Document Title:	**Evaluation of 3D Virtual Comparison Microscopy for Firearm Forensics within the Crime Lab**
Author(s):	**Ryan Lilien**
Document Number:	**255942**
Date Received:	**December 2020**
Award Number:	**2018-DU-BX-0216**

Final Research Report - Cover Page

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

Federal Grant or Other Identifying Number: 2018-DU-BX-0216

Project Title: Evaluation of 3D Virtual Comparison Microscopy for Firearm Forensics within the Crime Lab

PD/PI Name, Title, Contact Info: Ryan Lilien, Chief Scientific Officer, Cadre Research Labs; 500 Davis Street, Suite 500; Evanston, IL 60201

Name of Submitting Official: Ryan Lilien

Submission Date: May 20, 2020

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

Project/Grant Period: Start: 1/1/2019, End: 12/31/2019

Reporting Period End Date: 12/31/2019

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 into the establishment of error rates. The 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. In Aim 1 we investigated the effect of scan resolution on an examiner's ability to reach accurate conclusions. In Aim 2 we deployed scanning a systems to two crime labs and evaluated the use of virtual comparison microscopy within the lab. 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. 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 with symbols (*e.g.*, arrows) 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, 17, 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 (X3P). Comparison algorithms are being developed to analyze these 3D surface topographies [5, 10, 14, 15, 16, 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.* [11] 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 transition is understanding the effects of scan resolution on the conclusions reached via VCM. To the best of our knowledge, the study presented below is the first investigation into this question.

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 [8, 9]. 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 several negative influences of surface reflectivity on the measured topography. For example, the gel-based imaging approach allows our system to capture fine striations which may appear washed out when measured via other technologies.

1.2 Previous VCM Studies

The 2016 PCAST report was highly critical of firearms and toolmark examination claiming that error-rates have not been well established [7]. 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. [18]. 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.

Through previous NIJ awards we completed two large VCM studies for cartridge case examination. Our first study was completed in 2016 and full details appear in our 2018 JFS paper [6]. This first study (Study A), 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. The second study (Study B), was our Virtual Comparison Microscopy Error Rate Study (VCMERS) and was designed to assess overall VCM error rates. Study B involved 107 participants (97 qualified examiners, 10 trainees) from seven countries. The test structure included forty test fire triples (two known, one unknown). To respect each examiner's time, each participant was randomly given sixteen triples to evaluate. Study B was completed in 2018 and a paper summarizing the complete results has recently been submitted for review [4].

In both studies, participants worked through a training tutorial containing figures and step-by-step instructions on how to use each function of the VCM software. In addition to submitting their conclusions via electronic worksheet, particpants were asked to highlight (using the mouse as a paintbrush) any individual marks used as the basis of their conclusion. We produced a series of color *Annotation Image Maps* to illustrate the regions of similarity and dissimilarity identified by the participants. 'Combined' maps show a density map of annotations for a single cartridge case by combining the annotations from multiple individuals (*e.g.*, Figure 4). Regions of the surface that were not annotated by any examiner are uncolored; annotated regions appear in color. The colors range from blue to red and indicate the percent of annotations for the specified cartridge case that had the area marked (blue indicates that few participants annotated the region, red indicates that most participants annotated the region) - see the color scale on the right side of Figure 4. 'Individual' maps can also be created to show the parts of a single cartridge case surface marked by a single participant (not shown due to space limitations). Participants can annotate the surface as being either 'similar' or 'different'; therefore, we have Similarity Maps and Difference Maps each of which contain only marks of the respective type. Note that scans are displayed in their canonical orientations for the annotation image maps; however, the scans were initially presented to participants in a random orientation (as scanned).

Study A: There were no errors among the 368 results submitted by qualified examiners on the two test sets included in Study A. The annotation maps indicate that most participants utilized the aperture shear and about half utilized some scattered, yet consistent, breech-face impression filing marks for their identifications. 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.

Study B: Analysis of the results of Study B were broken down by region. Among the 76 US and Canadian qualified examiners there were three errors among 1216 results. This error rate of 0.2% compares favorably to the error rates reported for traditional microscopy (typically 0.0% - 1.6%) [1, 12, 13]. Our error rate may be even lower as we believe one of the three errors was a typographical error in that the examiner appeared to have compared the two knowns to each other rather than the known to the unknown. The remaining two errors were made by one individual, indicating that examiner training may be more culpable than the 3D technology. In fact, we know that all other examiners who saw the same test sets as this individual made correct conclusions which suggests 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 has been reported in most previous error rate studies. Overall, the results of Study B are consistent with those obtained by the FBI and RCMP in their own internal validation studies. That is, that VCM offers a more accurate method for determining source attribution than traditional light comparison microscopy. Three sample annotation maps from Study B are shown in Figure 4. Once again, the study demonstrated that the visualization tools were generally easy to learn and that the annotation mode provides valuable insights into the decision process.

2 Project Design

The one year project included two 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 resolution study (Aim 1). We named the study the Virtual Comparison Microscopy Topography Resolution Study and will refer to it by the acronym VCMTRS 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. Aim 1 was the primary aim of the proposal and its results will constitute the majority of this document. In the smaller Aim 2, we continued our interaction with the forensic community through two deployments to state police crime labs. During these deployments we trained local examiners providing them exposure and hands-on practice with emerging 3D VCM technology.

Our previous VCM work represent important first steps towards full validation of Cadre's VCM for the examination of cartridge cases. The work described here supports these efforts by investigating resolution requirements for successful VCM. VCM utilizes 3D topographies which represent the measured surface as a grid of sample points. For each (x, y) position there is a measured z height. Sampling resolution (studied here) is one of the many ways of defining resolution. Sampling resolution refers to the pixel size or the spacing between (x, y) pixels. In our previous work we measured surfaces with a square pixel size of between 1.4 and 1.8μm per pixel. Sampling resolution varies from technology to technology. For example, high-end confocal microscopes have sampling resolution better than 1.0μm per pixel whereas other scanning modalities capture at pixel sizes of 3μm, 6μm, 10μm, or larger. Some companies have not publicly declared the sampling resolution of their scanners; however, as each system adopts the X3P file format, it will be possible for organizations such as NIST to evaluate the resolution of the provided scans. In theory, high-resolution scans should capture significant surface detail and allow for accurate comparison while low-resolution scans may not capture sufficient resolution for accurate comparison. It is important to note that sampling resolution, while easier to measure, does not fully describe the types of toolmarks that can be captured and visualized with a technology. That is, a system with 3μm per pixel spacing may not be able to resolve features of that size. Each system deals with phenomenon such as specularities (*i.e.*, very shiny surface points), imaging aberrations, imperfect focus, artifacts of steep slope, and effects of different surface materials in different ways. For example, our system use of a gel pad to remove surface specularities greatly improves the quality of our measurement allowing the accurate capture of fine striated detail, steep slopes, and arbitrary surface materials. The take-home message is simply that not all $3\mu m$ measurements are the same. In this study we wanted to explore the effects of decreasing resolution on an examiner's ability to reach source conclusions.

