Final Research Performance Progress Report - Cover Page

Federal Agency and Organization Element: Department of Justice, Office of Justice Programs

Federal Grant or Other Identifying Number: 2017-IJ-CX-0024

Project Title: Firearm Forensics Black-Box Studies for Examiners and Algorithms using Measured 3D Surface Topographies

PD/PI Name, Title, Contact Info: Ryan Lilien, Chief Scientific Officer, Cadre Research Labs; 420 W Huron St, Suite 204; Chicago, IL 60654

Name of Submitting Official: Ryan Lilien

Submission Date: May 21, 2019

[redacted]

Recipient Organization: Cadre Research Labs, LLC (small business)

Recipient Identifying Number (if any): N/A

Project/Grant Period: Start: 1/1/2018, End: 12/31/2018

Reporting Period End Date: 12/31/2018

Report Term or Frequency: Twice a year

Signature of Submitting Official:

1 Project Purpose and Background

Over the past several years, advances in 3D surface metrology have made their way into the field of firearm and toolmark analysis. Accurate surface imaging coupled with high-resolution visualization tools and advanced algorithms are beginning to allow examiners to view, annotate, and share data between labs, to conduct blind verification, and to form a statistical basis for identification. In 2016, the President's Council of Advisors on Science and Technology (PCAST) issued a report critical of toolmark analysis and called for additional research. The three aims completed in this proposal address critical aspects of the recent PCAST report while advancing the field of 3D scanning and analysis for firearm forensics. First, we improved our Virtual Microscopy Viewer software to better facilitate its use in black-box studies. Second, we selected and scanned a large set of cartridge cases. Finally, we conducted a large black-box virtual microscopy study. The completed work includes critical steps towards further validating the field of toolmark examination and the use of 3D scanning technology in the forensic lab.

The comparison of cartridge cases is based on the observation that microscopic firearm imperfections can be transferred to ammunition during firing. The ability to certify two cartridge cases as similar is therefore a function of both the ability to capture and visualize a high-resolution measurement of each specimen and the ability to identify and match relevant structural features between the two. Courtroom challenges and recent reports have called for additional research into underlying error rates and performance measures for these comparative methods.

Firearm and toolmark examiners complete years of training to gain competency and proficiency in the examination and assessment of toolmarks. For over 100 years, these toolmarks have been manually examined using light-microscopy. Examiners document conclusions with written reports that contain image snapshots annotated to indicate regions of similarity. In the early 1990s, the examiner's ability to compare cartridge cases was augmented with the introduction of commercial database systems. The first systems combined traditional 2D light microscopy with a digital camera and software for image comparison and database search.

1.1 Transition to 3D Measurements

Several shortcomings of traditional (2D) toolmark examination can make comparison difficult [2]. For example, lighting effects (*i.e.,* shadows) can adversely affect 2D image interpretation. In addition, traditional comparison light-microscopy suffers from a physical access requirement. That is, examination requires physical access to the specimens. This may necessitate potentially burdensome chain-of-custody documentation and introduces the opportunity for evidence to be damaged or lost. When used as part of proficiency testing or error rate determination, the need to exchange and examine physical cartridge cases introduces test set to test set variability where different study participants each receive different sets of test fires (from the same set of firearms, but different non-identical test fires).

To address these issues, new technologies, capable of measuring 3D surfaces, are now being evaluated [3, 16, 19]. Some of these technologies, including our GelSight-based scanner, measure accurate 3D surface topographies in standard units resulting in a detailed heightmap of the cartridge case surface. These information-rich 3D surfaces typically offer examiners significantly more detail than traditional 2D images. In addition, these surfaces can be exchanged between systems using a common file format. Comparison algorithms are being developed to analyze these 3D surface topographies [4, 9, 13, 14, 15, 20, 21] and may soon provide statistical interpretations to their match scores (*e.g.*, a false match rate).

The topographic data acquired from 3D scanners can be used in the emerging application of Virtual Comparison Microscopy (VCM). Initially introduced by Senin *et al.* [10] in 2006, VCM describes the visual examination of a 3D microscopic representation of an object. In VCM, the examiner views and manipulates the object's measured 3D representation using a computer without physical access to the specimen. The lack of a physical access requirement allows several advantages across the areas of: Access & Archiving Evidence, Training, Proficiency/Error-Rate Studies, Verifications, and Algorithmic Comparison. For these reasons, the past few years have seen significant interest and movement towards 3D imaging. An important part of this shift is the validation of 3D microscopy and the establishment of error rates.

TopMatch (GelSight) Scan Acquisition. Over the past few years we've developed technology capable of measuring the 3D surface topographies of cartridge cases at micron-scale resolution (Fig. 1). Our approach utilizes advanced three-dimensional imaging algorithms (*e.g.*, shape from shading and photometric stereo) and the GelSight sensor [7, 8]. Our sensor is a block of optically clear elastomer with a thin layer of elastic paint on one side (Fig. 1). When an object is pressed into the elastomer, the layer of paint conforms to the shape of the surface. The paint removes the influence of the optical properties of

the surface on shape measurement. In contrast to confocal and focus-variation microscopy, this important feature of our system removes the influence of surface reflectivity on the measured topography. A particular strength of the technology is its ability to capture surfaces with significant slope. This provides an advantage over confocal microscopy whose signal can become unreliable for sloped surfaces [18].

1.2 Error Rates

At the heart of any error rate study is the recruitment of a large number of firearms examiners and the assessment of their analytic accuracy on a large set of test cartridge cases designed as closely as possible to emulate actual casework. The 2016 PCAST report was highly critical of firearms and toolmark examination claiming that error-rates have not been well established [6]. PCAST was critical of studies where comparisons were not fully independent. PCAST's claim is that non-independent tests might allow examiners to deconstruct the test design (*e.g.*, 'closed-set design'). PCAST looked most favorably on studies like that of Baldwin et al. [1] which is based off the latent-print study design of Ulery et al. [17]. These studies were structured as a large number of independent sample sets with only 2-4 samples per set. We note that most prior studies have been 'black-box' studies in that they are concerned with evaluating examiner accuracy (*e.g.*, their decisions) and not with the details of the decision making process. In contrast, 'white-box' studies are also interested in studying the decision making process.

We recently completed the first VCM study for cartridge case examination and summarized our results in our 2018 JFS paper [5]. This first study, evaluated the feasibility of using virtual microscopy for cartridge case examination. The study involved 56 participants (46 trained examiners and 10 trainees) from fifteen US labs. The study structure included two tests, each with three known test fires and four unknown test fires. Participants were asked if any of the unknowns identify to the known and were asked to mark the scan surfaces to indicate the individual marks used when reaching their conclusions. There were no errors among the 368 results submitted by qualified examiners. The study successfully demonstrated proof-of-concept that VCM could be used by examiners as a substitute for traditional comparison microscopy. It showed that similarity in both striated and impressed marks could be identified. We demonstrated that the visualization tools were generally easy to learn and that the annotation mode provides valuable insights into the decision process.

This proposal addresses continued validation of 3D Virtual Comparison Microscopy and the establishment of error rates for this new technology. The research work was completed by Cadre Research Labs, a leading research organization within the discipline of 3D imaging for firearm forensics. Cadre

worked closely with Todd Weller, a firearms examiner in private practice who was formerly affiliated with the Oakland Police Department and John Marshall, a firearms examiner from the Royal Canadian Mounted Police. The project team continued to collaborate with colleagues at NIST as well as federal, state, and local crime labs. These collaborators continued to be excellent partners and provide both scans and constructive feedback.

2 Project Design

The one year project included three aims which continued the R&D of our novel technology to advance 3D Virtual Comparison Microscopy. The core of the completed project is the large VCM validation study (Aim 3). We named the study the Virtual Comparison Microscopy Error Rate Study and will refer to it by the acronym VCMERS below. Completion of this study required assembly and design of the test datasets, participant recruitment, front-end VCM software development, back-end development of the server architecture to support data distribution and collection, support of participants, assembly of results, and summary of performance. In completion of Aim 1, we developed two software programs. The first is a significant update to our VCM software which was used in the validation study (Aim 3). The second piece of software is an algorithm testing tool designed to facilitate evaluation of different comparison algorithms. Although this proposal did not include the evaluation of different comparison algorithms (which would require the cooperation of different algorithm developers) we plan to conduct an algorithm comparison like this in the future. In Aim 2, we assembled the test sets used in the validation study (Aim 3). All 3 proposed aims were successfully completed during the project period.

3 Materials and Methods

In this section we describe the general approach for each aim. In the Results and Analysis section we describe the experimental performance and results of the project work. Methods have been abbreviated to conform to page limits.

3.1 VCM Testing Platform

We completed a rewrite of our core VCM software (Figure 2) to accomplish four main goals. First, we wanted to increase the range of supported computer hardware, second we wanted to create a more guided testing experience for participants, third we wanted to add network support to allow efficient distribution

of data and collection of results, finally we wanted to create an easy to use software installer. In our first (2016) VCM study, we loaned laptops to participants. Given the size and international geographic distribution of participants in VCMERS we were not able to purchase and distribute loaner laptops. Therefore, we needed to improve several aspects of the VCM software to allow it to run on participants computers. We also needed to develop a mechanism to get each participant the scan files corresponding to the test sets they were asked to consider. Finally, we needed a way to collect the test results from the participants when they had completed their analysis.

First, we rewrote the graphics (rendering) code to improve cross-platform compatibility. That is, we had reports that some types of graphics cards and some older computers were not able to run our older VCM software. The rewrite optimized the memory requirements and the use of the graphics libraries. We've found that with our updates the VCM software is able to run on virtually all tested Windows machines released after 2014. Despite this update there were still a few individuals who were unable to participate in the study because the newest computer they had access to did not meet these minimal requirements.

Second, a VCM testing mode was added to the software which guides the participant through the validation study. That is, an examiner is first presented with a set of training scans illustrating different firearm toolmark types. The examiner is guided through use of different software features such as adjusting the virtual light, the zoom, and the rotation. The examiner is then presented a mini proficiency-style test with three known test fires and four unknown test fires. The examiner is required to successfully complete an identification worksheet for these scans. Only after successfully demonstrating proficiency with the software and visualization is the examiner allowed to proceed to the study test sets.

Third, we developed a network server (our Nexus server), to host the scan data and results. As described below, the test was structured using a Balanced Incomplete Block Design (BIBD) structure. In a BIBD test, each individual receives a different set of tests to evaluate. Each participant was randomly assigned a participant ID and webcode (*i.e.*, their credentials). Each participant was emailed their ID and webcode. When the VCM software starts, it asks the user to login with their credentials. The first time a user enters this information the software requests permission to access the network to download the test sets assigned to that ID[1]. Each participant only requires access to the scans assigned to that individual. The software therefore only downloads the relevant scans. This 'as needed' approach minimizes the

[1]A backup option was provided for participants whose computer was not on the internet. These individuals could use a different computer to download their test sets from our website. The user could then copy the files to their testing computer (*e.g.*, via USB drive).

data transfer required for each computer. Although test sets varied slightly in size, the average size of each of the required sixteen test sets was approximately 100MB. Therefore each participant considered 1600MB (or 1.6GB) of data. Test sets were randomized, so while each participant is presented with test sets numbered one through sixteen the numbering is not consistent between participants. That is, test set one for participant A may be different than set one for participant B. The software keeps track of each test set and each participant. Our network server was also designed to accept the participant test results. After the user completes all sixteen test sets and are happy with their results, they select a menu option to submit their results. In addition to submitting the results (*i.e.*, conclusions and annotations), participants were given a short questionnaire and were asked if the software could record and upload their system's hardware specs. This provided us information on the types of machines on which the study was completed. There were no surprises in this information. As expected, individuals made use of both desktop and laptop computers with a range of screen resolutions.

Finally, we created a software installer to facilitate installation (and uninstallation) of the VCM software. The installer is downloaded from our website after the user enters their credentials. Double clicking the installer starts the automated installation process. Participants generally noted success with the software install process.

