20, 23, 25, and 28 (Figures 3–6). In each of these test sets, an aperture shear mark is present and the similarity annotation maps all show strong consensus in the use of this mark to support identification conclusions. A final example is illustrated by test set 9 (Figure 3). This test set involved the comparison of specimens from an Intratec Cat 9. The samples were well marked, with arch-shaped breech face marks carrying across nearly the entire surface. However, the green and yellow colors present in the similarity annotation maps show a preference by examiners to use breaks in these lines and other irregular-shaped marks. This may indicate that examiners are aware that milling, a source of arch-shaped toolmarks, can sometimes cause subclass toolmarks [17]. Most examiners did not rely on the entirety of the surface but instead used features that are likely to be from a random source (i.e., breaks in these toolmarks) and thus not likely to be found in the same geometric orientation on a non-matching sample.

TABLE 2 Summary of all reported conclusions by each participant group

	Conclusions				
Ground Truth	ID	INC-A	INC-B	INC-C	ELIM
USA and Canadian (USCAN) Qualified (76)					
KM	453	10	25	3	0
KNM	3	20	63	171	436
All Non-Qualified (10)					
KM	60	1	6	0	1
KNM	1	8	9	12	59
International Qualified (21)					
KM	137	1	4	1	3
KNM	4	4	6	10	156

The two KMs with lower identification rates are test sets 4 and 37. Test set 4 (Figure 3) involved the comparison of samples all fired from the same Kahr Arms pistol. The specimens from this test demonstrate a phenomenon that can occur where the aperture shear mark can be intermittent. The test was designed such that the knowns had no aperture shear mark and the unknown had an aperture shear mark present. This causes the knowns to appear different, especially with regard to overall surface geometry, to the unknown. Despite this difference, none of the USCAN qualified examiners reported a false elimination. This sample and the utility of individual annotation maps are further discussed in further discussed in Section 3.3. below. The second test set with a lower identification rate was number 37 (Figure 7), where examiners were provided specimens from a Beretta PX4 Storm. The Beretta is a firearm with a countersunk firing pin aperture. This leaves a small breech face area and typically no aperture shear. As such, the included samples

TABLE 3 Results by test set. The top is results of each test set for known matches (KM). The bottom is results for each test set for known non-matches (KNM). Test set 21 was removed because it was inadvertently labeled as 45 Auto caliber rather than 40 S&W caliber

Test Set	Caliber	Make/Model	ID	INC-A	INC-B	INC-C	ELIM	
KM								% ID
1	45 Auto	Springfield Armory XD 45	29	0	0	0	0	100
4	40 S&W	Kahr Arms K40	10	3	10	3	0	38.4
6	9 mm Luger	Glock 19	35	0	0	0	0	100
9	9 mm Luger	Intratec Cat 9	28	0	0	0	0	100
12	40 S&W	Kahr Arms CW−40	26	0	0	0	0	100
13	9 mm Luger	Hi-Point C9	29	0	0	0	0	100
14	9 mm Luger	Ruger SR9	29	0	0	0	0	100
19	40 S&W	Smith and Wesson SD40	29	0	0	0	0	100
20	45 Auto	Springfield Armory 1911-A1	32	0	0	0	0	100
23	9 mm Luger	Kel-Tec P−11	30	0	0	0	0	100
24	9 mm Luger	Norinco 213	29	0	0	0	0	100
25	9 mm Luger	Ruger SR9	25	0	0	0	0	100
28	40 S&W	Glock 22	26	0	0	0	0	100
29	9 mm Luger	FN Hi Power	28	0	0	0	0	100
30	40 S&W	Smith and Wesson SW40	31	0	0	0	0	100
32	45 Auto	Rock Island 1911	26	1	0	0	0	96.3
37	9 mm Luger	Beretta PX4 Storm	11	6	15	0	0	34.4
KNM								% Inc-C or ELIM
2	9 mm Luger	Kahr Arms MK9	0	2	1	10	16	89.7
3	9 mm Luger	Smith and Wesson 915	0	0	1	9	21	96.8
5	40 S&W	Glock 22	0	1	2	7	18	89.3
7	9 mm Luger	Smith and Wesson M&P 9	0	3	2	11	15	89.7
8	9 mm Luger	Glock 17	0	0	1	4	25	96.7
10	9 mm Luger	Smith and Wesson SW9VE	0	1	2	4	22	89.7
11	45 Auto	Glock 36	0	0	2	8	23	93.9
15	40 S&W	Star Bonifacio Firestar	0	0	0	6	26	100
16	40 S&W	Ruger P94	0	0	1	6	24	96.8
17	9 mm Luger	Taurus PT111	0	4	7	8	14	66.7
18	40 S&W	Smith and Wesson M&P 40	1	1	1	10	13	88.5
22	9 mm Luger	FEG PJK−9HP	0	0	0	6	28	100
26	9 mm Luger	Glock 26	0	1	0	7	24	96.9
27	9 mm Luger	Springfield Armory XD9	1	0	3	10	13	88.5
31	9 mm Luger	Glock 19	0	1	1	6	21	93.1
33	40 S&W	Glock 23	0	1	2	8	20	90.3
34	9 mm Luger	Sig Sauer 226	1	1	3	5	22	87.1
35	40 S&W	Smith and Wesson SD40	0	1	8	10	12	71.0
36	45 Auto	H&K USP Compact	0	1	16	5	4	34.6
38	9 mm Luger	Glock 19	0	1	3	8	19	87.1
39	40 S&W	Smith and Wesson SW40VE	0	0	3	12	18	90.9
40	9 mm Luger	Sig Sauer P938	0	0	0	5	18	100
41	40 S&W	Springfield Armory XD40	0	1	4	6	20	83.9