As VCM moves into the lab, each organization will adopt SOPs for their use. These SOPs will specify the requirements and conditions by which quality results can be achieved. These specifications will include scan resolution. To our knowledge there are no published studies which have investigated the effects of varying scan resolution. Our previous studies show that excellent performance can be obtained using surfaces with a lateral sampling resolution of 1.4-1.8μm/pixel on our gel-based measurement platform [6, 4]. Given that different labs are likely to purchase different scanning systems it is critically important to provide guidance regarding system resolution requirements. Low-resolution scans prevent visualization of small or fine features (see the many figures below, including Figures 9 and 10). We hypothesize that examiners will make more calls of 'inconclusive' or 'unsuitable for comparison' when examining low resolution scans. In this report we will describe the VCMTR Study designed to examine the effects of sampling resolution on an examiner's ability to reach same source conclusions. The study used a similar VCM testing platform as our previous successful studies. Test set design balanced the need to include the types of marks encountered in casework with the desire to keep the study a manageable size. Test sets and participant results were distributed and collected electronically. Results were summarized as presented below.

3 Materials and Methods

In this section we describe the general approach for VCMTR Study. In the Results and Analysis section we describe and interpret the obtained results.

3.1 VCM Testing Platform

We made use of our VCM testing software and our Nexus network server. Each participant was randomly assigned and provided a unique participant ID and webcode (*i.e.*, their credentials). We created a software installer to facilitate installation (and uninstallation) of the VCM software. The installer is downloaded from our website after the user is validated by entering their credentials. Double clicking the installer starts the automated installation process. When the VCM software starts, it once again 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 data transfer required for each computer. Although test sets vary slightly in size, the average size of each of the required test sets was approximately 38MB. Therefore each participant considered 640MB of testing data (plus 620MB of training data). Test sets were randomized, so while each participant is presented with test sets numbered one through seventeen 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 Nexus network server was also designed to receive participant test results.

Some of the improvements we made to our VCM testing software back when completing the VCMER Study were also useful in the current VCMTR Study. Most important of these features was the introduction of a VCM testing mode 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 (Figure 2). 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 (Figure 3). Only after successfully demonstrating proficiency with the software and visualization is the examiner allowed to proceed to the study test sets. Building from the VCMERS software, this year we made four changes to improve the software usability. First, and perhaps most importantly, the software now requires similarity or difference annotations for all test sets where an Identification or an Elimination is recorded. No annotations are required if an Inconclusive result is submitted. The reason for this change is to force the participant to "show their work" when a source conclusion is reached. In our previous VCMER Study all three observed errors had missing annotations on the unknown test fire. Second, because each test set only contains a pair of test fires we changed the software to auto-load the two scans when the participant selects a test set. Automatic loading of the two test specimens ensure that the correct surfaces are compared. Third, we improved the network communication protocol by which the VCM software communicates with our back-end software. This

[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).

improvement is designed to reduce communication issues that may occur for labs with a slow network connection. The software now attempts to reestablish broken connections and resume any in-progress file transfers. Finally, we improved the similarity and difference annotation tool. The mouse pointer is now a circle of size equal to the annotation brush and the annotation rendering is faster leading to less "jerkiness" in the annotation process. Based on these improvements and those implemented for the previous VCMER Study, we believe that the software should run on almost all Windows machines released after 2014. This is supported by the fact that only two people withdrew from the study due to computer compatibility issues.

After the user completes analysis of all seventeen 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.

3.2 Test Set Design

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). Over the past few years our interns have scanned over 8000 cartridge cases from more than 3000 different firearms. In collaboration with two firearms examiners, Todd Weller (Weller Forensics, Burlingame, California) and Zachary Carr (Johnson County Sheriff's Office Criminalistics Lab, Kansas), we considered test fires from over two 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 expected fewer inconclusive results for low complexity scans and more inconclusives for high complexity scans.

Given that participants were volunteers, we did not think it appropriate to ask individuals to evaluate a large number of test sets. We therefore based our study size on our previous VCMER Study design. In VCMERS, each participant considered sixteen test sets, each with two known and one unknown test fire. Participants needed to reach sixteen source conclusions. In VCMTRS, we created seventeen test fire pairs and asked participants to reach seventeen source conclusions. The study was designed such that it could be completed in less than six hours. Informal verbal feedback supports that we were successful in estimating maximum time requirements.

The selected cartridge cases represent a variety of tool manufacturing/finishing processes and class characteristics (Table 1). The study included both cartridge cases that are well marked and those that are minimally marked. The pairs range in complexity to represent the variability experienced in real casework. Of the seventeen selected pairs of test-fires, six (35%) were deemed low complexity, six (35%) were medium complexity, and five (30%) were high complexity. Of the sets, five (30%) have filed features, five (30%) have broached features, eight (47%) had granular features, six (35%) had partial or complete aperture shears. Several calibers were included, one was .357 Magnum, four were .38 SPL, one was .380 Auto, one was .40 S&W, three were .45 ACP, and seven were 9mm. Test fires came from eleven different firearm models. Of note, both test fires of a single questioned pair have the same class characteristics. In the case of the KNMs the two test fires were from the same firearm manufacturer (typically also the same model). Therefore it was not possible for participants to eliminate simply based on class. Images of test fires from all core test sets appear in Figures 5-8. One of the seventeen tests has strong subclass and is described in the next section.

3.3 Subclass Test Set

In addition to the sixteen pairs of core test fires we included one test fire pair with strong subclass marks. The test fires come from two .45 ACP Springfield XD45s with almost identical serial numbers. We included these test fires as a challenging example. Test fires with strong subclass marks can be encountered in casework and it is interesting to consider the effect of scan resolution on examiners ability to reach source conslusions. The two test fires are shown in Figure 11 at each of the four resolutions. It was our hypothesis that no participants would eliminate the test fires but that at low resolutions there may be some false positive errors. See Section 4.4 below for additional detail.