3.2 Algorithm Testing Tool

Separate from development of resources for the examiner VCM validation study we also created a software tool to facilitate evaluation of comparison algorithms. The software runs on Windows computers and allows the user to specify a set of scan files and an algorithm to use in comparison. The software uses the algorithm to compute a similarity score for all pairs of scan files in the test set. Any scan file measured and recorded in standard units and saved in the common X3P file format can be analyzed. Our tool can utilize two types of comparison algorithms, internal algorithms and external algorithms. Internal algorithms are those implemented in our core TopMatch software. The program code for this comparison is in our code base and can therefore be called directly from the algorithm testing tool. External algorithms are those created outside the TopMatch software (*e.g.*, by other vendors or research labs). This general functionality allows testing of algorithms from different vendors even if the vendor (understandably) does not release their program source code. That is, our algorithm testing tool can 'wrap' a compiled program without requiring access to the program's source code. The main requirement is that the program being wrapped have a command line interface. That is, the software must be invokable on

the windows command line using a templated start string. The software must then output the resulting score as the only output generated by the software. For example, consider a program "compareAlgoA" created by a researcher which accepts the names of two scan files to compare as well as a search parameter S which specifies some tunable detail of the comparison. In this example, the parameter S might specify a block-size into which each scan is divided. To invoke a comparison with this software the researcher might issue the following command on the Windows command prompt, "compareAlgoA -S 20 scanone.x3p scantwo.x3p". In this example, the S parameter is set to 20, and the scan with filename scanone.x3p is compared to the scan with filename scantwo.x3p. Upon pressing enter, the software would run and generate one number corresponding to the computed similarity of the two scans. Our algorithm testing tool is able to support any algorithm that functions like this. In the parameters section of the algorithm testing tool (Figure 3), the user would specify "Other" algorithm type and would then type "compareAlgoA -S 20 %1 %2" which tells the testing tool the general format that should be used when invoking the algorithm compareAlgoA. The values %1 and %2 are placeholders that will be filled in by the algorithm testing tool with the filenames of pairs of scans to compare. In total, the algorithm testing tool first assembles the complete list of scans to compare, then it computes the match score for each pair of scans by running the comparison algorithm according to the custom command line string specified by the user. In the above example, the algorithm compareAlgoA output a single numeric score for the comparison of the two scans; however, our tool can accept multiple outputs assuming they are separated by commas. For example, an algorithm might output two numeric scores and then a conclusion (*i.e.*, elimination, inconclusive, or identification). Our tool builds a spreadsheet (readable by Excel) with the names of the two compared scans and all generated outputs.

The algorithm testing tool has another important feature, which is that it can run multiple comparisons at the same time. Most modern computers are capable of running 2, 4, 8, or more processes at the same time. The user can specify the maximum number of simultaneous comparisons to perform by adjusting the Thread Count parameter. Increasing the thread count will reduce the overall time required to compare all pairs in the test set.

3.3 Test Set Design (Aim 2)

The first step in dataset creation involved acquiring and scanning test fires. We solicited test fire contributions from US crime labs via conference and seminar presentation and on the AFTE forums. Contributing labs included the San Francisco Police Department Crime Lab, the Palm Beach County Sheriff's

Office, the Corpus Christi Police Department, and the Virginia Department of Forensic Science. A few labs which contributed test fires asked to remain anonymous based on lab policy. Cartridge cases were scanned by our paid intern (a student in the masters program in forensic science at the University of Illinois at Chicago). Our intern scanned over 2000 cartridge cases from more than 500 different firearms. In collaboration with two firearms examiners, Todd Weller (Weller Forensics, Burlingame, California) and John Marshall (RCMP, Ottawa), we considered test fires from over three hundred different firearms. Test fires were attributed class characteristics and a level of 'complexity' (low, medium, or high). We use the term complexity to refer to the quality and quantity of individual marks present on the scan surface. Surfaces with low complexity are less complex to identify/eliminate whereas surfaces with high complexity are more complex to identify/eliminate. The expectation is that examiners should have no problem reaching the correct conclusions for low complexity cartridge cases. We expect fewer inconclusive results for low complexity scans.

Of the 40 selected test sets, 30% were deemed low complexity, 38% were medium complexity, and 32% were high complexity[2]. Of the knowns, 30% have filed features, 25% have broached features, 23% had granular features, 45% had partial or complete aperture shears. Three calibers were included, 55% were 9mm, 33% were .40 S&W, 12% were .45 ACP. Test fires from thirty different firearm models from fifteen different manufacturers were included. Of note, all test fires within a single test set had the same class characteristics. Therefore it was not possible for participants to eliminate simply based on class.

Given that participants were volunteers, we did not think it appropriate to ask individuals to evaluate a large number of test sets. However, to cover a range of scan complexity and class we knew the study would be stronger if it contained a large number of test sets. We therefore needed a test design that would allow us to evaluate performance of a large number of test sets while not requiring each participant to evaluate every test set. After discussion with Max Morris, a statistics professor at Iowa State University, we decided to structure the VCMER study using a Balanced Incomplete Block Design (BIBD). 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 our study, each participant examines a block (or group) of test sets (test fire triples). The term *incomplete* means that not all test sets are evaluated by each participant (*e.g.*, the blocks are not complete). This incompleteness satisfies our first criteria, as we did not want all participants to analyze all test sets. The term *balanced* means that every pair of test sets are seen by the

[2]Note that we initially selected 41 test sets; however, there was a typographical error in the description of one dataset where a .45 ACP caliber cartridge case was described as being a .40 S&W caliber. Although no participants made mistakes on this test set we decided to discard the set. This results in the 40 test sets described throughout the report.

same number of participants. Balancing the pairs allows better comparison of test set performance. This was not likely an issue for us as most test sets had perfect accuracy; however, the use of an established block design is the right thing to do.

A balance was struck between the total number of test sets and the number of test sets evaluated by each participant. We had approximately 200 individuals express interest in the study. We assumed that about 50-75% of those would actually complete the study. We therefore elected for 40 test sets with 16 sets evaluated per participant. In the end, we had 107 participants complete the study resulting in every test set having been examined by approximately 41 different individuals.

Each of our forty test sets consisted of three test fires (two knowns from the same firearm and an unknown). Each test set was either a known match (KM) or known non-match (KNM). Approximately forty percent of our test sets were KMs. It's possible through a random BIBD design that a participant (block) could end up randomly with zero KMs or zero KNMs. We felt this could affect the reported results and so we enforced that all participants had between 4 and 11 known-matches in their test sets. This does not affect the BIBD design criteria.

In practice, 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 signed up for the study (and were assigned test sets to analyze) but never completed the analysis. This slight incompleteness does not impair our ability to compute performance statistics.

3.4 VCM Study (Aim 3)

Many details of the study were presented in previous sections on software development and dataset design. Study participants were recruited via AFTE forums as well as conference and workshop presentations. One of these workshops was the VCM workshop we helped run at the 2018 AFTE meeting in Charleston, West Virginia. The workshop was well attended and those participating were able to work with our VCM software on their laptops. Many workshop participants completed the workshop excited and eager to participate in the VCMER study. Once the study began, participants were given approximately eight weeks to complete the study. The study was designed to require approximately five to eight hours to complete. Feedback suggests that we hit that mark. Therefore we believe that all participants were provided ample time to complete the study.

Training. All participants were provided a training booklet (in pdf format) which taught them how to use the software. All core software functionality was demonstrated through the visualization of a

number of test sample scans. The training materials also included a practice proficiency test (3 knowns, 4 unknowns) which needed to be completed successfully before the participant was allowed to advance to the actual test sets. We note that the majority of participants had not used our software before. As described in the next paragraph only 5% of participants reported regular use of 3D visualization tools; therefore 95% of participants were new or relatively new to 3D VCM.

Participant Demographic Breakdown. The 107 participants came from seven different countries. The USA had 63 participants, Canada had 21 participants, and the rest of the world had 23 participants. Note that we will use the term International to refer to countries other than the US and Canada. **Qualification:** 97 (91%) of the participants were self-reported to be qualified to perform independent casework, 10 (9%) were self-reported to be not qualified to perform independent casework (*e.g.*, they were trainees). **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, 35 (33%) had more than ten years of experience. **Hardware:** 80 (75%) utilized desktop computers while 27 (25%) used laptop computers. There was an interesting breakdown by country. 72% of Americans used desktops, 100% of Canadians used desktops, and 61% of International participants used desktops. **VCM Experience:** USA: 3% Use Routinely, 73% Used VCM a Few Times, 24% No Experience. Canada: 5% Use Routinely, 62% Used VCM a Few Times, 33% No Experience. International: 9% Use Routinely, 43% Used VCM a Few Times, 48% No Experience. **Confidence:** Upon completion of the study and at the time of result submission, participants were asked to rate their confidence in their conclusions. USA: 79% Very Confident, 21% Somewhat Confident, 0% Not Confident. Canada: 52% Very Confident, 43% Somewhat Confident, 5% Not Confident. International: 70% Very Confident, 26% Somewhat Confident, 4% Not Confident. **Lab Policy Allows Elimination on Individual Marks[3]:** USA:87% Yes. Canada: 33% Yes. International: 87% Yes.

4 Data Results and Analysis

In this section we summarize the experimental results. The work products of Aim 1 and 2 went into the completion of the VCMER Study. The results of the VCMER Study (Aim 3) are presented in this section.

The primary group of participants whose results are most important to our study is the group of 76

[3]This is an important detail as individuals from labs which are not allowed to eliminate on individual marks will not be able to eliminate any tests in our set. The strongest negative statement of association they could make is an Inconclusive C. This will be relevant later.

qualified examiners from the US and Canada (56 from the US, 20 from Canada). This core group of participants represents the primary users of our VCM technology. Trainees (and others not qualified to perform independent casework) offer interesting insight into the use of new technology; however, their lack of experience within the discipline can cause them to make errors that would not be made by those who are qualified. We refer to the set of Qualified US and Canadian examiners as USCAN examiners below. Unless otherwise specified the described results are for the group of USCAN examiners.

4.1 Statistics

US and Canadian qualified examiners demonstrated a total of three errors (of 1184 comparisons). One participant made two errors and a second participant made the other error. All three errors were false positives. The overall error rate was therefore 3 of 1184 or 0.2%. It's difficult to compare error rates across different studies due to a number of test design variables. For example, the included firearm makes/models, the ammunition used, the complexity of the samples, the explicit samples included, the examiners that participated, and the overall test design. However, previous studies using traditional light comparison microscopy (not VCM) suggest error rates between 0.0 and 1.6% [1, 11, 12]. Therefore the error rate achieved with our technology falls towards the lower-end of that range.

The overall positive predictive value, defined as the number of ID calls which are actually KM is 453/456 = 99.3%. The overall negative predictive value, defined as the number of Elimination calls which are actually KNM is 436/436 = 100.0%. The sensitivity, defined as the number of KM called as ID is 453/491 = 92.2%. Because some labs are not able to eliminate on individual marks, the specificity can be defined using two options for the negative call, either 'Elimination only' or 'Elimination or Inconclusive-C'. The specificity defined as the percentage of KNM called as either Inc-C or Elimination is 607/693 = 87.6% and when defined as the percentage of KNM called as Elimination is 436/693 = 62.9%. These numbers support the hypothesis that VCM using our technology is an excellent alternative to traditional microscopy.

Participant A (2 Errors). Participant A is a qualified examiner with 2 years of experience from the US or Canada. This individual made 9 correct conclusions, 2 errors, and listed 5 inconclusives. They come from a lab which is able to eliminate on individual characteristics. Therefore Participant A's inconclusive rate is higher than other USCAN examiners.

This person made a false positive on test set 18 (.40 S&W). The knowns and unknowns of set 18 were both from M&P 40s. The knowns used Winchester ammo with a nickel primer, the unknowns used

Winchester ammo with a brass primer. No other USCAN examiner issued a false positive conclusion on this set. 88% of USCAN examiners reported Elimination or Inc-C for this set. Participant A did not indicate any areas of similarity on the unknown scan for test set 18; therefore, we are unable to determine the basis for his/her conclusion. The summary similarity and difference maps for test set 18 are shown in Figure 12.