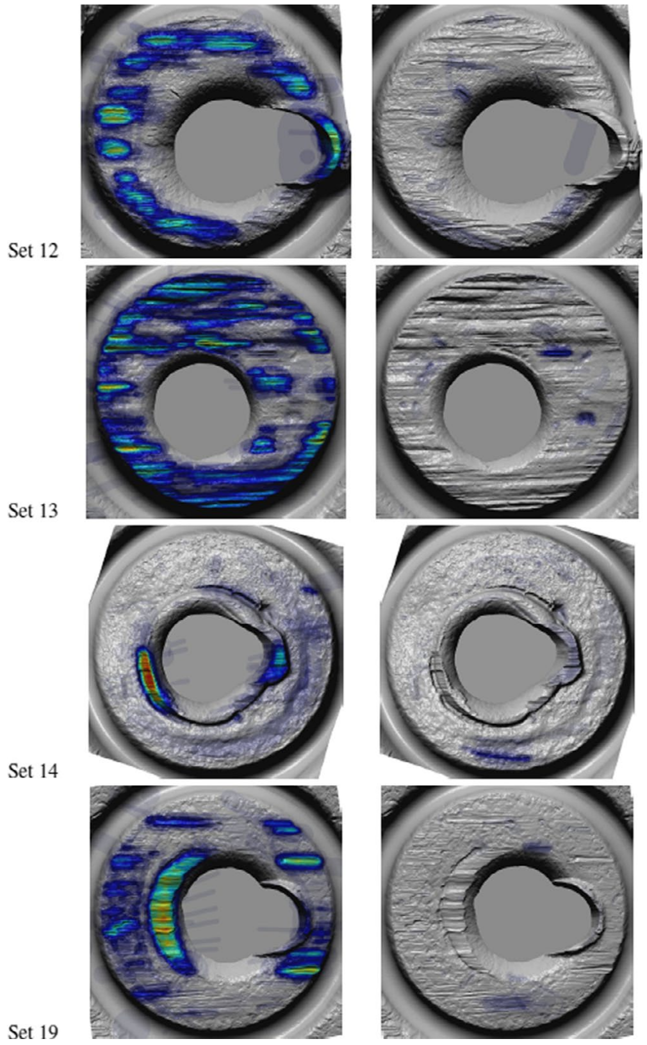


FIGURE 4 Summary annotation maps for four KM test sets for USCAN Qualified Examiners. Left is the similarity map, and right is the dissimilarity annotation map. Surface color indicates the portion of participants that marked the region as similar (left) or different (right). Color scale is shown next to Set 1 in Figure 3 [Color figure can be viewed at wileyonlinelibrary.com]

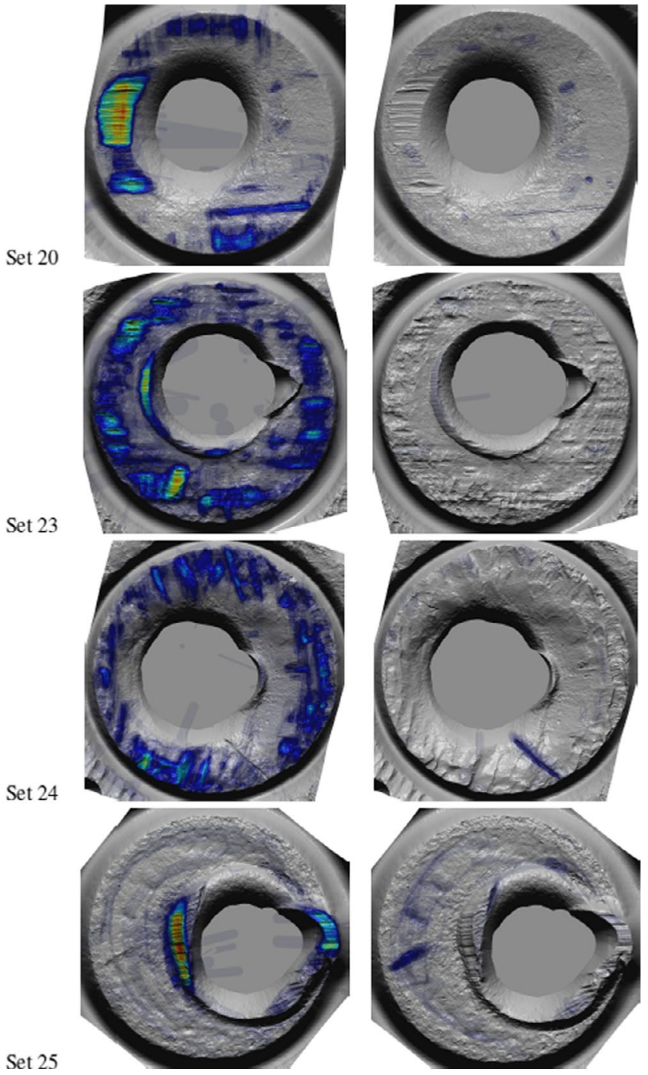


FIGURE 5 Summary annotation maps for four KM test sets for USCAN Qualified Examiners. Left is the similarity map, and right is the dissimilarity annotation map. Surface color indicates the portion of participants that marked the region as similar (left) or different (right). Color scale is shown next to Set 1 in Figure 3 [Color figure can be viewed at wileyonlinelibrary.com]

were minimally marked specimens. The lack of marks found useful for comparison are indicated by only blue colors present on the summary similar annotation map and almost no annotations present on the difference annotation map.

The source conclusions for individual KNM test sets are shown in Table 3 (bottom). Similarity and difference annotation maps are presented in Figures 8 through 13. The majority of KNM test sets are reported as Inconclusive-C or Elimination. Similar to the matches, the non-matches demonstrated consistency among annotated individual marks used in examination. For example, examiners show consensus in observing the differences of aperture shear marks as indicated in the difference annotation maps on test sets 5, 8, 10, 11, 18, 26, 31, 33, 35, 38, and 39 (Figures 8–13). Test set 15 (Figure 9) shows another example of independent and consistent agreement in the marks annotated as being different between knowns and unknown. In this case, the sample is a Star Bonifacio Firestar with linear marks across the entire breech face. The difference annotation map indicates examiners relied upon differences found near the 9 o'clock region when 100% of the USCAN examiners concluded either Elimination or Inconclusive-C.