3.4 Scan Processing

The initial scans for each pair of test fires were processed to produce four test sets. All scans were acquired at finer than 2 micrometer per pixel (*e.g.*, typically 1.4 or 1.8 μm/px) lateral sampling. Each input scan was first downsampled via bilinear interpolation to 2.0μm/pixel which we then define as the highest resolution scan for the study. We then took this high resolution scan and performed a Gaussian blur with each of three different Gaussians (of half-width 4μm, 8μm, and 12μm) resulting in three lower resolution versions of the scans. Simply applying a Gaussian filter results in very glossy looking surfaces so we added in a bit of surface texture by adding $\pm 0.3\mu$m random noise. This tiny amount of noise made the surfaces appear less glossy without affecting the information content. In the rest of this report we refer the full resolution scan as 00 (since no Gaussian was applied) and the downsampled scans as 04, 08, and 12 corresponding to the three different Gaussian filters. Examples of the four resolutions for several different classes of test fires are shown in Figures 9 and 10. The images illustrate the loss of fine detail at lower resolutions. Note that it is extremely difficult to generate downsampled surfaces that are representative of scans acquired using different imaging technologies. For example, the gel-pad based imaging technique used by our TopMatch-3D system allows us to capture fine striated marks such as aperture shear. These shears are still present in many of our downsampled images; however, other scanning technologies often can not image fine striated marks resulting in a loss of information and a blurred area of the scan surface[2]. That is, other systems may measure a blurred aperture shear when sampling at 3μm/px despite the fact that when our system images at this same lateral sampling interval we are able to measure fine aperture shear detail. Thus, it's not possible to directly compare the scans used in our study to those obtained by other technologies. What can be concluded is the increase in inconclusives as scan resolution drops. See the results below for more details.

Four test sets were created from each of the seventeen pairs. The first set consisted of high-resolution scans (approximately 2.0 micrometer per pixel). The second, third, and fourth sets contained the 04, 08, and 12 down-sampled versions of the originals. The seventeen original test fire pairs thus resulted in 68 scaled test sets. Participants analyzed all seventeen pairs of test fires but each participant only analyzed each pair at one of the four (randomly selected) resolutions. Each participant therefore saw different scan pairs at different resolutions from high-resolution to low-resolution. For each test set, participants were asked both to reach a source conclusion (utilizing the 5 Point AFTE range of conclusions) and to

[2] One cause of this blurring is the effects of surface reflectance specularities (*e.g.*, shiny spots) on the metal surface. Our use of a gel-pad removes these specular effects thereby allowing the measurement of fine detail.

annotate areas of similarities and differences that were used when reaching their conclusion. In practice, the number of test responses for each test set will not be perfectly balanced. That is, some test sets will be evaluated by slightly more participants than others. This imbalance is due to the random setup process and the fact that some individuals signed up for the study (and were assigned test sets to analyze) but never completed the analysis.

3.5 Participants

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 2019 AFTE meeting in Nashville. 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 VCMTR 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.

Participant Demographic Breakdown. The 102 participants came from six different countries. The USA had 87 participants, Canada had 11 participants, and the rest of the world had 4 participants. To avoid deanonyimizing the international participants we will not include their results below. The remainder of this breakdown includes the USA and Canadian participants only. **Qualification:** 77 (79%) of the participants were self-reported to be qualified to perform independent casework, 21 (21%) were self-reported to be not qualified to perform independent casework (*e.g.*, they were trainees). The remainder of the demographics are for USA and Canadian examiners qualified to perform independent casework. **Experience:** 18 (23%) had three or fewer years of experience in firearm and toolmark examination, 27 (35%) had between three and ten years of experience, 32 (42%) had more than ten years of experience. **Hardware:** 56 (73%) utilized desktop computers while 21 (27%) used laptop computers. **VCM Experience:** 3% Use Routinely, 82% Used VCM a Few Times, 15% 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: 61% Very Confident, 36% Somewhat Confident, 3% Not Confident. **Lab Uses 5-Point Range of Conclusions:** 42% Yes, 56% No, 2% Unsure. **Participated in the 2018 VCMER Study:** 56% Yes, 40% No, 4% Unsure. **Lab Policy Allows Elimination on Individual Marks**[3]**:** 82%

[3] This is an important detail as individuals from labs which are not allowed to eliminate on individual marks will not be able

Yes.

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 3% of participants reported regular use of 3D visualization tools; therefore 97% of participants were new or relatively new to 3D VCM.

4 Data Results and Analysis

In this section we summarize the experimental results. The main focus of the study was examining inconclusive rates as a function of scan resolution. Section 4.1 describes the overall and individual test set inconclusive rates. Section 4.2 describes the use of a 5-point range of conclusions. Section 4.3 presents the three false positive results. Section 4.4 summarizes the results of the one subclass test set.

The primary group of participants whose results are most important to our study is the group of 77 qualified examiners from the US and Canada (66 from the US, 11 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 may cause them to make errors that would not be made by those who are qualified. Only four international participants completed the study this year and their submissions were not included in the analysis below.

We are finishing the creation of a website where each participant can log in and access their results. It will show each test set they saw, along with the results submitted, ground truth source information, and annotation maps. Links to this resource will be emailed to all participants.

4.1 Inconclusive Rates

In this section we summarize the inconclusive rates for all sixteen core test fire pairs. Section 4.1.1 summarizes the overall inconclusive rates and section 4.1.2 breaks the analysis down by test set.

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.

4.1.1 Overall Inconclusive Rates

The overall inconclusive rates for the core sets is shown in Figure 12. For KM the inconclusive rate increases from 12% to 17% to 25% to 40% as the resolution decreases. Although the inconclusive rate starts at a much higher value, a similar trend is seen for the KNMs. The KNM inconclusive rate increases from 71% to 75% to 91% to 97%. Remember that all test fire pairs are from the same make/model of firearm and exhibit the same class characteristics. The total number of submitted responses for KM were 255, 202, 198, and 192 at full resolution, 04, 08, and 12 resoution respectively. For KNM the number of submitted responses were 98, 76, 110, and 101 (from full resolution to lowest resolution).

4.1.2 Inconclusive Rates by Set

The trends observed in Section 4.1.1 can be broken down by test set. Figures 5-8 show the two full resolution items for each test set along with a bar chart of the submitted conclusions at each of the four resolutions. The results for each test set including their annotation maps will be described below. For each set we also compute the 'Inconclusive Slope' which is simply the slope the inconclusive vs resolution line. A positive slope indicates that inconclusives increase with decreasing resolution. A slope of say 10 means that there was an approximate 10% increase in inconclusives for each drop in our scan resolution. Because scan resolution is difficult to quantify (see discussion above) there is no strict interpretation of the Inconclusive Slope; however, test sets with larger slopes correspond to sets where there is a strong correlation between inconclusive rate and resolution. See discussion above.

Set 1: (Conclusions: Fig. 5, Annotation Maps: Fig. 13) Set 1 is a KM with strong aperture shear and parallel marks visible at each resolution. The annotation maps show a strong reliance on the aperture shear. Interestingly a small patch of breech face impression at the 2 o'clock position only becomes marked as similar at the lower resolution scans. Very few submissions were marked inconclusive. The Inconclusive slope was 4.5. All four resolutions are shown in Figure 9 (left).