This person also made a false positive on test set 27 (9mm Luger). The knowns of set 27 came from a Springfield Armory XD9 sub-compact (Remington ammo with nickel primer), and the unknown came from a Smith & Wesson 3913 (Winchester ammo with nickel primer). No other USCAN examiner issued a false positive conclusion on this set. 85% of USCAN examiners reported Elimination or Inc-C for this set. Once again, Participant A did not indicate any areas of similarity on the unknown for test set 27; therefore, we are unable to determine the basis for his/her conclusion. Test set 27 annotation maps are shown in Figures 13.

Participant B (1 Error). Participant B is a qualified examiner with over 10 years of experience from the US or Canada. This individual made 14 correct conclusions, 1 error, and 0 inconclusives[4]. The individual comes from a lab which is able to eliminate on individual characteristics. Participant B's inconclusive rate is lower than other USCAN examiners.

This person had a false positive on test set 34 (9mm Luger). The knowns of set 34 came from a Sig Sauer 226 (Remington ammo with nickel primer), and the unknown came from a Smith & Wesson 639 (Winchester ammo with nickel primer). No other USCAN examiner issued a false positive conclusion on this set. 84% of USCAN examiners reported Elimination or Inc-C for this set. The annotation maps for test set 34 are shown in Figure 14. Participant B was very diligent in completion and annotation of their test sets. They provided annotations between the unknown and a known for all test sets except test set 34 for which they only provided an annotation between the two knowns. We hypothesize that the examiner made a mistake when using the software and reported a conclusion for the comparison of the two knowns to each other and not to the unknown. In hindsight, we see how this could happen. That is, the participant could mistakenly load the two knowns when they thought they had loaded one of the unknowns. This is something we can prevent in software. We can change the software so that participants can only record a conclusion when the software is displaying a known and an unknown. We note that specimen 'mix-up' is something that can happen with traditional (physical) proficiency tests and error rate studies. Updating the software to reduce the risk of such a mixup could be another advantage of VCM over traditional

[4]Participant B did complete analysis of 16 test sets; however, one of the sets was the set we discarded for all participants because of a typo.

microscopy.

We note that none of the three false positives as submitted would pass a laboratory quality assurance procedure where the examiner is required to 'show their work' by indicating the individual characteristics utilized to reach their conclusions. Participant A's errors would be flagged because they did not mark any areas on the unknown and Participant B's error would be flagged because they didn't mark any areas of the known or unknown. In both cases, the examiners would be required to go back and justify their conclusions and a second examiner might be called in to review the conclusion.

4.2 Test Set Results

The results of the KM test sets are shown in the top half of Table 1. For KM test sets we also show the percentage of ID responses. Note that almost all KM test sets had 100% ID recall. The two exceptions are test sets 4 and 37. Test set 4 is a Kahr K40 which produces an inconsistent aperture shear. That is, the two known test fires do not have an aperture shear while the unknown does show an aperture shear. In fact, one trainee made a false elimination on this test set. The annotations provided by the trainee show that the aperture shear was indeed the primary reason for their false elimination (Figure 4). The annotation maps for test set 4 are shown in Figure 5. Test set 37 is a Beretta PX4. Berettas have a countersunk firing pin aperture which typically results in a very small area for breech-face impression toolmark transfer. Test set 37 is no exception. We expected almost all participants to indicate inconclusive; however, a number of participants found regions of agreement between the knowns and unknown and were able to correctly ID the unknown. The annotation maps for test set 37 are shown in Figure 9.

The results of the KNM are shown in the bottom half of Table 1. For the KNM test sets we also list the percentage of Inconclusive-C or Elimination responses (grouped because some labs are not able to eliminate on individual marks). The two test sets with the least recall were test sets 17 and 36. Test set 17 is a Taurus PT111 with thin filing marks. The difference map shows that participants cued into only one real area of difference (Figure 12). The small quantity of individual differences likely was not enough support for most examiners to eliminate. Test set 36 is an H&K USP Compact. The firearm appears to be minimally marking and the difference map shows just two small areas of noted difference (Figure 14). Once again, the minimal areas of geometric difference did not likely provide enough support for most examiners to eliminate.

The vast majority of KM test sets were correctly identified as same source by USCAN examiners. Figures 5-9 show that examiners marked significant areas of similarity. The majority of KNM test sets

were also correctly recognized as different source. Figures 10-15 show that examiners were able to recognize consistent areas of differences for KNMs. We did not notice a consistent trend between the number of inconclusive results among the low, medium, and high complexity samples with the exception of the test sets noted above. That is, test sets 4, 36, and 37 had higher inconclusive rates and were all listed as high complexity. Interestingly the other test set described above, 17, was classified as a low complexity sample.

4.3 Use of a 5-Point Reporting Scale

Although not an explicit aim of this study, we can investigate the use of a five-point scale. The study instructions included descriptions for each of five conclusions (AFTE range of conclusions) and the text (Fig 16) was also available through the software. Unfortunately, our questionnaire did not ask if participants routinely use a 5-point scale and we note that many of our participants likely do not routinely utilize three inconclusives. This lack of uniformity in use of the 5-point scale may lead to different interpretations of the scale.

The use of the 5-point scale among US and Canadian qualified examiners is shown in Table 2. The use of Identification and Elimination were as expected. Of the comparisons called as Identification, 453 of 456 (99.3%) were indeed KM. Of the comparisons called as Elimination, 436 of 436 (100%) were indeed KNM. The use of Inconclusive-C is also as expected. Many labs were not allowed to eliminate on individual marks and as such, Inc-C may be the strongest statement of non-association allowed by those participants. We therefore expect a large number of Inc-C which might otherwise be called as Elimination. Inconclusive-A provides an interesting insight into the use of a 5-point scale. Inc-A is intended to indicate some agreement of individual characteristics but insufficient for identification. However, only 1/3 of the comparisons labeled as Inc-A were indeed from the same firearm. This result is consistent with other (not yet published) research for traditional comparison microscopy (*i.e.*, not 3D VCM). These results suggest that additional work, perhaps in terms of education or framing of conclusions, may be required to ensure appropriate consumption of information contained in the label of Inc-A for both traditional and virtual comparison microscopy.

4.4 Results Summary

The overall performance of USCAN examiners was excellent. The overall error rate of 0.2% compares favorably with the error rates reported for traditional microscopy (typically 0.0% - 1.6%) [1, 11, 12].

Our error rate may be even lower as we hypothesize that one of the errors was an operational error in which the examiner appeared to have compared the two knowns to each other rather than the known to the unknown. Interestingly, the remaining two errors were made by one individual, indicating that examiner training may be more culpable than the 3D technology. In fact, there were no other errors among examiners who saw the same test sets suggesting that the scans contained sufficient information for a reliable conclusion. The phenomenon whereby most errors tend to be made by a small number of participants is supported by previous error rate studies.

The annotation maps provided significant insight into the examiner decision process. Errors among trainees can be visualized and understood (*e.g.*, test set 4 Figure 4). Among trained examiners, annotations can form a valuable part of the verification process. Conclusions submitted without supporting annotations may be flagged by a QA/QC process. All three errors made by qualified examiners would have been flagged by such a process. Annotation maps also support the hypothesis that examiners follow a common, rigorous, and consistent examination process. The consistency seen in the maps of Figures 5-15 represent the assembled independent annotations of 76 different examiners.

Three of the four test sets which reported increased inconclusive rates were as expected. These test sets were minimally or inconsistently marked. We noted no trend in inconclusive rates among the other test sets labeled as low, medium, or high complexity. As expected, the KM results were impressive with virtually 100% recall (except for the two high complexity sets as described in Section 4.2). Also as expected, the KNM results had a higher inconclusive rate than the KMs with a large number of Inconclusive-C calls. There were no false eliminations.

4.5 Continued Deployment Study

As we have during each of our previous awards, we continue to collaborate with crime labs. Through most of the project period we had a machine setup with Blake Reta and his lab at the West Virginia State Police crime lab in Charleston. At the beginning of the project period, Ryan Lilien went down and provided a day of hands-on training to all examiners in the lab. Then during the deployment period Blake assumed the main point-of-contact within the lab. The WV State Police collected over 400 scans during the deployment period. Upon completion, we brought the scanner back to Chicago and added the scans to our growing research set. During the deployment Blake provided feature suggestions and usability feedback. Through deployments like these we continue to collect scan data, to elicit excellent feedback from practitioners, and to train examiners and trainees.

5 Scholarly Products Produced

The primary product of the proposed research is the presentation of our results and progress. At the May 2018 AFTE national meeting we gave two technical presentations. One presentation took place during the main technical session and was entitled "Recent Progress Towards 3D Virtual Comparison Microscopy". At the same meeting we co-ran a virtual microscopy workshop titled "Implementation of 3D Technology, Analysis, and Statistics for FA/TM Examinations". During the full-day workshop participants had hands-on time with our virtual microscopy software. They worked through a training tutorial and a virtual CTS test. During the project period Lilien also presented our work on validating virtual microscopy at the Eastern Regional AFTE meeting (FBI Organized, Fredericksburg, VA), the Midwest Firearm Examiner Training Seminar (Indianapolis, IN), the London Metropolitan Police Crime Lab Firearms Unit (London, UK), the European Network of Forensic Science Institutes Firearms Meeting (Copenhagen, Denmark), the Netherlands Forensic Institute (The Hague, Netherlands), and the National Firearms Examiner Academy (NFEA) (Gaithersburg, MD). A shortened version of this final report is being submitted for publication as a paper titled "Results of the 3D Virtual Comparison Microscopy Error Rate Study (VCMERS) for Firearm Forensics". During the project period our paper on our previous VCM study was published, "Development and Validation of a Virtual Examination Tool for Firearm Forensics" [5]. The above publications and presentations continue our pattern of disseminating our research results. Over the past several years, we have presented at more than 25 forensic conferences and run training sessions at fourteen local, state, and federal crime labs.

6 Summary

We successfully completed the proposed aims during the project period. In Aim 1, we developed an updated VCM testing platform capable of supporting most modern Windows computers. The updated software has a testing mode which is designed to facilitate implementation of validation studies. The software has a network access option which simplifies the acquisition and submission of validation study data. Also in Aim 1 we developed an algorithm testing software tool to assist in the future evaluation of different comparison algorithms. In Aim 2 we selected test fires for our forty test sets paying attention to include firearms with a range of toolmark types and complexity. Test fires were scanned and assembled for the validation study. Finally in Aim 3, we completed the largest VCM validation and error rate study to date. Our study involved 107 participants including 76 qualified examiners from the US and Canada.

These examiners demonstrated highly accurate analysis. The error rate of 3 in 1184 (0.2%) compares favorably with previous studies for traditional light comparison microscopy (0.0-1.6%). Of these three errors, two came from the same examiner (suggesting a training issue) and one is believed to be an operations error (where the examiner did not compare the known to the unknown).

Overall the study provides extremely strong support for the use of Cadre's hardware and software tools for 3D Virtual Comparison Microscopy. It is important to note that the validation accomplished via the VCMER Study and the error rates reported only apply to the use of Cadre VCM, using our 3D scanning hardware and our VCM software. Different scanning technologies produce different quality scans and the results presented in this report do not necessarily generalize to other technologies. For example, some 3D scanners are known to have difficulty scanning striated lines such as the aperture shear. The annotation maps shown above suggest that the aperture shear is a very important toolmark for both identification and elimination. Therefore, a system which does not accurately measure the aperture shear will likely have trouble with those test sets. Each different VCM platform needs to be validated on its own.