One KNM test set (number 36) had the lowest Inconclusive-C or Elimination rate. Test set number 36 (Figure 12) had an Inconclusive-C/Elimination rate of 34.6%. The known specimens in this test set were from a Heckler and Koch (H&K) model USP. These samples were inconsistently and minimally marked. The lack of annotation differences (or similarity) indicates examiners independently found few toolmarks useful for comparison purposes and most concluded Inconclusive-B for this test set.

Overall, the USCAN performance statistics and corresponding consistency in annotation maps support the hypothesis that VCM can be a useful and accurate tool for reaching source conclusions.

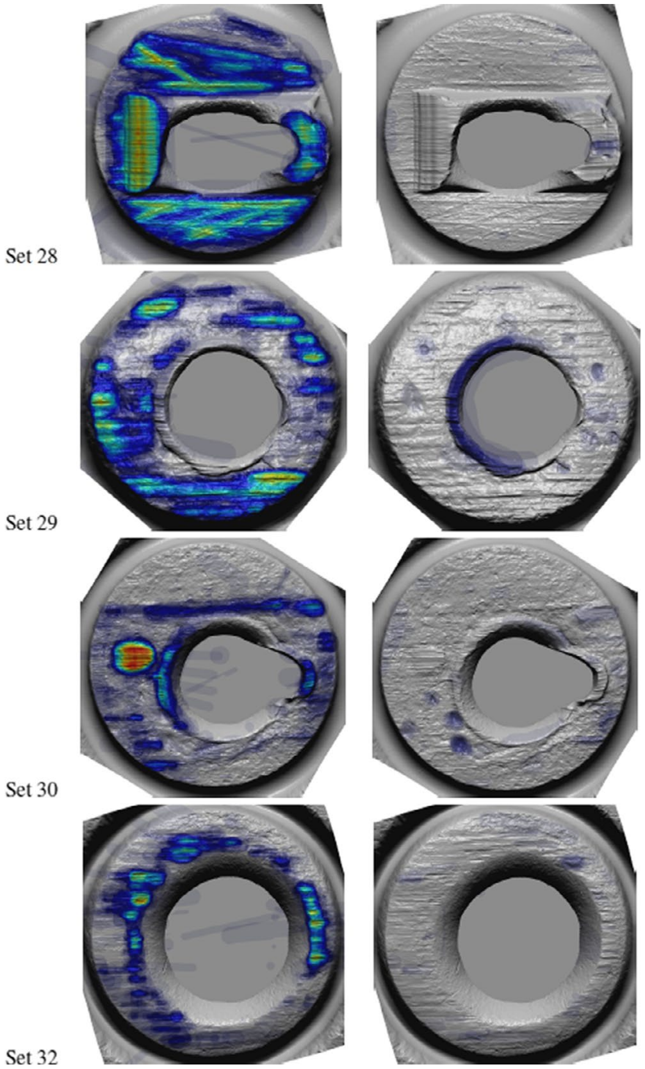


FIGURE 6 Summary annotation maps for four KM test sets for USCAN Qualified Examiners. Left is the similarity map, and right is the dissimilarity annotation map. Surface color indicates the portion of participants that marked the region as similar (left) or different (right). Color scale is shown next to Set 1 in Figure 3 [Color figure can be viewed at wileyonlinelibrary.com]

Participants were clearly able to recognize similar individual characteristics on the KM as indicated by the increased amount of similarity annotated on KM when compared to KNM. The converse was also true in that participants annotated more regions of dissimilarity on the KNM than the KM.

3.2 | USCAN errors

Of the three errors submitted by USCAN examiners, one participant (with less than three years of experience) contributed two errors and a second participant (with over ten years of experience) contributed one error. All three errors were false positives. The individual annotation maps were examined for insight into these errors. Unfortunately, the first examiner did not annotate the unknown sample for either of the two test sets (18 and 27) that resulted in their false identifications. Given the lack of annotations on the unknowns for both of these test sets, standard laboratory quality control procedures, such as verification and technical review, could catch the lack of proper documentation and prevent their reporting.

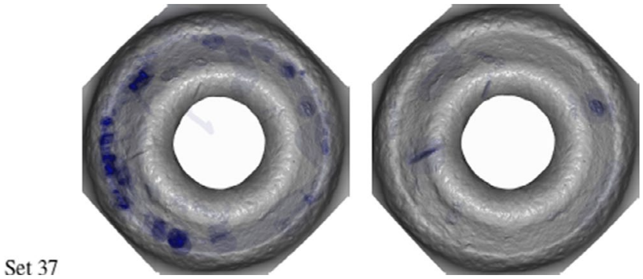


FIGURE 7 Summary annotation maps for the final KM test set for USCAN Qualified Examiners. Left is the similarity map, and right is the dissimilarity annotation map. Surface color indicates the portion of participants that marked the region as similar (left) or different (right). Color scale is shown next to Set 1 in Figure 3 [Color figure can be viewed at wileyonlinelibrary.com]

The examiner with ten years of experience reported one false positive. All other test sets submitted by this individual were correctly identified or eliminated, and they did not report any inconclusives. This examiner diligently annotated similarities and differences between the known and unknown for all test sets except the one with an error. On the error test set, the individual only annotated similarity between the two knowns. It is possible that the participant mistakenly compared the two knowns when they thought they were comparing a known to an unknown. This type of specimen “mix-up” is beyond the control of the designer of these types of error rate studies, regardless if VCM or LCM is used as the comparison method. Future versions of VCM software can minimize the risk of this phenomenon by restricting participants to record conclusions only when a known and an unknown are displayed. Similar to the other errors, this false positive could have been caught by standard laboratory quality assurance and quality control procedures such as verification and technical review.