Set 2: (Conclusions: Fig. 5, Annotation Maps: Fig. 14) Set 2 is a KNM that has areas of minimal marking without significant differences between the two test fires. There is a patch of similar parallel marks at the 2 o'clock position. The pair was marked inconclusive across all resolutions with 66 or 74 (89%) of the inconclusives being either Inconclusive B or C. One false positive was reported for set 2 (see Section 4.3). The Inconclusive slope was 3.7.

Set 3: (Conclusions: Fig. 5, Annotation Maps: Fig. 15) Set 3 is a KM with strong parallel and possibly broach marks. A slight trend towards increasing inconclusives at lower resolutions is seen; however, because the marks were very strong the trend is small. The Inconclusive slope was 3.1.

Set 4: (Conclusions: Fig. 5, Annotation Maps: Fig. 16) Set 4 is a KM with shallow parallel lines. A strong trend is seen of increased inconclusives as lower resolution. The Inconclusive slope was 14.1.

Set 5: (Conclusions: Fig. 6, Annotation Maps: Fig. 17) Set 5 is a KNM with shallow parallel lines. There is an interesting increase in Elimnations at the 04 resolution. This is likely random variance as the number of eliminations was 3 at full resolution and 5 at the 04 resolution. The annotation maps do not support the hypothesis that there was some similarity seen at the full resolution that was not visible on the 04 surface. There were almost no eliminations at the 08 and 12 resolutions. The Inconclusive slope was 5.7.

Set 6: (Conclusions: Fig. 6, Annotation Maps: Fig. 18) Set 6 is a minimally marked granular KM. There were 2 identifications at the full resolution scan. There were no identifications at any other resolution. Across all resolutions, the vast majority of conclusions were Inconclusive B. The Inconclusive slope was 2.6.

Set 7: (Conclusions: Fig. 6, Annotation Maps: Fig. 19) Set 7 is a KM with minimal breech face impression surface area from a Beretta firearm. The similarity map shows that the aperture shear is the main feature used in identification and that the shear is most visible on the high resolution scans. There are zero inconclusives at the top two resolutions, then 5% and 23.5% inconclusives at the lower resolutions. The Inconclusive slope was 7.6.

Set 8: (Conclusions: Fig. 6, Annotation Maps: Fig. 20) Set 8 is a KNM with some strong parallel potential subclass lines. The similarity maps show that the line (along the bottom of the scan) is marked as similar at low resolutions but not as frequently at high resolutions. A number of different marks are annotated on Item 2 at the higher resolutions, but not the lowest resolution. Among the 77 submissions for this test set, there were 2 eliminations at resolution 04, 1 elimination at resolution 08. The rest were marked inconclusive. One false positive was reported for set 8 (see Section 4.3). The Inconclusive slope was 1.5.

Set 9: (Conclusions: Fig. 7, Annotation Maps: Fig. 21) Set 9 is a KM with circular class breech face impression. Only a few small patches of the breech face impression were marked as similar at any of the resolutions. A trend is seen of increasing inconclusives at lower resolutions. The Inconclusive slope was 8.2.

Set <u>10:</u> (Conclusions: Fig. 7, Annotation Maps: Fig. 22) Set 10 is a KM with a possible broached breech face and fine aperture shear. A good amount of similiarity is recognized at the full resolution scan. Some areas of the breech face impressionare only marked similar on the lower resolution scans. A trend of increasing inconclusive rates is seen. The Inconclusive slope was 10.8.

Set <u>11:</u> (Conclusions: Fig. 7, Annotation Maps: Fig. 23) Set 11 is a KM with prominent aperture shear and granular and cross-hatch breech face impression. Similarity of the aperture shear is apparent at all resolutions; however, despite this trend the inconclusive rates for the lower two resolutions are 22.2% and 37.5% respectively. One hypothesis is that at lower resolutions, the aperture shear, while similar, did not contain enough fine striations for examiners to make the identification. The Inconclusive slope was 13.5. All four resolutions are shown in Figure 9 (right).

<u>Set 13:</u> (Conclusions: Fig. 7, Annotation Maps: Fig. 25) Set 13 is a KM with small breech face impression area with granular and cross-hatch type marks. The number of small lines marked as similar on the right side of the scan are only present in the full and 04 resolution scans. This is reflected in the inconclusive rates of 0% and 4.3% for the highest resolution scans and 15.0% and 45.5% for the lower resolution scans. The Inconclusive slope was 14.7.

<u>Set 14:</u> (Conclusions: Fig. 8, Annotation Maps: Fig. 26) Set 14 is a KM with granular breech face impression with distinct aperture shear. We see two phenomenon here that we've seen before. The aperture shear similarity is highest in the full resolution scan with very little aperture shear marked on the lower resolution scans. All four resolutions are shown in Figure 10 (left). Examiners marked more globular areas as similar on the lower resolution scans. A clear trend is seen in increasing inconclusives for decreasing resolution, from highest to lowest resolution: 0.0%, 4.3%, 14.3%, 50.0%. The Inconclusive slope was 16.0.

<u>Set 15:</u> (Conclusions: Fig. 8, Annotation Maps: Fig. 27) Set 15 is a KNM with areas of minimal granular breech face impression and well marked yet fine aperture shear. Examiners annotated differences in the aperture shear only at the highest two resolutions with minimal to no differences marked for the same areas on the lower resolution scans. Item 2 also has a defect towards the top of the scan which was marked at all resolutions. The Elimination rate dropped from highest to lowest resolution: 44.0%, 25.0%, 17.4%, 0.0%. One false positive was reported for set 15 (see Section 4.3). The Inconclusive slope was 15.2.

<u>Set 16:</u> (Conclusions: Fig. 8, Annotation Maps: Fig. 28) Set 16 is a well marked granular KM. The depths and scale of the granular marks on these test fires allowed equivalent levels of similarity to be annotated at each resolution. There was a slight dropoff in Identification rate; however it is worth noting of the 5 inconclusives reported, 4 of them were Inconclusive A. The Inconclusive slope was 6.5.

<u>**Set 17:**</u> (Conclusions: Fig. 8, Annotation Maps: Fig. 29) Set 17 is a KNM with transient parallel lines and a well marked yet fine aperture shear. The results are similar to Set 15 which also had a fine aperture shear. Examiners annotated differences in the aperture shear only at the highest resolution with minimal to no differences marked for the same areas on the lower three resolutions. All four resolutions are shown in Figure 10 (right). Item 2 has a defect towards the bottom of the scan which was more frequently marked at full resolution. The Elimination rate dropped from highest to lowest resolution: 75.0%, 31.6%, 16.7%, 10.0%. The Inconclusive slope was 21.0.