Appendix

Implications for Criminal Justice Policy and Practice

Our primary impact has been the continuing development of a novel 3D imaging and analysis system with reduced cost and improved accuracy compared to existing solutions. Our work directly addresses several aims of the NIJ's Applied Research and Development in Forensic Science for Criminal Justice Purposes program. Through direct collaboration, networking, talks, seminars, and publications we have made many forensic labs (local, state, and federal), practitioners, and policy makers within the criminal justice system aware of this work. The completed project increases the quality and efficiency of forensic analysis, develops new instrumentation systems, and provides a novel approach to enhancing the analysis and interpretation of forensic data derived from physical evidence. The ability to utilize 3D Virtual Comparison Microscopy in actual casework provides examiners a number of functional advantages. Evidence supports the hypothesis that high-quality 3D VCM examination requires less time and results in more accurate conclusions than traditional microscopy. Our work developing 3D scanning and visualization tools and then validating this technology through large examiner-based studies ensures the successful adoption of this technology. As 3D VCM becomes more mainstream it will increasingly benefit the criminal justice system and its ability to present firearm identification and toolmark evidence in the courtroom.

Additional impact will be made as more crime labs become aware of the work and as we continue to disseminate results. At least ten crime laboratories have had access to our 3D scanning hardware and now close to two hundred practitioners have had access to our VCM software. This would not have been possible prior to receiving recent NIJ awards. For labs that currently have 2D imaging systems, our 3D system provides a significant improvement in imaging and match accuracy. For labs that currently have competing 3D imaging systems, we feel our system offers more flexibility and transparency with respect to how the scanner works as well as validated hardware and software tools on which conclusions can be based.

Figures and Tables

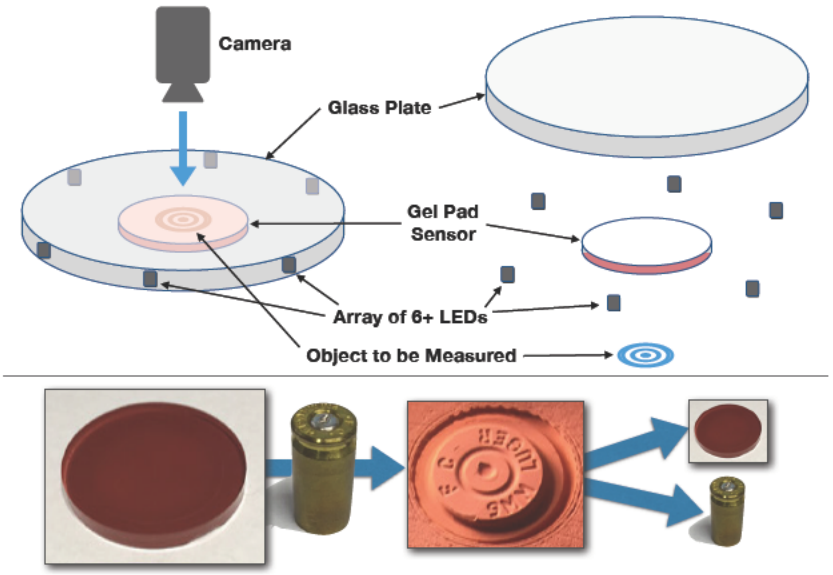


Figure 1: **GelSight Scanning Setup.** Our 3D scanning technique (GelSight) is based on the use of a silicone elastomeric pad with embedded micron-scale thick layer of pigment. (Top Row) The Gel Pad sensor is placed between a glass plate and the item being imaged. When the object to be measured is raised 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 our current scanners, 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 (5mm thick, 38mm diameter) allowing the pigment to conform to the cartridge surface. After scanning the cartridge is removed and the gel can be used again.

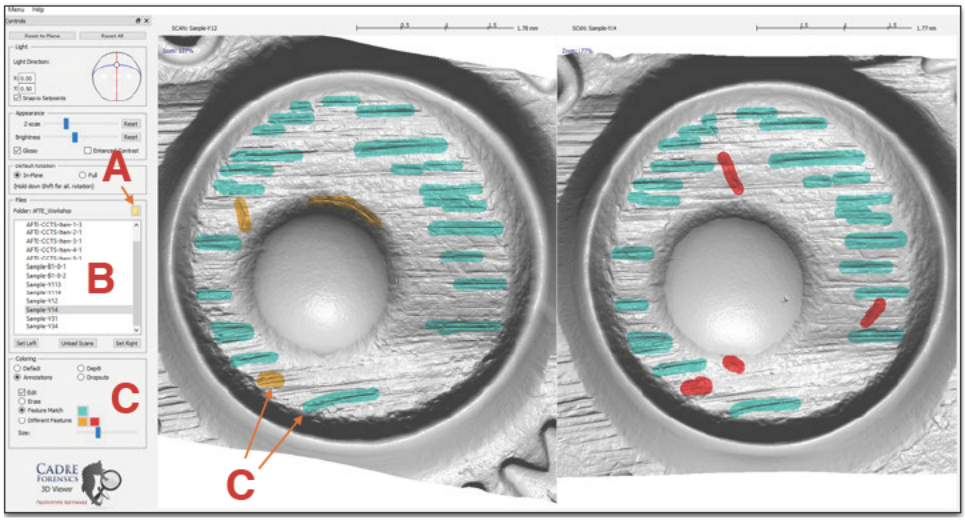


Figure 2: **Virtual Comparison Microscopy (VCM) Software.** The VCM software provides a virtual comparison scope. Examiners can adjust the virtual light position, manipulate the cartridge case orientation, position, and zoom (locked or unlocked). In a typical workflow, the user first selects a folder 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 can also talk with our Nexus internet server to retrieve scans uploaded from other locations.

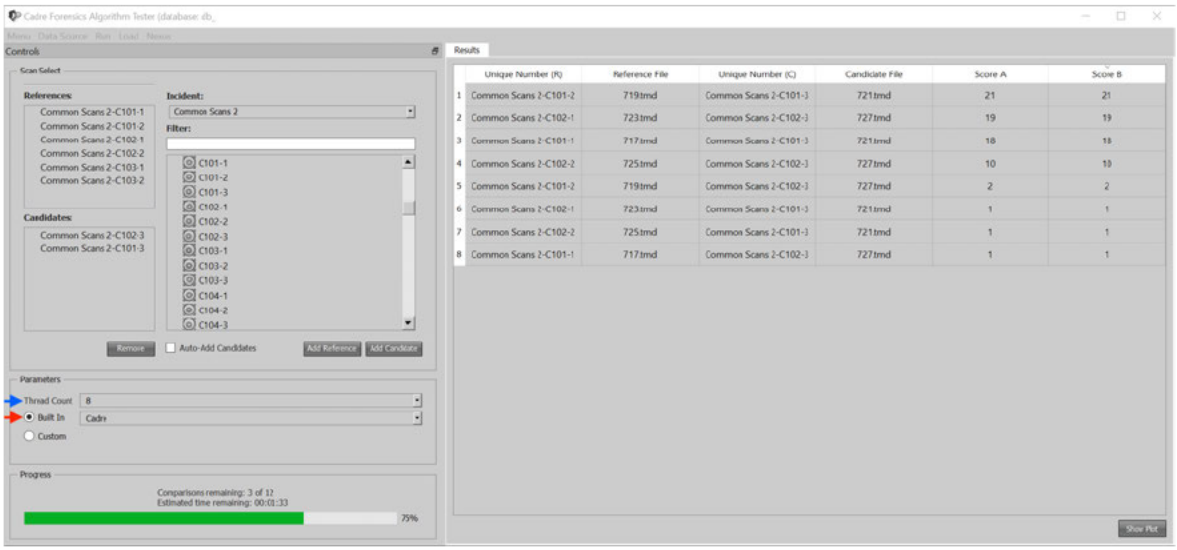


Figure 3: **Algorithm Testing Tool.** The developed algorithm testing tool allows the user to run a specified comparison algorithm against a specified dataset. The user can select the algorithm (red arrow) and the number of parallel jobs to run (blue arrow). Results are shown on the right side of the window.

Test Set 4

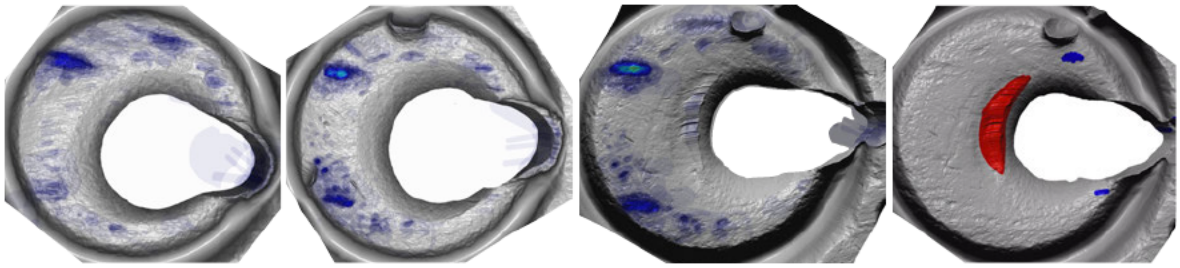


Figure 4: **Annotations for Test Set 4.** Test set 4 is a known match from a Kahr K40. The appearance of the firearms's aperture shear is inconsistent. The aperture shear is missing from the two known test fires but it appears in the unknown. The similarity maps for the known and unknown test fires are shown (legend for the three left images is shown in the first row of Figure 5). On the far right is the annotation of a trainee who made a false elimination. In the trainee annotation, red indicates a region of annotated difference and blue indicates an annotated similarity. The trainee marked the aperture shear (in red) as the basis for their elimination. This illustrates an excellent teaching opportunity for both this examiner and others. Other trainees can be presented this test set as part of their education program. The full similarity and difference maps for test set 4 are in Figure 5.

Test Set	Caliber	Make/Model	ID	INC-A	INC-B	INC-C	ELIM	
KM								**% ID**
1	.45 ACP	Springfield XD45	29	0	0	0	0	100.0
4	.40 S&W	Kahr K40	10	3	10	3	0	38.4
6	9mm Luger	Glock 19	35	0	0	0	0	100.0
9	9mm Luger	Intratec CAT 9	28	0	0	0	0	100.0
12	.40 S&W	Kahr CW-40	26	0	0	0	0	100.0
13	9mm Luger	Hi-Point C9	29	0	0	0	0	100.0
14	9mm Luger	Ruger SR9	29	0	0	0	0	100.0
19	.40 S&W	S&W SD40	29	0	0	0	0	100.0
20	.45 ACP	Springfield 1911-A1	32	0	0	0	0	100.0
23	9mm Luger	Kel-Tec P-11	30	0	0	0	0	100.0
24	9mm Luger	Norinco 213	29	0	0	0	0	100.0
25	9mm Luger	Ruger SR9	25	0	0	0	0	100.0
28	.40 S&W	Glock 22	26	0	0	0	0	100.0
29	9mm Luger	Fabrique Nationale	28	0	0	0	0	100.0
30	.40 S&W	S&W SW40	31	0	0	0	0	100.0
32	.45 ACP	Rock Island 1911	26	1	0	0	0	96.3
37	9mm Luger	Beretta PX4	11	6	15	0	0	34.4
KNM								**% INC-C or ELIM**
2	9mm Luger	Kahr MK9	0	2	1	10	16	89.7
3	9mm Luger	S&W 915	0	0	1	9	21	96.8
5	.40 S&W	Glock 22	0	1	2	7	18	89.3
7	9mm Luger	S&W M&P9	0	3	2	11	15	89.7
8	9mm Luger	Glock 17	0	0	1	4	25	96.7
10	9mm Luger	S&W SW9	0	1	2	4	22	89.7
11	.45 ACP	Glock 36	0	0	2	8	23	93.9
15	.40 S&W	Star Bonifacio Firestar	0	0	0	6	26	100.0
16	.40 S&W	Ruger P94	0	0	1	6	24	96.8
17	9mm Luger	Taurus PT111	0	4	7	8	14	66.7
18	.40 S&W	S&W M&P40	**1**	1	1	10	13	88.5
22	9mm Luger	FEG PJK-9HP	0	0	0	6	28	100.0
26	9mm Luger	Glock 26	0	1	0	7	24	96.9
27	9mm Luger	Springfield XD9	**1**	0	3	10	13	88.5
31	9mm Luger	Glock 19	0	1	1	6	21	93.1
33	.40 S&W	Glock 23	0	1	2	8	20	90.3
34	9mm Luger	Sig Sauer 226	**1**	1	3	5	22	87.1
35	.40 S&W	S&W SD40	0	1	8	10	12	71.0
36	.45 ACP	H&K USP Compact	0	1	16	5	4	34.6
38	9mm Luger	Glock 19	0	1	3	8	19	87.1
39	.40 S&W	S&W SW40	0	0	3	12	18	90.9
40	9mm Luger	Sig Sauer P938	0	0	0	5	18	100.0
41	.40 S&W	Springfield XD40	0	1	4	6	20	83.9

Table 1: **Results by Test Set.** Note that test set with number 21 was removed from the study because of a typo in the description listing it as .45 rather than .40 S&W.