The distribution of the errors is consistent with previous error rate studies where small numbers of individuals are responsible for multiple errors [18,19]. Because the same error was not repeated within any test set, the results suggest that a VCM quality control verification process may catch the source attribution errors reported. A laboratory with quality control procedures that require annotations to support reported conclusions could have caught all three errors through internal quality checks, such as a verification and technical review, given the lack of proper documentation (i.e., a lack of annotations of the comparison) by the participants for these test sets. Further study comparing the effectiveness of review with and without annotations would be of interest. However, given the low numbers of errors detected [3] in our study, it may be a challenge to obtain statistically meaningful results.

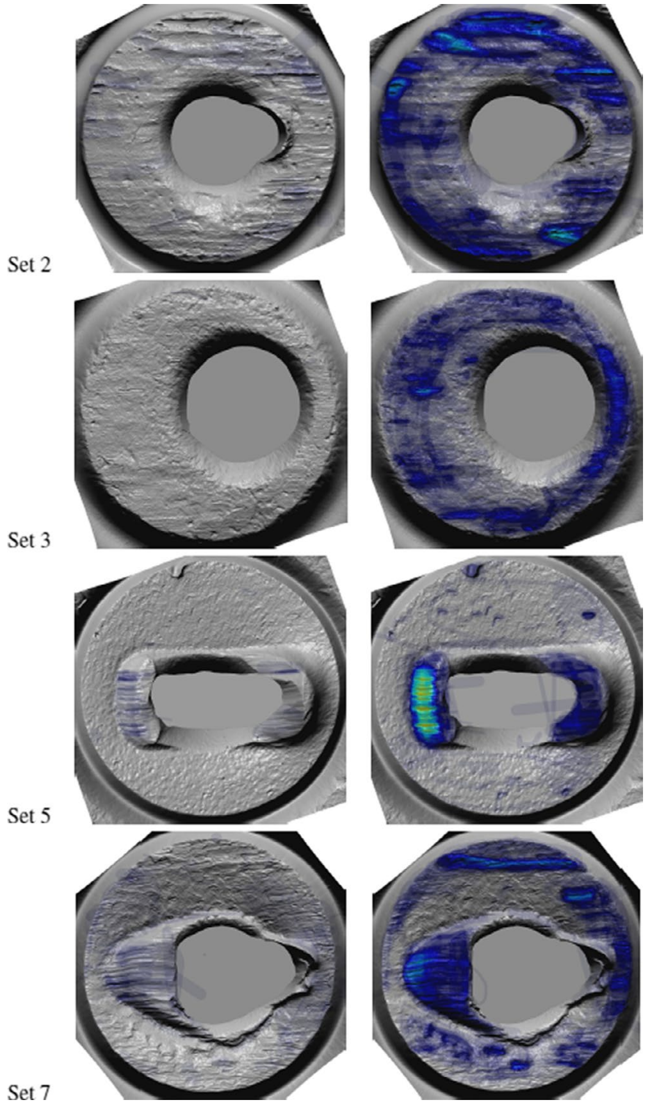


FIGURE 8 Summary annotation maps for four KNM test sets for USCAN Qualified Examiners. Left is the similarity map, and right is the dissimilarity annotation map. Surface color indicates the portion of participants that marked the region as similar (left) or different (right). Color scale is shown next to Set 1 in Figure 3 [Color figure can be viewed at wileyonlinelibrary.com]

3.3 | Error insight from individual surface annotations

Another use of the annotations is demonstrated by a false negative reported by a non-qualified examiner (e.g., trainee). On test set 4, the trainee annotated the absence and presence of an intermittent aperture shear mark which was absent on the knowns and present on the unknown. The individual annotation map indicates that the aperture shear was the basis of their false elimination conclusion (Figure 14). USCAN qualified examiners, demonstrated through their performance, their knowledge that the absence or presence of an intermittent aperture shear should not form the basis of an elimination. This highlights a way in which VCM technology can serve as a useful teaching tool for trainers and trainees. Not only can VCM ensure that a specific toolmark phenomenon appears in each participant's test set but the individual annotation maps allow instructors to understand the basis for any errors that might occur.

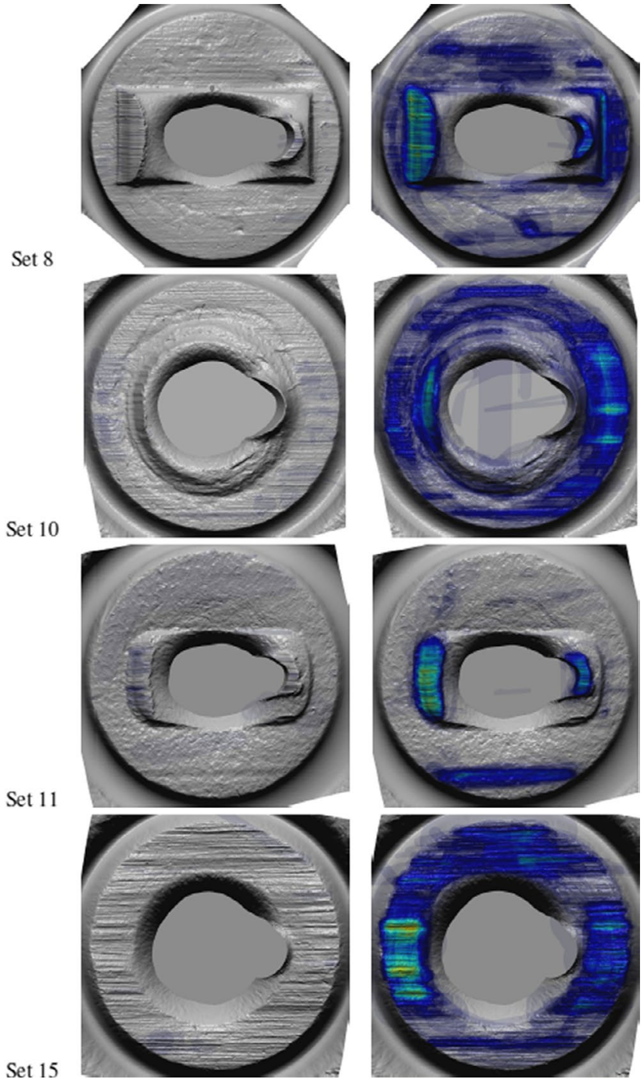


FIGURE 9 Summary annotation maps for four KNM test sets for USCAN Qualified Examiners. Left is the similarity map, and right is the dissimilarity annotation map. Surface color indicates the portion of participants that marked the region as similar (left) or different (right). Color scale is shown next to Set 1 in Figure 3 [Color figure can be viewed at wileyonlinelibrary.com]