4.2 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 30) was also available through the software. Our background questionnaire indicated that only 42% of participants use a 5-point scale in their labs. The remaining individuals may differ in their interpretation and use of the 5-point scale. The use of the 5-point scale among US and Canadian qualified examiners is shown in Table 2. The structure of this study, designed to investigate the effect of scan resolution, makes it difficult to draw strong conclusions from these numbers. Note that among the sixteen core sets, eleven were KM and five were KNM. Of the comparisons called as Identification, 662 of 665 (99.5%) were indeed KM. Of the comparisons called as Elimination, 60 of 60 (100%) were indeed KNM. The use of Inconclusive-C is also as expected with 91.4% (107 of 117) of these conclusions being KNMs. 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. Approximately 63% of the comparisons labeled as Inc-A were indeed from the same firearm; however this number should be corrected for the number of KM and KNM. Eleven (69%) of the sixteen sets were KM which implies that a KNM is slightly more likely to end up with an Inconclusive A result than a KM. This result is consistent with our VCMER Study as well as 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.

4.3 False Positives

Not including the subclass example described below (Section 4.4), the vast majority of KM test sets were correctly identified as same source by study participants. There were no false negatives (KM called as Elimination) and there were three false positives (KNM called as Identification) reported qualified examiners.

Set 2: A false positive was called on the full-resolution scans of Set 2 by an examiner with between 3 and 10 years of experience using a laptop computer. The annotation map is shown in Figure 31 (top). This individual submitted 7 correct conclusions, had this false positive, and had 8 inconclusives. For Set 2, the examiner's confidence on the submitted conclusion was 5 (on a scale of 1-5 with 5 being most confident). This examiner submitted an average confidence of 5 for all test sets. No additional detail was provided in the comment field for this test set. This examiner used inconclusive on 1 of 8 test sets presented at full or 04 resolution. They used inconclusive on 7 of 8 test sets presented at the lowest two resolutions. Set 2 has areas of minimal marking without significant differences between the two test fires. Of the 77 conclusions submitted for Set 2, there were only 2 correct eliminations. There were 74 inconclusives. The regions of the scans marked as similar in this false positive include the parallel lines at the 2 o'clock position and a few start/end points of parallel lines on the left side of the scan.

Set 8: A false positive was called on the 04 resolution scans of Set 8 by an examiner with less than three years of experience using a desktop computer. The annotation map is shown in Figure 31 (middle). This individual submitted 9 correct conclusions, had this false positive, and had 6 inconclusives. For Set 8, the examiner's confidence on the submitted conclusion was 4 (out of 5). This examiner submitted an average confidence of 4.38 for all test sets. No additional detail was provided in the comment field for this test set. This examiner used inconclusive on 4 of 8 test sets presented at full or 04 resolution. They used inconclusive on 2 of 8 test sets presented at the lowest two resolutions. Set 8 has some strong parallel potential subclass lines. As described in Section 4.1.2 and shown in Figure 20 many participants indicated the strong parallel line as being similar between the two scans. Examiners were more likely to annotate this similarity on the lower resolution scans. The regions of the scans marked as similar in this false positive include this potential subclass line on the bottom half of the test fires as well as a few parallel lines at the 3 o'clock position. Of the 77 conclusions submitted for Set 8, there were only 3 correct eliminations. There were 73 inconclusives.

Set 15: The final false positive was called on the full resolution scans of Set 15 by an examiner with between 3 and 10 years experience using a desktop computer. The annotation map is shown in Figure 31 (bottom). This individual submitted 9 correct conclusions, had this false positive, and had 6 inconclusives. For test set 15, the examiner's confidence on the submitted conclusion was 3 (out of 5). This examiner submitted an average confidence of 4.44 for all test sets. No additional detail was provided in the comment field for this test set. This examiner used inconclusive on 2 of 8 test sets presented at full or 04 resolution. They used inconclusive on 4 of 8 test sets presented at the lowest two resolutions. Set 15 has areas of minimal granular breech face impression and well marked yet fine aperture shear. As described in Section 4.1.2 and shown in Figure 27 examiners marked the aperture shears as different for the full and 04 resolution scans. The regions of the scans marked as similar in this false positive include a few select stria of the aperture shear. Interestingly, the examiner marked large areas of the breech face impression as different. This includes the right side outer perimeter of the primer, an area which includes a number of parallel lines marked as different by most examiners (Fig. 27). Of the 77 conclusions submitted for Set 15, there were 18 (23%) correct eliminations. There were 58 inconclusives. On the full resolution scans (on which this false positive was made), 44% of the participants correctly eliminated the pair while 52% were inconclusive.

Although the surface annotations provide some insight, none of the three false positives in this study have clear explanations. All three were made for test sets with high complexity on which most other participants went inconclusive. Examiners who submitted the false positives made use of 'inconclusive' meaning that they could have used that conclusion here. It is interesting to note that all three errors were on the top two resolutions. One possible explanation for this is that at low resolutions examiners were likely to go inconclusive rather than identify or eliminate. Therefore, it is unlikely that there would be a false positive or negative at low resolutions. This is a good finding. The hypothesis is that low resolution scans may be less likely to induce a false positive or false negative; however, this comes at the cost of an increase in inconclusives.

4.4 Subclass Data Set

As described above, we included a KNM test set with strong subclass marks from two .45 ACP Springfield XD45 pistols. The two pistols have nearly identical serial numbers. Figure 11 shows both test fires

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

at all four resolutions. Both test fires have similarly arranged parallel lines. The annotation maps for the set (Set 12) are shown in Figure 24. The maps show very little annotated differences between the scans and a fair amount of similarity. On the full resolution scan, the similarity consists primarily of three parallel lines. On the lower resolution scans the entire left side of the breech face impression is labeled similar. In other words, it was more difficult to identify subtle differences at low resolutions leading to more area being marked as similar. At high resolution, subtle positioning differences between the lines could be identified and thus the lines were not marked as similar.

The conclusion bar chart is shown in Figure 32 (top). Note the number of participants who completed each resolution (28, 18, 6, and 25). This was due to the random nature of test set distribution and test set completion. It is unfortunate that only 6 participants received and submitted results for the 08 resolution scans. At the highest resolution, half the participants falsely identified the pair while the other half went inconclusive. At the 04 resolution level there were 83% false identifications and at the 08 resolution level all six participants falsely identified the pair. Although the trend is not likely statistically significant, there is a trend where more false identifications were made at lower resolution scans. The lowest resolution scans (12) have fewer false positives than the 04 and 08 resolutions. This might be within the statistical uncertainty (due to the number of participants) or it might reflect a phenomenon where the resolution is too low for a conclusion to be made. It is interesting to examine the levels of inconclusive submitted (Fig. 32 (bottom)). At the two highest resolutions there is a split between Inconclusive A and B whereas all inconclusives at the lowest resolution are Inconclusive A. This further supports the hypothesis that fine positioning differences and other differences in individual marks may only be visible at the highest resolutions.