Ground Truth	ID	INC-A	INC-B	INC-C	ELIM
KM	453	10	25	3	0
KNM	3	20	63	171	436

Table 2: **Use of 5-Point Scale.** Use of 5-point scale among US and Canadian qualified examiners. Of the comparisons called as Identification, 453 of 456 (99.3%) were indeed KM. Of the comparisons called as Elimination, 436 of 436 (100%) were indeed KNM.

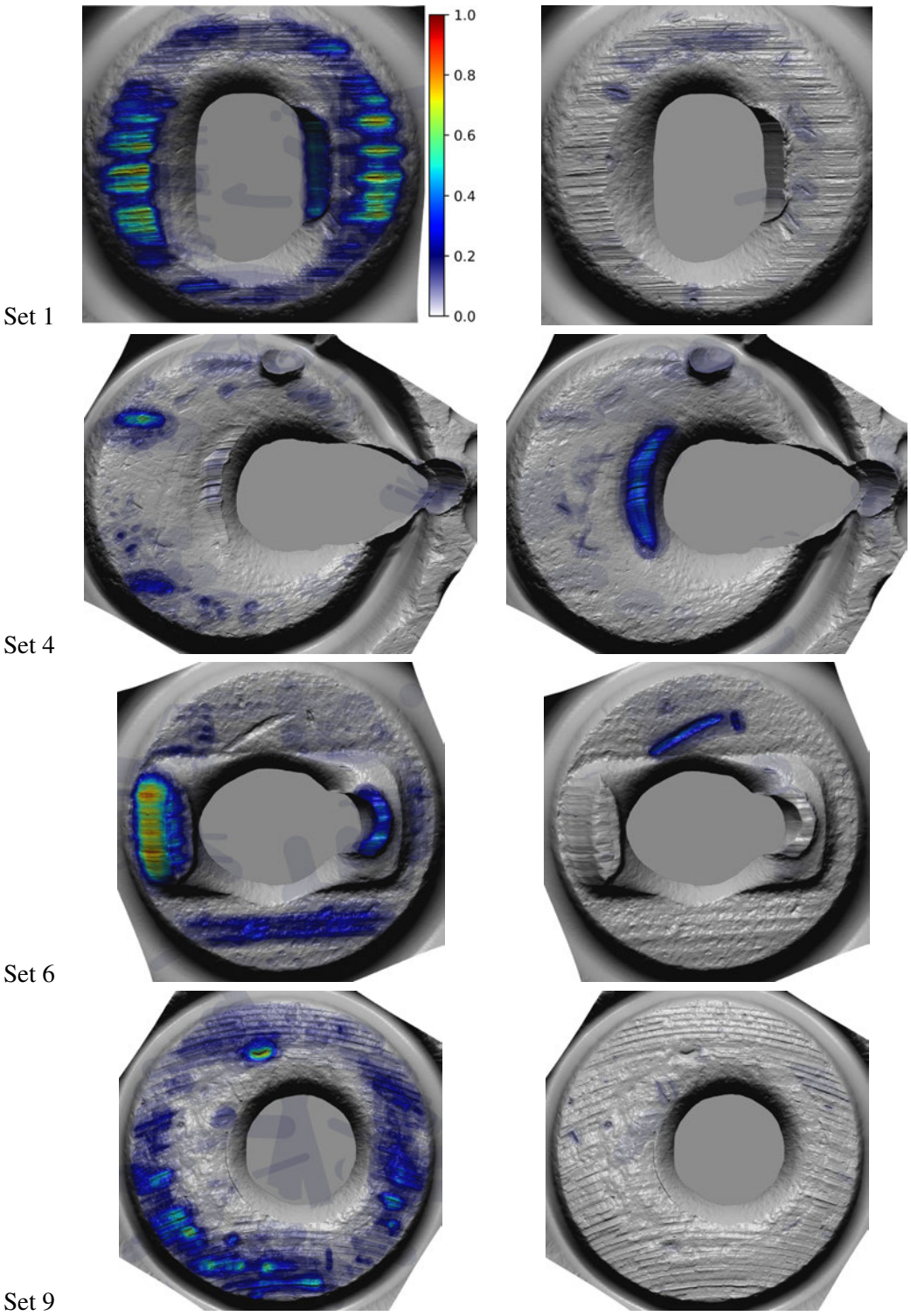


Figure 5: **Annotation Maps (KM).** Group 1 of Known Matches. (Left) similarity annotation map, (Right) difference annotation map. Unknown scan shown for both. Surface coloring indicates the fraction of participants that marked the indicated region as similar (or different). Color scale bar is shown next to Set 1 in Figure 5.

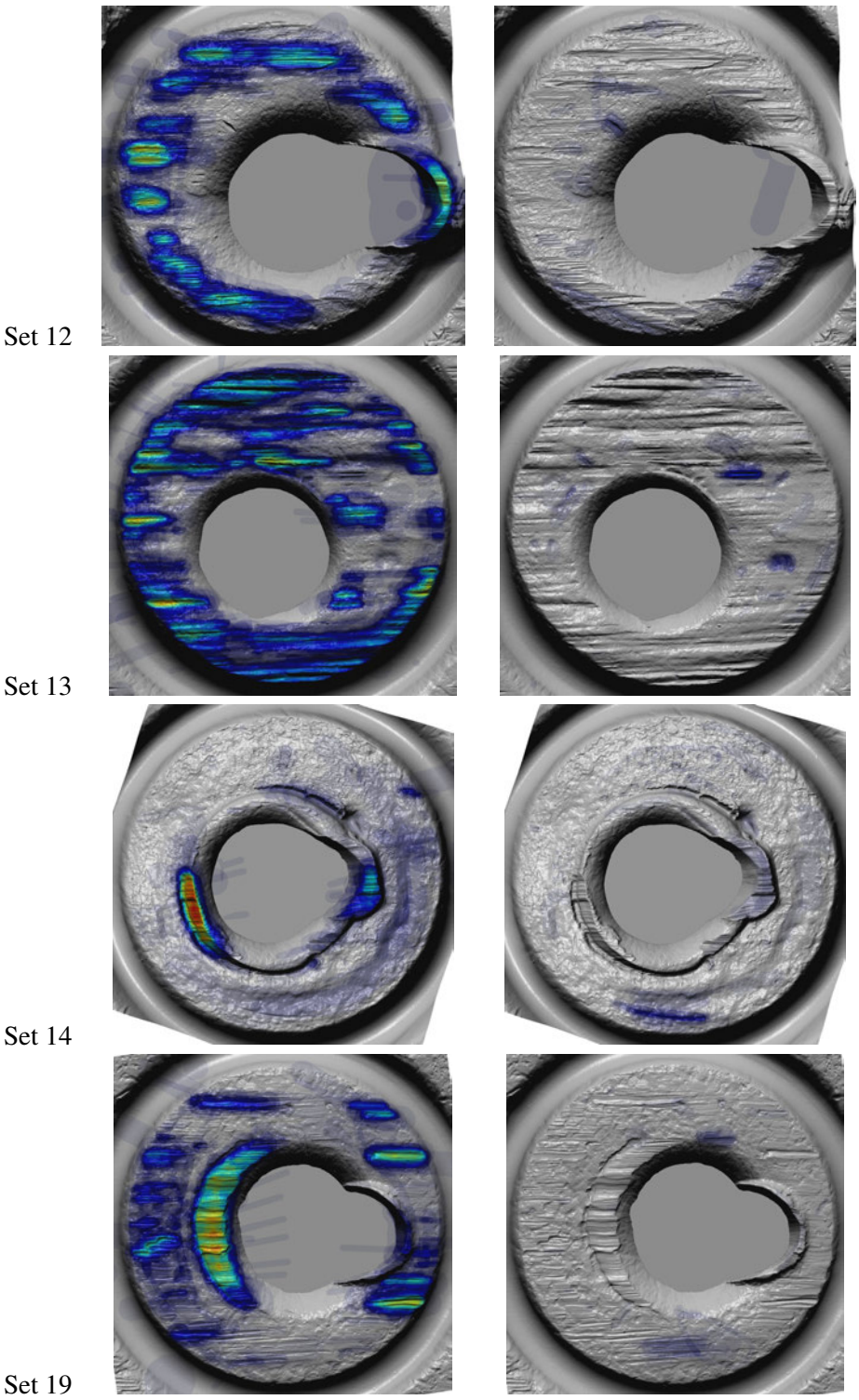


Figure 6: **Annotation Maps (KM).** Group 2 of Known Matches. (Left) similarity annotation map, (Right) difference annotation map. Unknown scan shown for both. Surface coloring indicates the fraction of participants that marked the indicated region as similar (or different). Color scale bar is shown next to Set 1 in Figure 5.

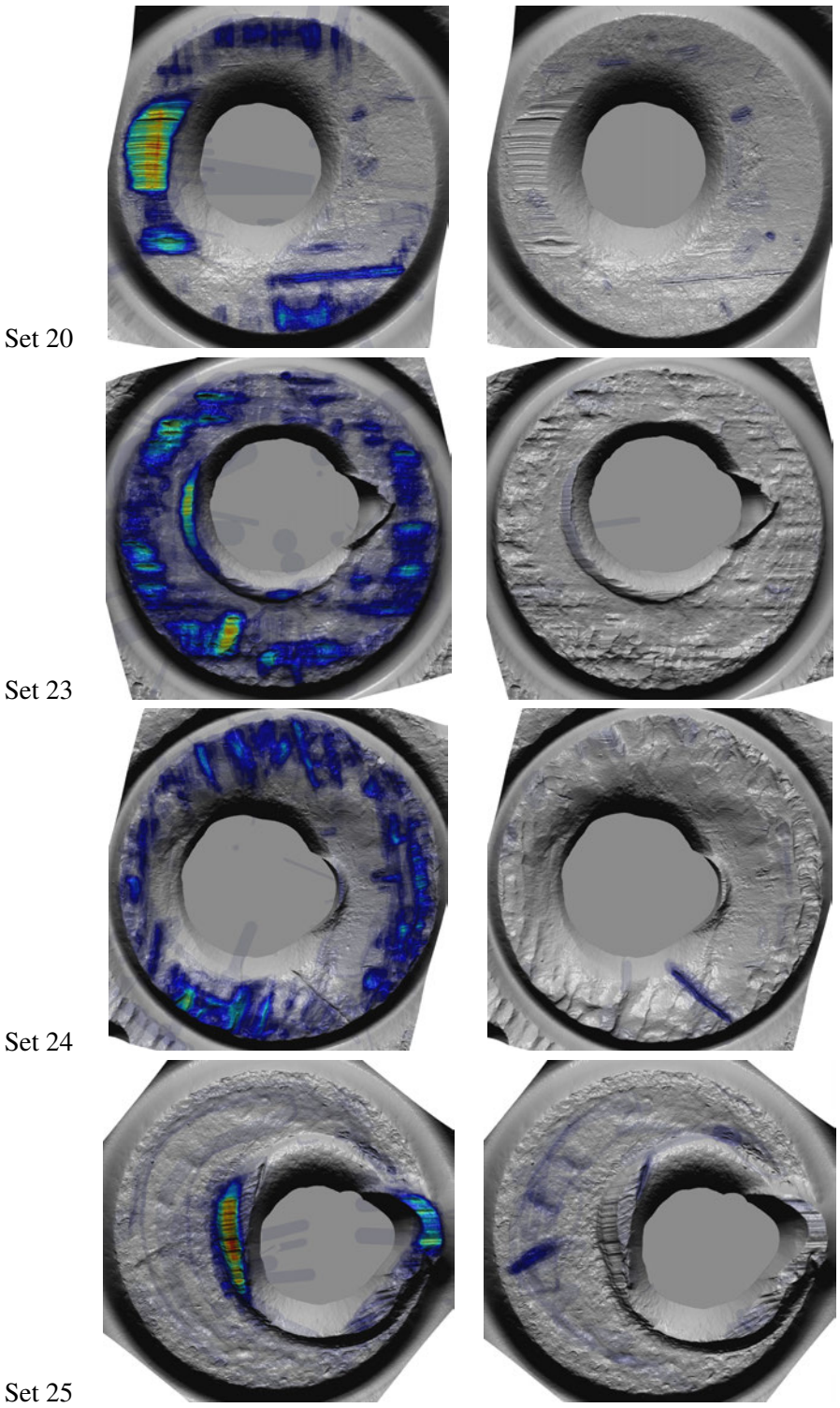


Figure 7: **Annotation Maps (KM).** Group 3 of Known Matches. (Left) similarity annotation map, (Right) difference annotation map. Unknown scan shown for both. Surface coloring indicates the fraction of participants that marked the indicated region as similar (or different). Color scale bar is shown next to Set 1 in Figure 5.