3.4 | Additional statistics

Additional statistics were computed for the USCAN, the non-qualified (e.g., trainee), and the international groups. The USCAN overall positive predictive value defined as the number of ID calls which are actually KM is 453/456 = 99.3% (98.1%–99.9%). The USCAN overall negative predictive value defined as the number of Elimination calls which are actually KNM is 436/436 = 100.0% (99.2%–100.0%). The USCAN sensitivity defined as the number of KM called as Identifications is 453/491 = 92.2% (89.5%–94.5%).

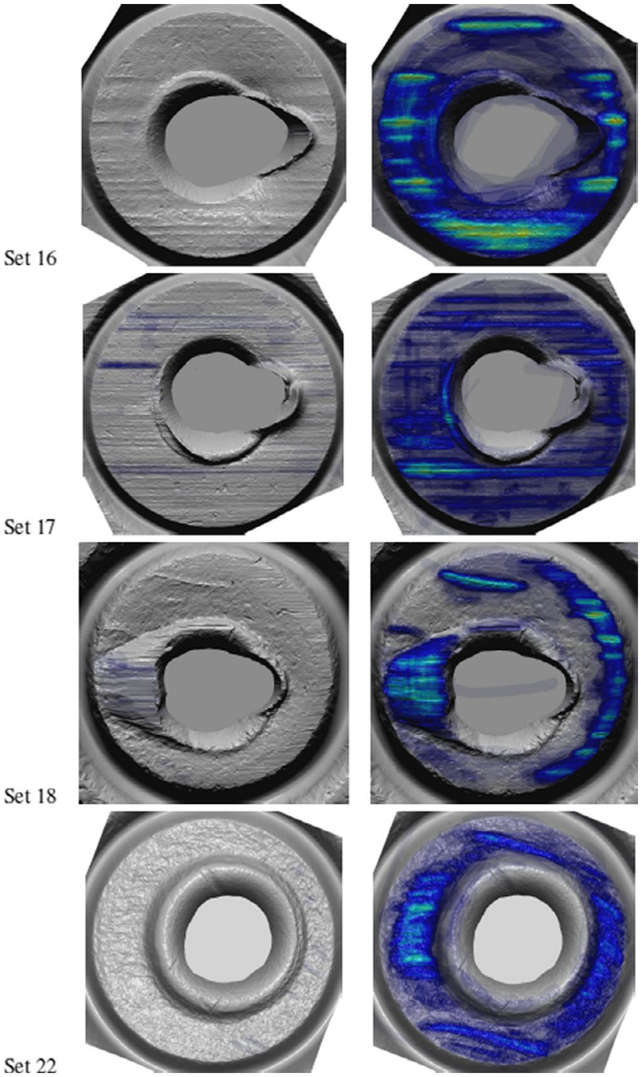


FIGURE 10 Summary annotation maps for four KNM test sets for USCAN Qualified Examiners. Left is the similarity map, and right is the dissimilarity annotation map. Surface color indicates the portion of participants that marked the region as similar (left) or different (right). Color scale is shown next to Set 1 in Figure 3 [Color figure can be viewed at wileyonlinelibrary.com]

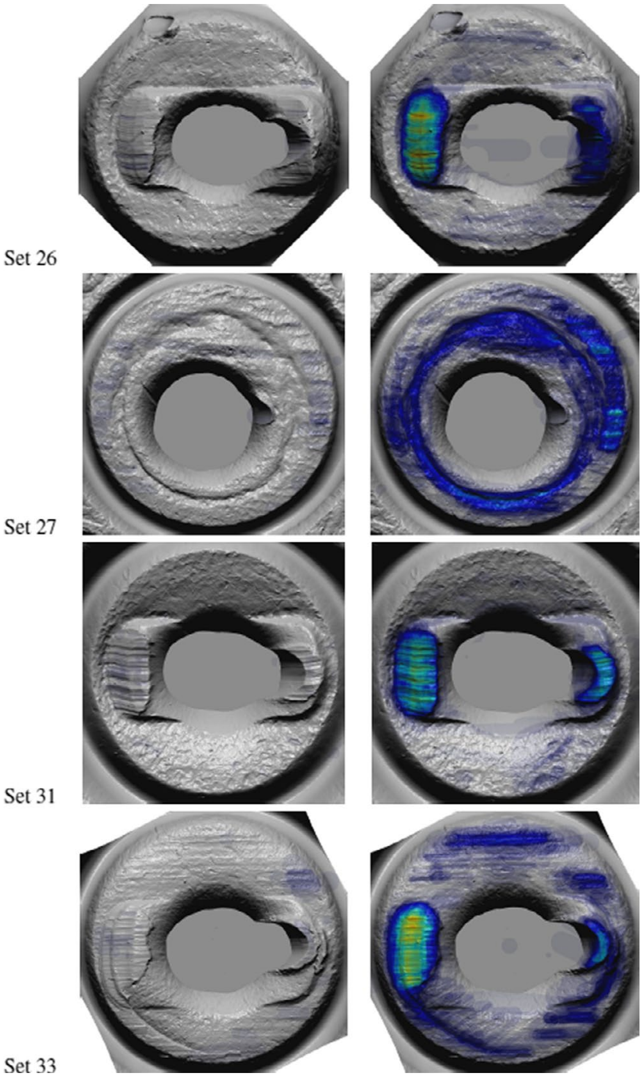


FIGURE 11 Summary annotation maps for four KNM test sets for USCAN Qualified Examiners. Left is the similarity map, and right is the dissimilarity annotation map. Surface color indicates the portion of participants that marked the region as similar (left) or different (right). Color scale is shown next to Set 1 in Figure 3 [Color figure can be viewed at wileyonlinelibrary.com]

Some laboratories are not able to eliminate on individual marks or without being able to inspect the firearm for alteration, or other contextual information (such as elapsed time between evidence and firearm collection). Therefore, it is useful to consider defining specificity using two options for the negative call, either “Elimination only” or “Elimination or Inconclusive-C”. Specificity is defined as the number of KNM called with the negative call. The USCAN specificity defined with a negative call as either Inconclusive-C or Elimination is 607/693 = 87.6% (84.9%–90.0%) and when defined using only Elimination is 436/693 = 62.9% (59.2%–66.5%).