The comments for Set 12 were very informative. For completeness here are all comments. The conclusion is listed along with the confidence on a 1-5 scale (with 5 being the most confident).

Full Resolution

- (Identification) (4) Used enhanced contrast
- (Inconclusive A) (3) I'm very tempted to call this an ID but with the resolution and not being able to distinctly make out all the fine features, I have to go with inconclusive. Item 2 is not as well marked as Item 1.
- (Inconclusive A) (5) Long gross features in agreement (I was instructed to disregard the possibility of subclass and consider them as individual), however the finer features are not well marked on Item 1 (better marked on Item 2). I would conclude to an inconclusive unless I can obtain more tests.

- (Inconclusive A) (5) Correspondence of class characteristics and coarse individual detail.
- (Inconclusive A) (4) at first glance I believed this was an ID but after filling in the annotations I realized there weren't as many similarities as I originally thought
- (Inconclusive B) (5) Minimal coarse agreement. No clear area of individual agreement
- (Inconclusive B) (4) Some agreement and disagreement of individual characteristics on the breech-face.
- (Inconclusive B) (1) Not sure what is bothering me with this one, feels like I want to look at other marks (FPI, chamber etc.) to decide
- (Inconclusive B) (4) Possible subclass in BFM.

04 Resolution

- (Identification) (2) resolution too poor to feel high confidence in common features. presuming no subclass. differences seem to be related to how well reproduced are the BF marks. too many points of congruence to be happenstance
- (Identification) (4) I would also like a second well reproducing area (e.g. fpi or ch mks).
- (Inconclusive A) (5) Areas of agreement in the gross breech face marks, however, some areas of disagreement. Due to the potential for subclass in the BFM, I would need to examine other areas of the cartridge cases for areas of agreement to reach an identification conclusion.
- (Inconclusive B) (3) While there are some vague agreement (and disagreement) of horizontal potential subclass marks, the lack of resolution makes confident identification or elimination not reasonable.

08 Resolution

- (Identification) (4) The focus on these images is borderline–not sure if I would proceed using an optical comparison microscope under these conditions.
- (Identification) (3) I would like to see better resolution.

12 Resolution

- (Identification) (2) Appears to be lots of good agreement, but individual characteristics are very gross and blurry, not a lot of fine detail.
- (Identification) (3) Images are lower quality
- (Identification) (4) potential subclass
- (Identification) (5) The images were out of focus, but I was confident they were fired from the same firearm.
- (Identification) (2) resolution not good

- (Identification) (5) Item 1 vs. 2 exhibit sufficient agreement of all discernible class characteristics with significant agreement of coarse parallel impressed linear breechface mark detail, and corresponding fine granular detail scattered amongst the linear breechface marks.
- (Inconclusive A) (3) apparent sufficient individual agreement observed. I will not call it an identification due to the blurry nature of the resolution.
- (Inconclusive A) (5) Subclass influence detected; difficult to rule it out due to poor resolution (require access to finer detail to eliminate possibility of subclass).
- (Inconclusive A) (5) Potential for subclass here so went with 4 on confidence scale.
- (Inconclusive A) (5) need a little more detail to reach a more definitive conclusion
- (Inconclusive A) (4) Low resolution image limits ability to evaluate possible subclass influence
- (Inconclusive A) (4) Like other comparisons, the poor resolution makes me want to focus the images. I see agreement of gross contour in position and shape, but poor resolution make it not possible to see if there are finer striae present. In some areas see what looks like agreement, but there is less detail on one primer vs the other or there are subtle variations/differences in width. Areas of disagreement look to be from poor reproducibility or resolution, but hard to tell without changing the resolution. No counts made since fine striae are not visible.
- (Inconclusive A) (3) The image quality isn't the clearest, and while the gross lines fall in, there are few individual details that are in agreement.
- (Inconclusive A) (1) Significant agreement observed. However, resolution of image is low and observed parallel marks are gross in nature and travel from one side of cartridge case primer to the other. Cannot rule out subclass and there are not enough fine, irregular marks present to establish an identification.

Many comments mentioned the concern about subclass. The concern for subclass was even mentioned by examiners who reached an ID conclusion. This suggests that the participants might not have expected us to include a subclass example. That is, the participants felt that they might reach a different conclusion if this was actual casework instead of a research study. What is clear from this test set is that VCM makes this type of study possible, where all participants examine the exact same surface. VCM allowed us to gather insights into the decision process, allowing us to see exactly which surface features were used as the basis for the submitted conclusion. Additional studies into subclass using VCM will likely be very informative. Furthermore, subclass training using VCM may be a powerful training tool for new examiners.

4.5 Results Summary

The completed study aimed to investigate the effect of scan resolution on inconclusive rates and the examiners ability to reach source conclusions. The performance of 77 qualified examiners from the US and Canada led strong support to the hypothesis that the inconclusive rate increases as the scan resolution decreases. The annotation maps and inconclusive bar charts for Sets 4, 7, 10, 11, and 13 demonstrate the loss of ability to identify (*i.e.*, annotate) similarity in fine surface details. Several additional conclusions can be drawn:

- The detrimental effects of low scan resolution on source conclusions are particularly significant for cartridge cases with aperture shear. The annotation maps and inconclusive bar charts for KNM Sets 15 and 17 and KM Sets 7, 10, 11, and 14 clearly demonstrate this phenomenon.
- Low resolution scans cause loss of high resolution detail with the remaining lower resolution blob-like structures looking more similar. This phenomenon can be seen in the annotation maps of Sets 10, 11, and 14.
- Because lower resolution scans may start to look similar, participants may be more likely to annotate their surfaces as similar using broad strokes. The recognition that surfaces may look similar when they are blurred does not necessarily translate into false identifications. That is, examiners seem to weight the quality of the similarity with the resolution placing more confidence in similarity observed in high resolution scans than low resolution scans.
- Several participants commented on the challenge of reaching conclusions for the blurry low-resolution images. The common theme was that they would not utilize such low resolution images in casework. If the view was from a traditional light comparison microscope they would have adjusted the objectives to bring the images into focus. If the surfaces were from a low-resolution microscope not capable of measuring the target features they would not be able to use those machines for source conclusion.
- Participants noted that they made more frequent use of inconclusive than they might under normal conditions.
- Some examiners did not appear to understand that the purpose of the study was to create and use low resolution images for comparison. Several participants commented that the images appeared blurry.
- As expected, examiners found it difficult to differentiate small differences on lower resolution scans. At the same time very low resolution scans contain so little structural information that examiners often decided to reach an inconclusive decision.