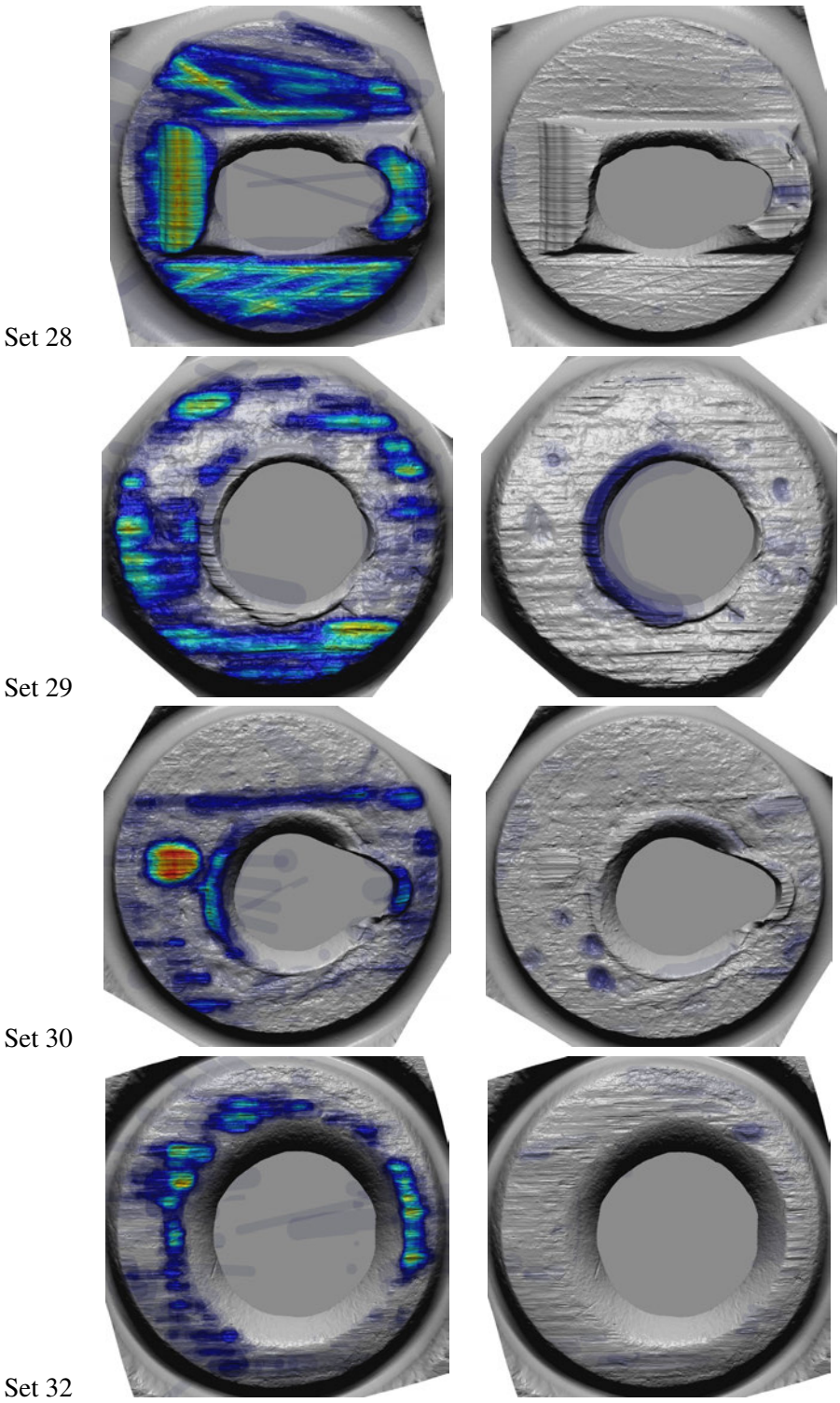
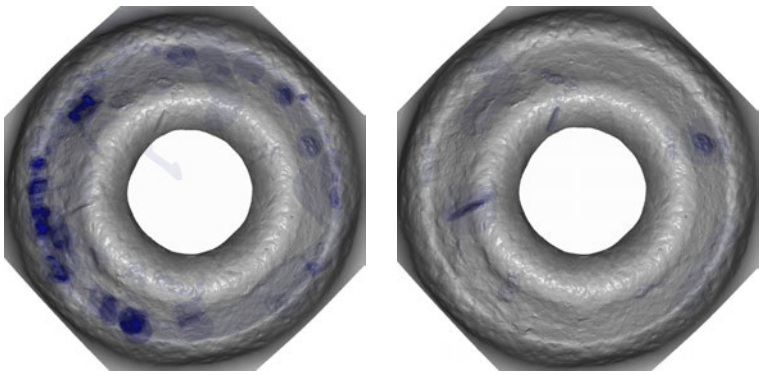


Figure 8: **Annotation Maps (KM).** Group 4 of Known Matches. (Left) similarity annotation map, (Right) difference annotation map. Unknown scan shown for both. Surface coloring indicates the fraction of participants that marked the indicated region as similar (or different). Color scale bar is shown next to Set 1 in Figure 5.



Set 37

Figure 9: **Annotation Maps (KM).** Group 5 of Known Matches. (Left) similarity annotation map, (Right) difference annotation map. Unknown scan shown for both. Surface coloring indicates the fraction of participants that marked the indicated region as similar (or different). Color scale bar is shown next to Set 1 in Figure 5.

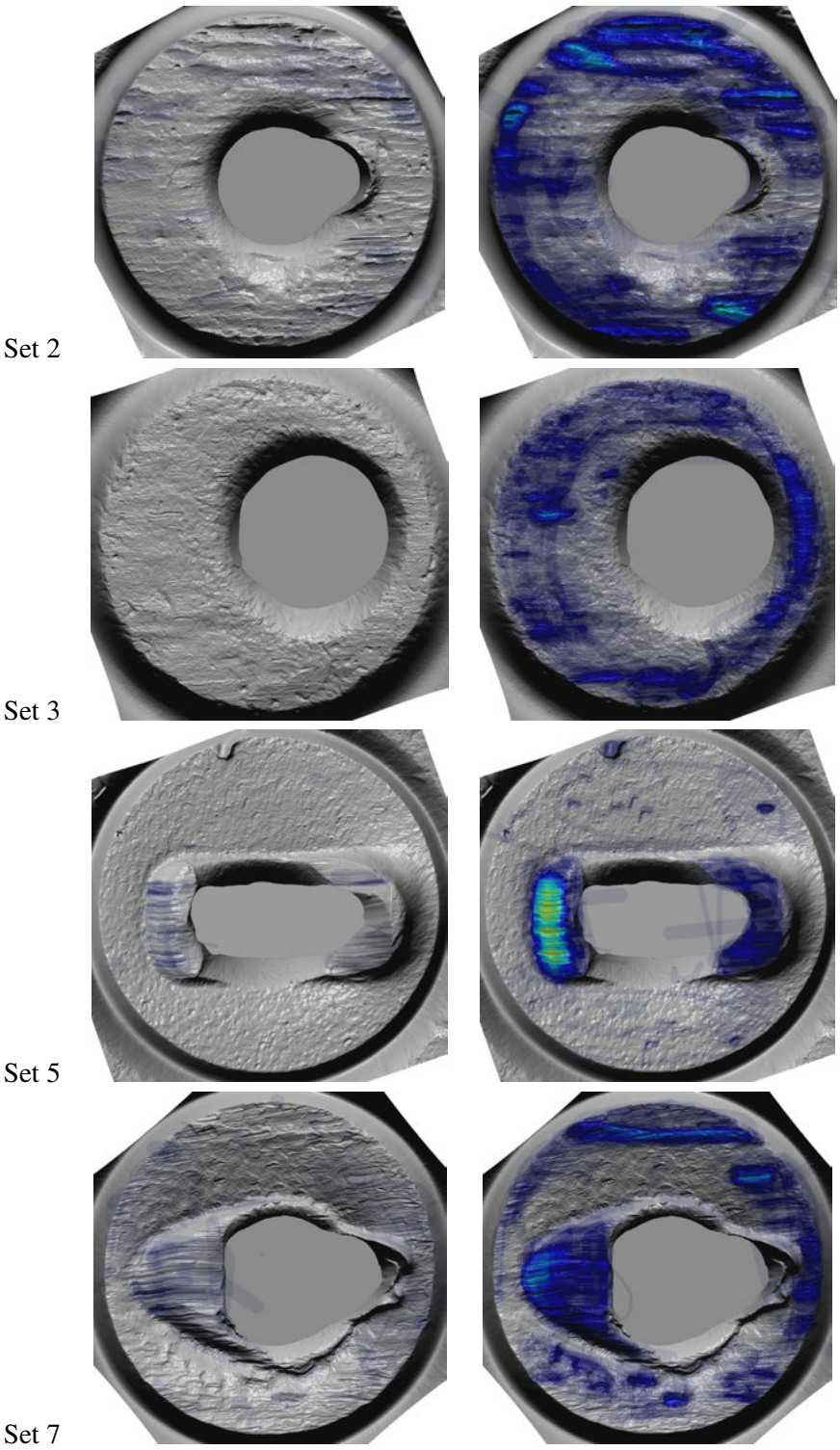


Figure 10: **Annotation Maps (KNM).** Group 1 of Known Non-Matches. (Left) similarity annotation map, (Right) difference annotation map. Unknown scan shown for both. Surface coloring indicates the fraction of participants that marked the indicated region as similar (or different). Color scale bar is shown next to Set 1 in Figure 5.