The ten non-qualified participants (entire world) had a higher proportion of false-positive and false-negative errors when compared to USCAN examiners (Table 2). The non-qualified participants had one false positive (of 89 known non-matches) and one false negative (of 68 known matches). The results for the twenty-one qualified examiners in the international group are shown in Table 2. The international qualified examiners had four false positives (of 180 known non-matches) and three false negatives (of 146 known matches). Unlike the USCAN group where there was a relatively large number of participants, the non-qualified and International qualified groups have a small number of test takers. Therefore, it is important not to over interpret statistics on these small groups representing several countries.

4 | DISCUSSION

The goal of this study was both to validate VCM and to establish VCM error rates for the TopMatch-3D system. One way to validate the use

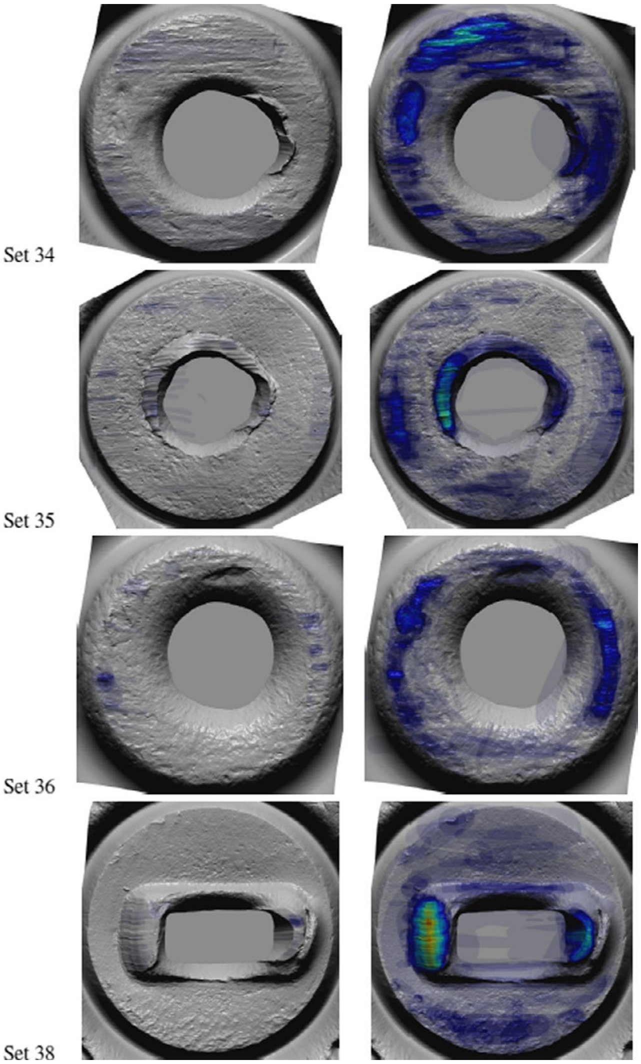


FIGURE 12 Summary annotation maps for four KNM test sets for USCAN Qualified Examiners. Left is the similarity map, and right is the dissimilarity annotation map. Surface color indicates the portion of participants that marked the region as similar (left) or different (right). Color scale is shown next to Set 1 in Figure 3 [Color figure can be viewed at wileyonlinelibrary.com]

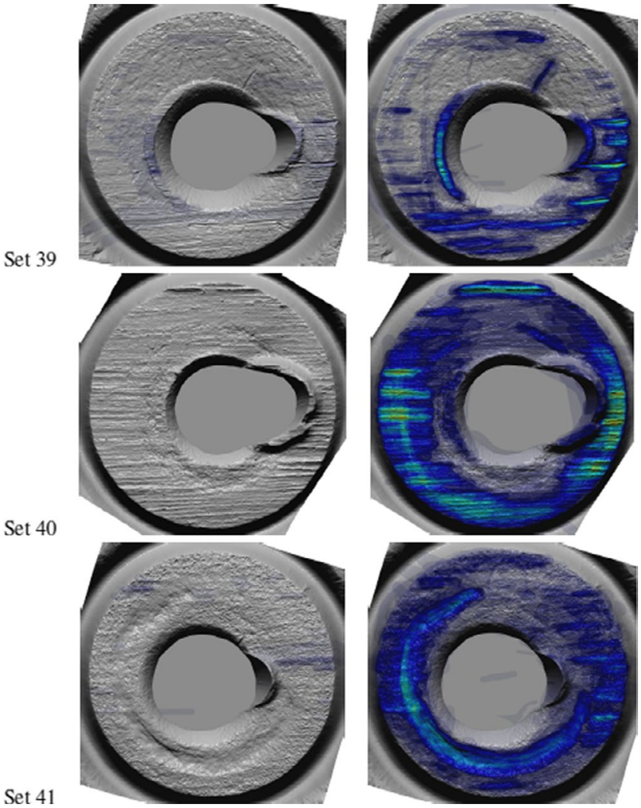


FIGURE 13 Summary annotation maps for three KNM test sets for USCAN Qualified Examiners. Left is the similarity map, and right is the dissimilarity annotation map. Surface color indicates the portion of participants that marked the region as similar (left) or different (right). Color scale is shown next to Set 1 in Figure 3 [Color figure can be viewed at wileyonlinelibrary.com]