- The use of the annotation maps provided significant insight into the examiner decision process. Details are described above for each test set. 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.
- Although this was not an error rate study, it is worth noting that three false positives occurred among the sixteen core sets. Details of these false positives are presented above. It's possible that all three errors made by qualified examiners would have been flagged by such a process. This is supported by the finding that none of the other participants made errors on these test sets and the annotated areas of similarity do not agree with the consensus annotations of the other examiners. The positive predictive value of an ID was 99.5% and negative predictive value of an elimination was 100%.
- The subclass dataset proved particularly challenging. The lowest error rate was seen with the highest resolution scans.
- Because the selected test fires included those of high complexity we would not expect a 100% ID rate for KM regardless of the 3D scanning or visualization platform. The 88% ID rate for full resolution scans suggests that there may be limited benefit to acquiring scans at a resolution better than the full resolution scans used in this study. This remains an open question (*i.e.*, could we get even fewer inconclusives with a higher resolution scan, or are the inconclusive scans simply going to be inconclusive). For example, would a higher resolution scan help ID the scans in Set 6?

Overall, there was a clear trend. For KMs, moving from the highest to lowest resolution scans the inconclusive rate increased from 12% to 17% to 25% to 40%. Similarly, for KNMs, moving from the highest to lowest resolution scans the inconclusive rate increased from 71% to 75% to 91% to 97%. Note that the test sets selected for this study were designed to be challenging. All pairs of test fires came from the same make/model of firearm and exhibited the same class characteristics. We therefore did not expect a high Elimination rate for the KNMs. Taken together, the use of low resolution scans may result in a loss of source attribution. Even moving between the two highest resolution scans caused an inconclusive increase of approximately 5%. Over the course of a year this is quite a large number of cases. The phenomenon leading to this decrease is typically a loss of fine detail, often in the form of parallel lines and stria on aperture shear.

4.6 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 the Michigan State Police in Grand Rapids and at the Indiana State Police in Indianapolis. At the beginning of each deployment, Ryan Lilien went down and provided a day of hands-on training to all examiners in the lab. The Michigan State Police collected several hundred scans. Both labs provided useful feedback to our development team. 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 2019 AFTE national meeting we gave three technical presentations. One presentation took place during the main technical session and was entitled "Results of the 2018 3D Virtual Comparison Microscopy Error Rate Study for Firearm Forensics". 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 California Association of Criminalists Training Seminar (Oakland, CA), the Northeast Area Firearms Training Seminar (Waterbury, VT), the National Firearms Examiner Academy (NFEA) (Gaithersburg, MD), and CSAFE (Iowa State University). A shortened version of this final report is being submitted for publication as a paper titled "Results of the 3D Virtual Comparison Microscopy Topography Resolution Study (VCMTRS) for Firearm Forensics". The above publications and presentations continue our pattern of disseminating our research results. Over the past several years, we have presented at more than 30 forensic conferences and run training sessions at sixteen local, state, and federal crime labs.

6 Summary

We successfully completed the proposed aims during the project period. In Aim 1, we completed a large VCM study entitled the Virtual Comparison Microscopy Topography Resolution Study (VCMTR

Study). The study involved over one hundred participants including 77 qualified examiners in the US and Canada. The results of the study support the proposed hypothesis that decreasing scan resolution results in increased inconclusive rates. The details of this phenomenon were presented above. The findings have potential to make great impact on the discipline. As Firearm and Toolmarks move to adopt VCM into standard casework it is important to understand the effects of scan resolution on source conclusions. The work presented here is the first of what will hopefully be many studies into this phenomenon. To complete the aim, we developed software, assembled test fires, acquired 3D surface topographies, built out a back-end server infrastructure to support data distribution and results collection, and assembled the results presented here. The results obtained in the VCMTR Study provide another pillar of support for the use our VCM hardware and software within the crime lab. In Aim 2 we continued our collaborating with local crime labs. We trained the examiners at two state police labs and collected design feedback on the use of VCM.

Appendix

Implications for Criminal Justice Policy and Practice

The specific question investigated, how inconclusive rates vary with scan resolution, is of critical importance to the Firearm and Toolmark discipline and the Criminal Justice System. Understanding how technical procedures effect the amount of extracted information, in the form of source conclusions, helps a forensic examiner objectively evaluate the presented evidence. As labs move to adopt VCM, they need to know the technical requirements necessary for producing the highest quality result. It is clear from this study that low resolution scans, particularly those that do not capture fine lines, stria, and aperture shear are likely to result in higher inconclusive rates than those that would be achieved from high resolution scans.

Through this project and our previous NIJ grant awards 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 twelve crime laboratories have had access to our 3D scanning hardware and now close to three 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

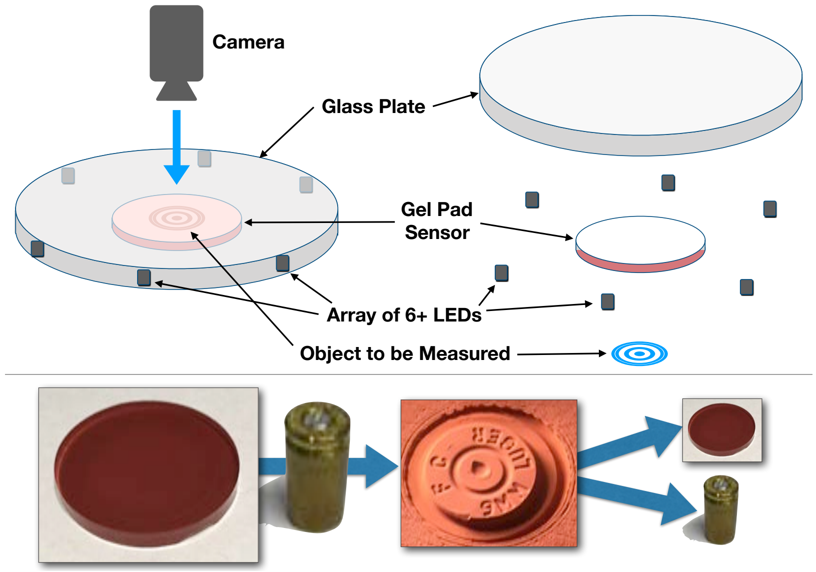


Figure 1: **GelSight Scanning Setup.** Our 3D scanning technique (GelSight) is based on the use of a silicone elastomeric pad with embedded micron-scale thick layer of pigment. (Top Row) The Gel Pad sensor is placed between a glass plate and the item being imaged. When the object to be measured is raised 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.