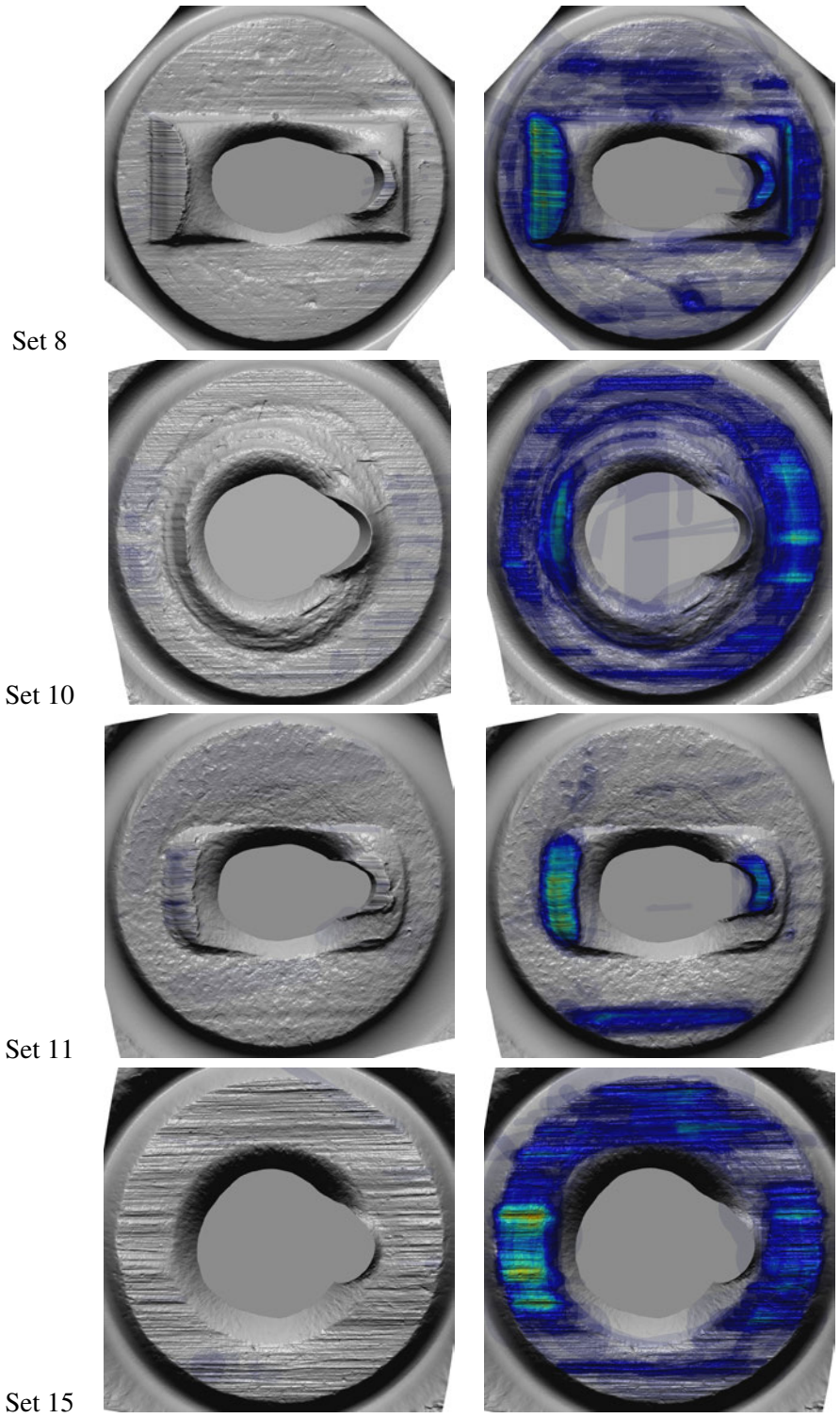


Figure 11: **Annotation Maps (KNM).** Group 2 of Known Non-Matches. (Left) similarity annotation map, (Right) difference annotation map. Unknown scan shown for both. Surface coloring indicates the fraction of participants that marked the indicated region as similar (or different). Color scale bar is shown next to Set 1 in Figure 5.

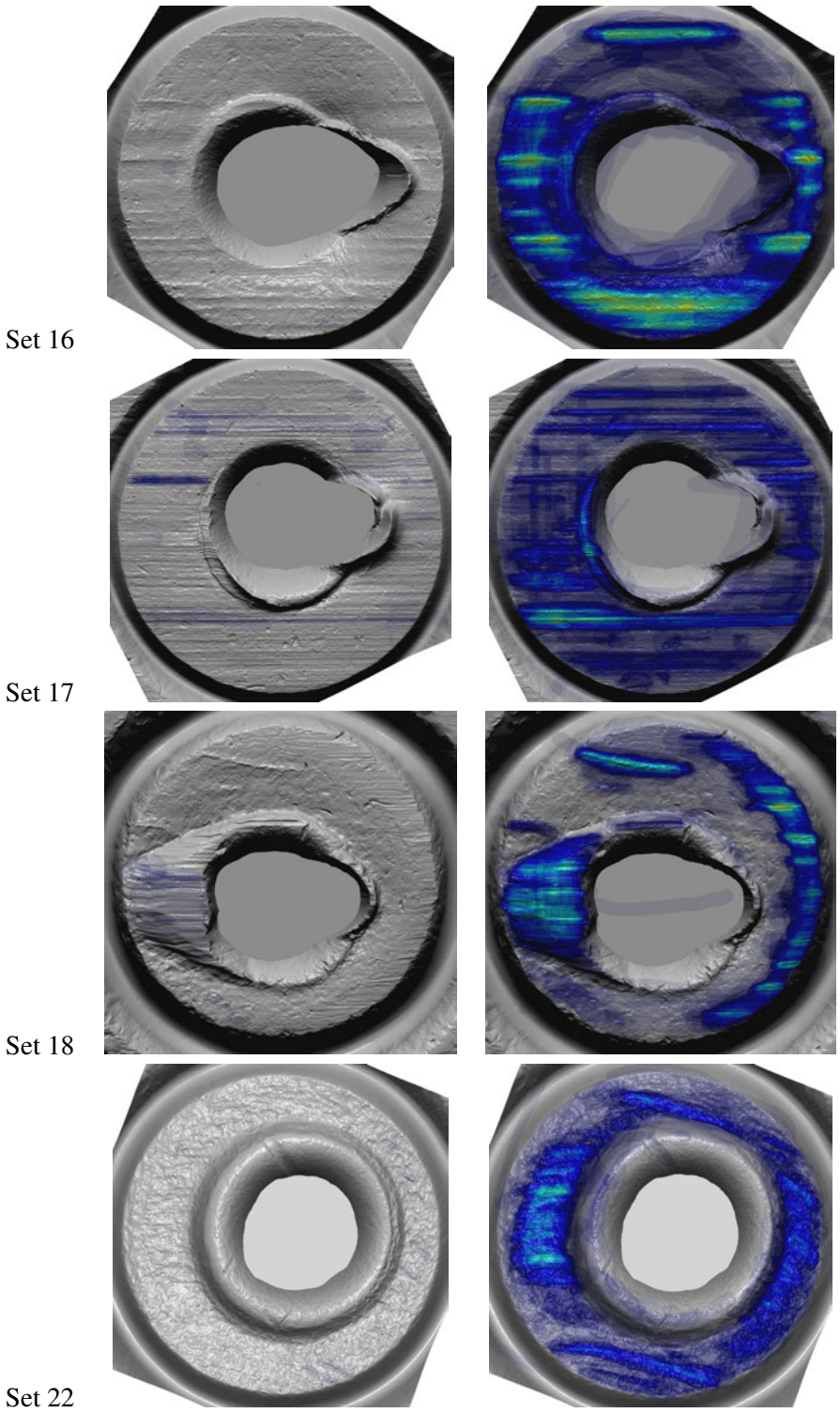


Figure 12: **Annotation Maps (KNM).** Group 3 of Known Non-Matches. (Left) similarity annotation map, (Right) difference annotation map. Unknown scan shown for both. Surface coloring indicates the fraction of participants that marked the indicated region as similar (or different). Color scale bar is shown next to Set 1 in Figure 5.

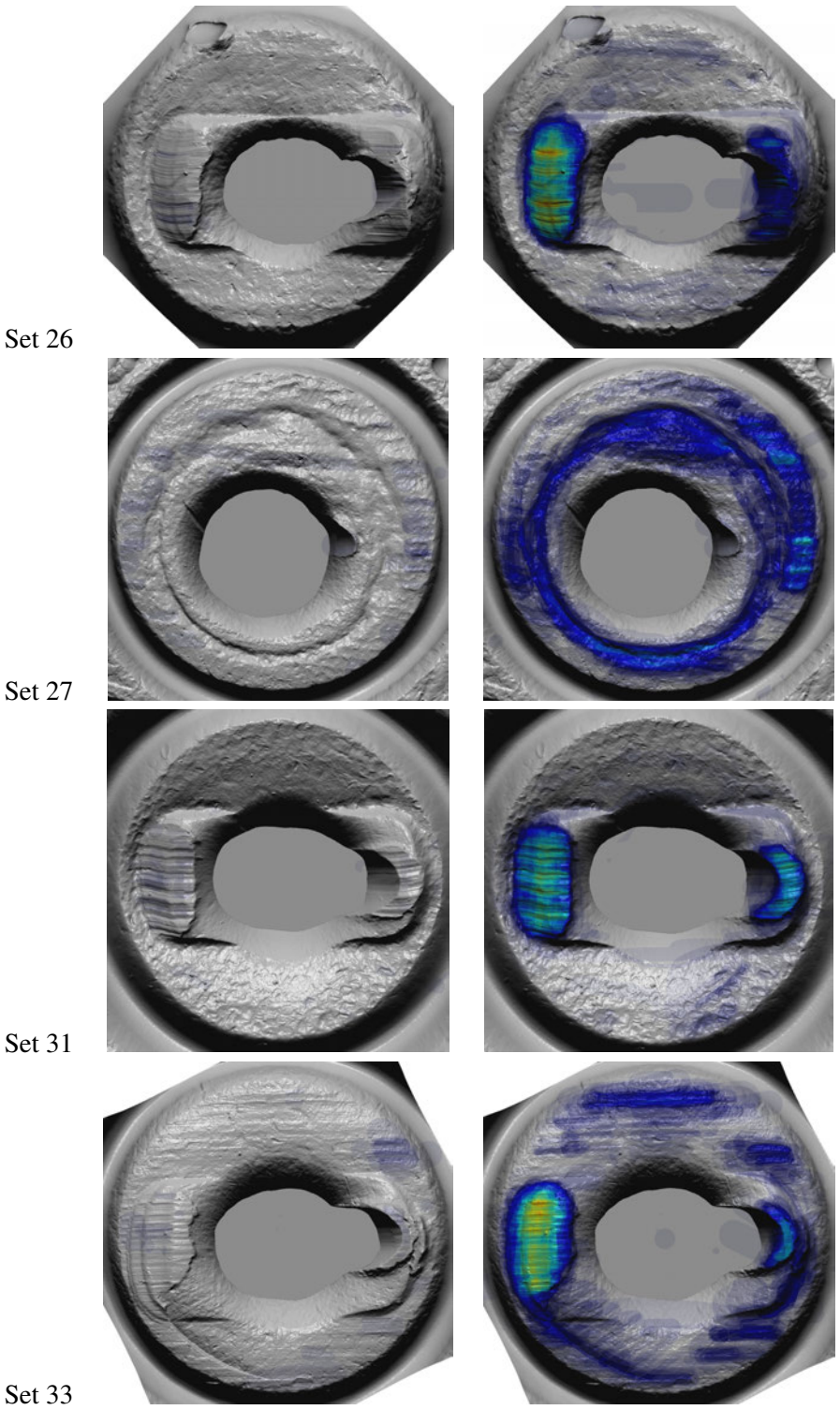


Figure 13: **Annotation Maps (KNM).** Group 4 of Known Non-Matches. (Left) similarity annotation map, (Right) difference annotation map. Unknown scan shown for both. Surface coloring indicates the fraction of participants that marked the indicated region as similar (or different). Color scale bar is shown next to Set 1 in Figure 5.