of VCM within a crime laboratory is to establish that VCM achieves error rates similar to or better than currently utilized microscopic examination methods. Studies should involve a large number of participants such that confidence intervals around calculated error rates be small. The VCMER study was completed by over one hundred participants and had a core group of 76 qualified examiners from the USA and Canada. Most of the participants had minimal or no prior VCM experience. It is also important that validation studies include a broad range of samples representative of casework. Prior error rate studies have typically used only one or a few makes and models of firearms [10,18–26]. The VCMER study involved 40 test sets covering a range of common firearm makes, models, and calibers. These sets include both well and minimally marked cartridge cases spanning a range of expected comparative complexity. To reflect a heterogeneous laboratory setting, participants completed the study using VCM software on their own Windows-based desktop or laptop computers with data downloaded from a common network server. All participants worked through a training booklet and proficiency test prior to beginning the study. Source conclusions were reported using the AFTE 5-point Range of Conclusions and documented with surface annotations indicating areas of similarity and dissimilarity.

Error rate studies inherently include a number of design choices such as the firearm makes, models, and calibers included, the ammunition used for test fires, the individual samples selected, the examiners that elected to participate, and the test design implemented. These design variables make it difficult to directly compare error rates from any two individual studies. Two previous studies using traditional light comparison microscopy (not VCM) reported false-positive error rates between 0.0% and 0.939% [18,20,21] and false-negative rates between 0.0% and 0.367%. The false-positive and negative rates in the presently reported VCMER study of 0.43% and 0.0% are well within the ranges reported in published LCM studies. As noted above, most VCMER study participants had little or no previous experience with VCM; therefore, one expects lower error rates as the profession gains additional experience with VCM. The error rates provide strong support that VCM is as reliable as traditional LCM.

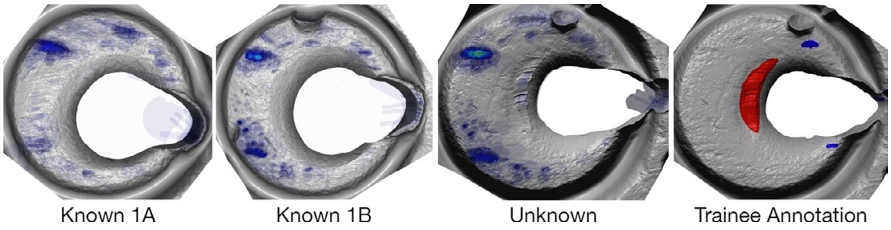


FIGURE 14 Summary annotation maps (left three images) and one individual annotation map from a non-qualified (trainee) participant (right image) for test set 4, a known match comparison. The presence of the aperture shear for this sample is inconsistent. The aperture shear is missing from the two known test fires but it appears in the unknown. On the far right are the individual annotations of a trainee who reported a false elimination. In the right image, red indicates a region annotated as different and blue indicates a region annotated as similar. The trainee annotated the aperture shear (in red) indicating they used this as the basis for their elimination. This illustrates an excellent teaching opportunity for both this trainee and others who have not seen this intermittent aperture shear phenomenon [Color figure can be viewed at wileyonlinelibrary.com]

The annotation maps provide insights into both individual and overall examiner decision-making processes. Summary annotation maps represent a consensus of approximately 30 independent examinations. Examiners had a high amount of agreement with regard to the areas useful for identification and elimination as well as those areas which should be avoided for definitive source attribution. It is worth reiterating that examiners worked independently and that the described annotation map patterns emerged when these independent submissions were combined. This consistency reinforces the fact that examiners typically agree on the toolmarks most important and most reliable for reaching source conclusions. Additionally, the use of individual annotation maps can provide insight into the source of errors. The individual annotation maps for the three errors reported above did not support the submitted source conclusion. If a laboratory's QA system required examiners to support conclusions with proper annotations, then the erroneous conclusions could have been caught by standard laboratory quality control procedures. Used in this manner, annotation maps may form an important part of the verification and technical review process. Finally, annotation maps coupled with virtual samples provide excellent opportunities for training in that trainees can be exposed to specific and important toolmark phenomena such as the intermittent aperture shear mark described in test set 4.

It is important to note that not all VCM systems are the same. Each has different strengths and weaknesses. For example, some 3D microscopes are unable to measure aperture shear which the presented summary annotation maps strongly suggest are critical for reaching correct conclusions on many cartridge cases. The Cadre system used in the VCMER study accurately measured this aperture shear. Similarly, some software does not allow surface annotations which support the meticulousness of the examination process. Therefore, the VCMER study validated the TopMatch-3D system from Cadre Forensics (Chicago, IL) and not those from other vendors.

Although the VCMER study was designed to establish VCM error rates and validate VCM technology, it also supports a more general conclusion. Among qualified examiners, the low error rates and the consistency observed in annotation maps lend support to the foundational principles of the firearm and toolmark examination discipline. The VCMER study establishes low VCM error rates and strongly supports the use of VCM in determining source conclusions. The adoption of VCM within the crime laboratory and the establishment of best practices for its use will be further strengthened by the publication of additional VCM studies.

ACKNOWLEDGEMENTS

The authors are grateful to Max Morris (Iowa State University) and Erich Smith (FBI, Quantico VA) for helpful discussions, assistance, and experiment planning. The authors thank all participants and participating laboratories for volunteering their time and test fires to this research.

CONFLICTS OF INTEREST

The TopMatch system and Virtual Microscopy Viewer software described in this manuscript are projects of Cadre Research Labs. Authors Chapnick, Weller, Meschke, Duez, and Lilien are supported in part by Cadre Research Labs.

REFERENCES

1. Hamby JE. The history of firearm and toolmark identification. AFTE J. 1999;31(3):226–84.
2. Vorburger TV, Yen JH, Bachrach B, Renegar TB, Filliben JJ, Ma L, et al. Surface topography analysis for a feasibility assessment of a national ballistics imaging database. A report prepared for the National Academies Committee to assess the feasibility, accuracy, and technical capability of a National Ballistics Database. NISTIR 7362. Gaithersburg, MD: National Institute of Standards and Technology; 2007. p. 1–171.
3. Vorburger TV, Song J, Petraco N. Topography measurements and applications in ballistics and tool mark identifications. Surf Topogr. 2016;4(1):013002. https://doi.org/10.1088/2051-672X/4/1/013002.
4. Petraco ND, Shenkin P, Speir J, Diaczuk P, Pizzola PA, Gambino C, et al. Addressing the National Academy of Sciences' challenge: a method for statistical pattern comparison of striated tool marks. J Forensic Sci. 2012;57(4):900–11. https://doi.org/10.1111/j.1556-4029.2012.02115.x.
5. Hadler JR, Morris MD. An improved version of a tool mark comparison algorithm. J Forensic Sci. 2018;63(3):849–55. https://doi.org/10.1111/1556-4029.13640.
6. Song J. Proposed, "Congruent Matching Cells (CMC)" method for ballistic identification and error rate estimation. AFTE J. 2015;47(3):177–85.

7. Yang M, Mou L, Fu YM, Wang Y, Wang JF. Quantitative statistics and identification of tool-marks. J Forensic Sci. 2019;64(5):1324–34. https://doi.org/10.1111/1556-4029.14040.
8. Chumbley LS, Morris MD, Kreiser MJ, Fisher C, Craft J, Genalo L, et al. Validation of tool mark comparisons obtained using a quantitative, comparative, statistical algorithm. J Forensic Sci. 2010;55(4):953–61. https://doi.org/10.1111/j.1556-4029.2010.01424.x.
9. Ekstrand L, Zhang S, Grieve T, Chumbley LS, Kreiser MJ. Virtual tool mark generation for efficient striation analysis. J Forensic Sci. 2014;59(4):950–9. https://doi.org/10.1111/1556-4029.12435.
10. Duez P, Weller T, Brubaker M, Hockensmith RE 2nd, Lilien R. Development and validation of a virtual examination tool for firearm forensics. J Forensic Sci. 2018;63(4):1069–84. https://doi.org/10.1111/1556-4029.13668.
11. Weller T, Brubaker M, Duez P, Lilien R. Introduction and initial evaluation of a novel three dimensional imaging and analysis system for firearm forensics. AFTE J. 2015;47(4):198–208.
12. Senin N, Groppetti R, Garofano L, Fratini P, Pierni M. Three-dimensional surface topography acquisition and analysis for firearm identification. J Forensic Sci. 2006;51(2):282–95. https://doi.org/10.1111/j.1556-4029.2006.00048.x.
13. Johnson MK, Adelson EH. Retrographic sensing for the measurement of surface texture and shape. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 June 20–25; Miami, FL. Piscataway, NJ: IEEE; 2009. p. 1070–7. https://doi.org/10.1109/CVPR.2009.5206534.
14. Johnson MK, Cole F, Raj A, Adelson EH. Microgeometry capture using an elastomeric sensor. ACM Trans Graph. 2011;30(4):1–8. https://doi.org/10.1145/1964921.1964941.
15. Montgomery DC. Design and analysis of experiments. 8th ed. Hoboken, NJ: John Wiley & Sons; 2013. p. 168–76.
16. AFTE Standardization and Training Committee. Association of Firearm and Tool Mark Examiners Glossary. 6th ed. Chicago, IL: AFTE Standardization and Training Committee; 2013. https://afte.org/resources/afte-glossary. Accessed 3 Sept 2020.
17. Nichols R. Subclass characteristics: from origin to evaluation. AFTE J. 2018;50(20):68–88.
18. Baldwin DP, Bajic SJ, Morris M, Zamzow D.A study of false-positive and false-negative error rates in cartridge case comparisons. Ames Laboratory, USDOE. 2014. Technical Report #IS-5207. https://afte.org/uploads/documents/swggun-false-postive-false-negative-usdoe.pdf. Accessed 3 Sept 2020.
19. Mayland B, Tucker C. Validation of obturation marks in consecutively reamed chambers. AFTE J. 2012;44(2):167–9.
20. Keisler MA, Hartman S, Kilmon A, Oberg M, Templeton M. Isolated pairs research study. AFTE J. 2018;50(1):56–8.
21. Keisler MA. ,Letter to the Editor: “Isolated pairs research study”. by Keisler et al. AFTE J 2018;50(1):56–8. AFTE J. 2018;50(3):131.
22. Fadul TG Jr, Hernandez GA, Stoiloff S, Gulat S. An empirical study to improve the scientific foundation of forensic firearm and tool mark identification utilizing 10 consecutively manufactured slides. AFTE J. 2013;45(4):376–89.
23. Fadul TG, Hernandez GA, Wilson E, Stoiloff S, Gulati S.An empirical study to improve the scientific foundation of forensic firearm and tool mark identification utilizing consecutively manufactured Glock EBIS barrels with the same EBIS pattern. National Institute of Justice Grant #2010-DN-BX-K269; 2013. p. 1–51. https://www.ncjrs.gov/pdffiles1/nij/grants/244232.pdf. Accessed 3 Sept 2020.
24. 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. https://doi.org/10.1111/1556-4029.13916.
25. Smith TP, Andrew Smith G, Snipes JB. A validation study of bullet and cartridge case comparisons using samples representative of actual casework. J Forensic Sci. 2016;61(4):939–46. https://doi.org/10.1111/1556-4029.13093.
26. Stroman A. Empirically determined frequency of error in cartridge case examinations using a declared double-blind format. AFTE J. 2014;46(2):157–75.

How to cite this article: Chapnick C, Weller TJ, Duez P, Meschke E, Marshall J, Lilien R. Results of the 3D Virtual Comparison Microscopy Error Rate (VCMER) Study for firearm forensics. *J Forensic Sci*. 2021;66:557–570. https://doi.org/10.1111/1556-4029.14602