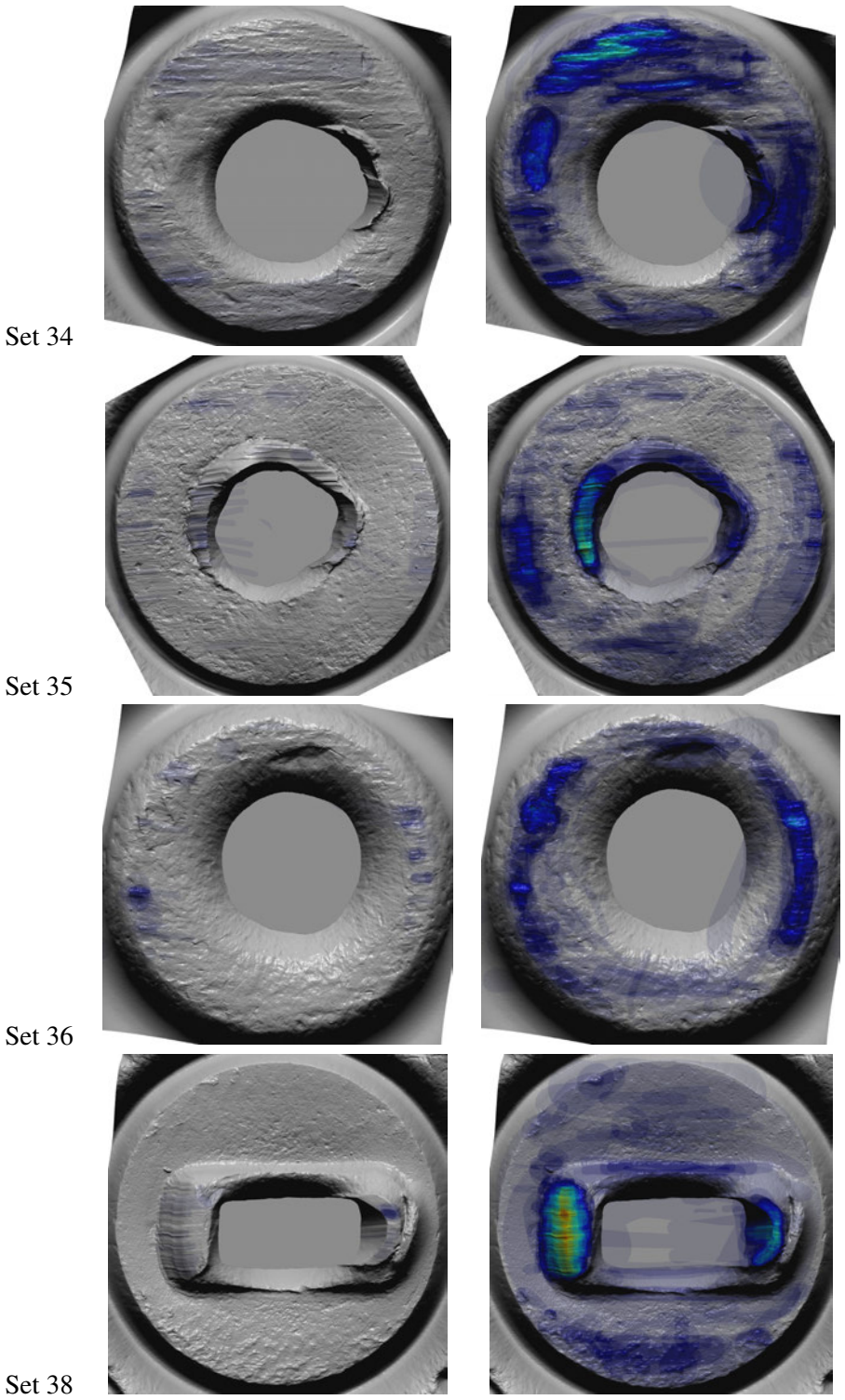


Figure 14: **Annotation Maps (KNM).** Group 5 of Known Non-Matches. (Left) similarity annotation map, (Right) difference annotation map. Unknown scan shown for both. Surface coloring indicates the fraction of participants that marked the indicated region as similar (or different). Color scale bar is shown next to Set 1 in Figure 5.

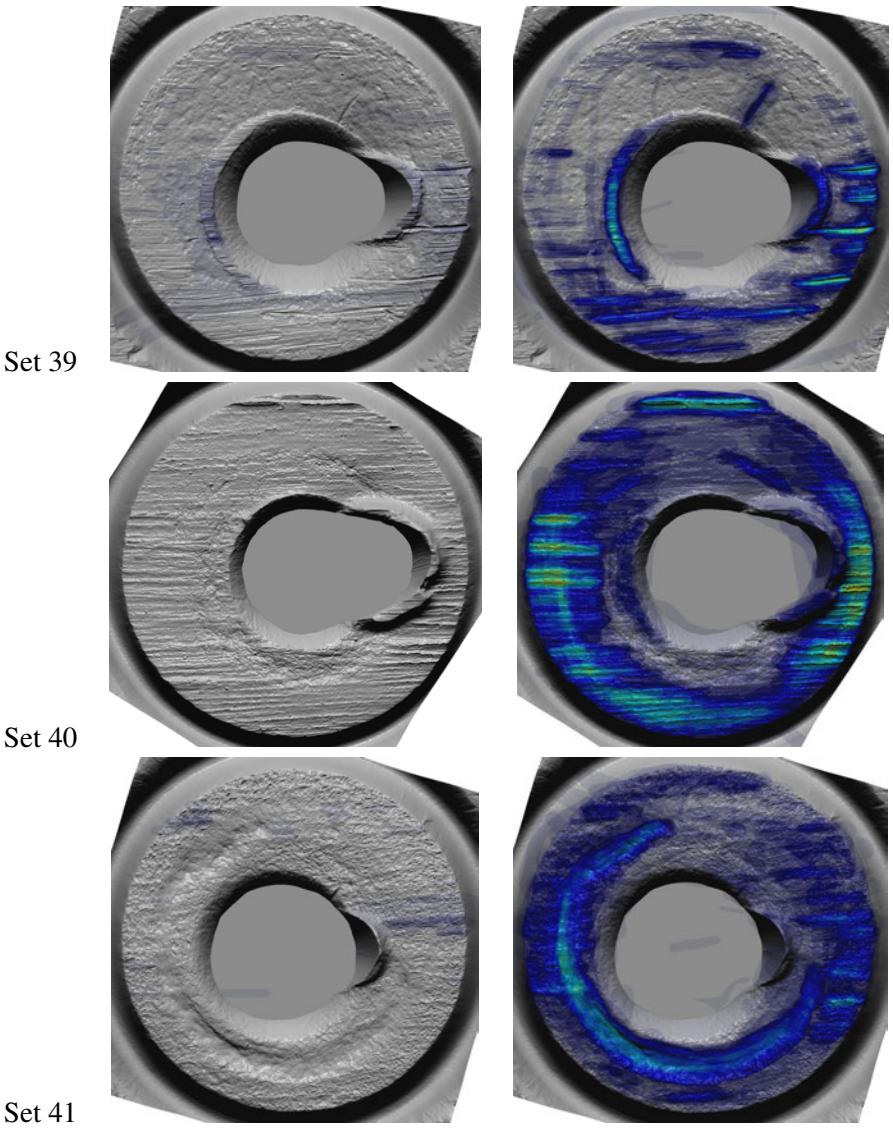


Figure 15: **Annotation Maps (KNM).** Group 6 of Known Non-Matches. (Left) similarity annotation map, (Right) difference annotation map. Unknown scan shown for both. Surface coloring indicates the fraction of participants that marked the indicated region as similar (or different). Color scale bar is shown next to Set 1 in Figure 5.

Please use the AFTE Range of Conclusions when indicating your results on the test worksheets. If your lab utilizes a different scale, please adopt the scale below as best you can. You may indicate additional clarification or qualification information in the 'comments' section of each worksheet.

Identification:

Agreement of a combination of individual characteristics and all discernible class 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: Some agreement of individual characteristics and all discernible class characteristics, but insufficient for an identification.

B: Agreement of all discernible class characteristics without agreement or disagreement of individual characteristics due to an absence, insufficiency, or lack of reproducibility.

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.

Figure 16: **Five-Point Range of Conclusions.** The five-point range of conclusions as presented to each participant.

References

[1] D. Baldwin, S. Bajic, M. Morris, and D. Zamzow. A study of false-positive and false-negative error rates in cartridge case comparisons. Technical Report IS-5207, Ames Laboratory USDOE, 2014.

[2] A. Banno, T. Masuda, and K. Ikeuchi. Three dimensional visualization and comparison of impressions on fired bullets. *Forensic Science International*, 140:233–40, 2004.

[3] R. Bolton-King, P. Evans, C. Smith, J. Painter, D. Allsop, and W. Cranton. What are the prospects of 3D profiling systems applied to firearms and toolmark identification. *AFTE Journal*, 42:23–33, 2010.

[4] W. Chu, M. Tong, and J. Song. Validation tests for the Congruent Matching Cells (CMC) method using cartridge cases fired with consecutively manufactured pistol slides. *AFTE Journal*, 45:361–6, 2013.

[5] P. Duez, T. Weller, M. Brubaker, R. Hockensmith, and R. Lilien. Development and validation of a virtual examination tool for firearm forensics. *J. Forensic Sciences*, (in press), 2018.

[6] J. Holdren, E. Lander, W. Press, and M. Savitz. Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods. Technical report, President's Committee of Advisors on Science and Technology, 2016.

[7] M. Johnson and E. Adelson. Retrographic sensing for the measurement of surface texture and shape. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1070–7, 2009.

[8] M. Johnson, F. Cole, A. Raj, and H. Adelson. Microgeometry capture using an elastomeric sensor. *ACM Trans. on Graphics, Proc. of SIGGRAPH*, 30:139–44, 2011.

[9] J. Roth, A. Carriveau, X. Liu, and A. Jain. Learning-based ballistic breech face impression image matching. *Proc. of the IEEE Conference on Biometrics (BTAS)*, 2015.

[10] N. Senin, R. Groppetti, L. Garofano, P. Fratini, and M. Pierni. Three-dimensional surface topography acquisition and analysis for firearm identification. *J. Forensic Sciences*, 51:282–95, 2006.

[11] E. Smith. Cartridge case and bullet comparison validation study with firearms submitted in casework. *AFTE Journal*, 37:130–5, 2005.

[12] T. Smith, A. Smith, and J. Snipes. A validation study of the bullet and cartridge case comparisons using samples representative of actual casework. *J. Forensic Sciences*, 61:939–46, 2016.

[13] J. Song. Proposed "NIST ballistics identification system (NBIS)" based on 3d topography measurements on correlation cells. *AFTE Journal*, 45:184–94, 2013.

[14] J. Song. Proposed "congruent matching cells (CMC)" method for ballistic identification and error rate estimation. *AFTE Journal*, 47:177–85, 2015.

[15] J. Song, T. Vorburger, W. Chu, J. Yen, J. Soons, D. Ott, and N. Zhang. Estimating error rates for firearm evidence identifications in forensic science. *Forensic Science International*, 284:15–32, 2018.

[16] M. Stocker, R. Thompson, J. Soons, T. Renegar, and A. Zheng. Addressing challenges in quality assurance of 3d topography measurements for firearm and toolmark identification. *AFTE Journal*, 50:104–11, 2018.

[17] B. Ulery, A. Hicklin, J. Buscaglia, and M. Roberts. Accuracy and reliability of forensic latent fingerprint decisions. *PNAS*, 108:7733–8, 2011.

[18] T. Vorburger. Optical methods of surface measurement. Presented at the NIST Measurement Science and Standards in Forensic Firearms Analysis Conference, 2012.

[19] T. Vorburger, J. Song, and N. Petraco. Topography measurements and applications in ballistics and tool mark identifications. *Surface Topography: Metrology and Properties*, 4:1–35, 2015.

[20] T. Weller, M. Brubaker, P. Duez, and R. Lilien. Introduction and initial evaluation of a novel three-dimensional imaging and analysis system for firearm forensics. *AFTE Journal*, 47:198–208, 2015.

[21] A. Zheng, J. Soons, R. Thompson, J. Villanova, and T. Kakal. 2D and 3D topography comparisons of toolmarks produced from consecutively manufactured chisels and punches. *AFTE Journal*, 46:143–7, 2014.