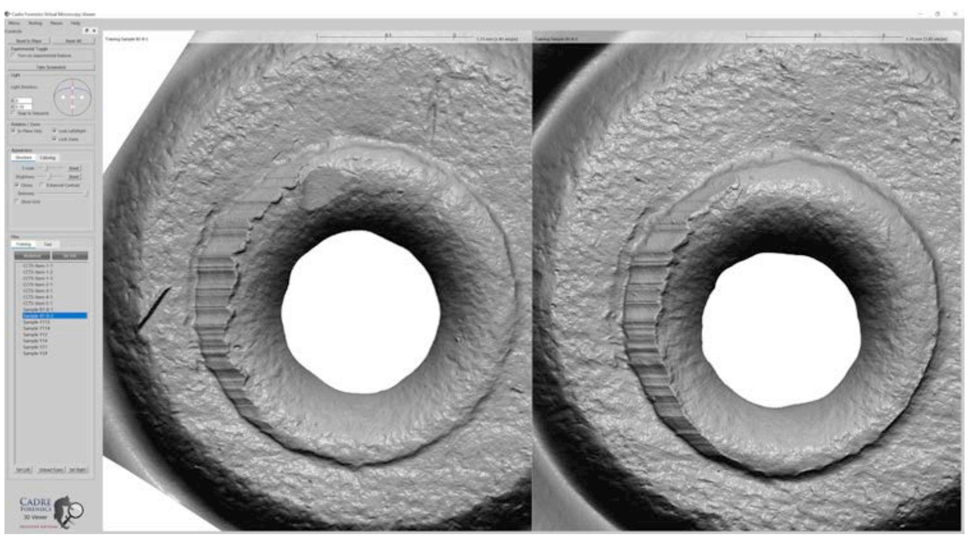


Figure 2: **Virtual Comparison Microscopy (VCM) Software.** The VCM software provides a virtual comparison scope. Examiners can adjust the virtual light position, manipulate the cartridge case orientation, position, and zoom (locked or unlocked). Training or Testing test sets can be selected on the left navigation bar. Pairs of cartridge cases can be annotated to indicate regions of similarity or difference. The VCM software can talk with our Nexus internet server to retrieve training and test data as well as submit study results.

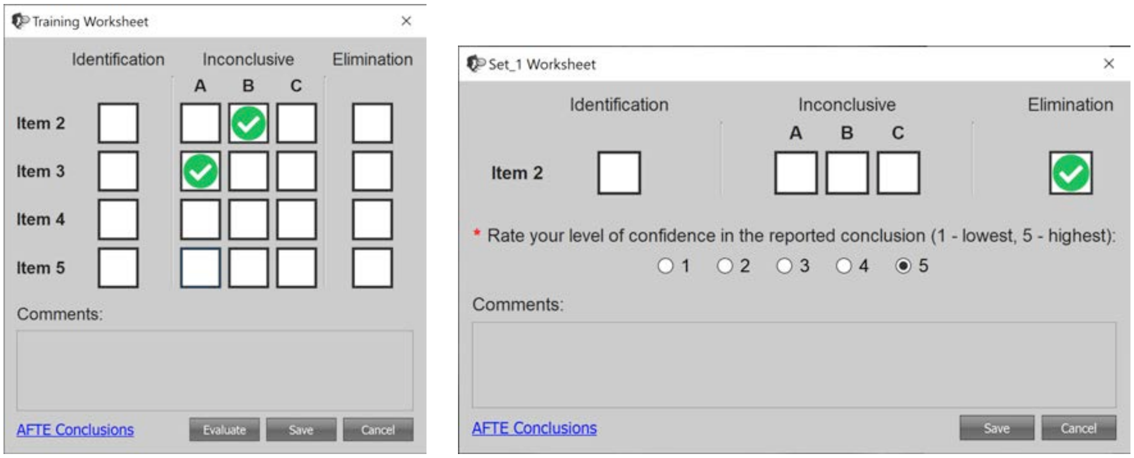


Figure 3: **Conclusion Worksheets.** (Left) The worksheet for the training scans. (Right) The worksheet for each of the test sets. Participants indicate their conclusion and confidence by clicking elements. They can also enter free text comments.

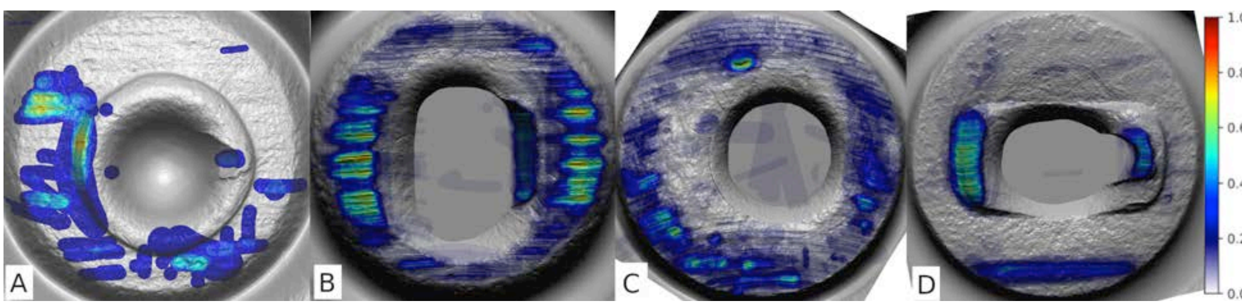


Figure 4: **Annotation Maps.** Each surface is colored by the percentage of participants annotating this item that marked the corresponding surface area. Image A is from Study A. Images B, C, and D are from Study B. Images A, B, and C are similarity maps. Image D is a difference map. These maps illustrate the consistency in the examiner thought process. Examiners did not work together to reach a consensus; each map summarizes the work of 30-40 examiners. Images such as these provide never before obtained insight into the examiner decision process. For example, it's easy to see the individualizing filing marks (B), milled marks (C), and aperture shear (D) used as the basis of each conclusion. The color legend is shown at the far right.

Test Set	KM/KNM	Caliber	Item 1	Item 2	Inc Slope
1	KM	9mm Luger	Taurus PT111 Millennium		4.5
3	KM	.38 SPL	Taurus 85		3.1
4	KM	.38 SPL	Taurus 84		14.1
6	KM	.45 ACP	H&K UMP		2.6
7	KM	9mm Luger	Beretta PX4 Storm		7.6
9	KM	.38 SPL	Smith & Wesson 12-3		8.2
10	KM	9mm Luger	Kel-Tec P11		10.8
11	KM	.40 S&W	Glock 23		13.5
13	KM	.380 ACP	Browning BDA-380		14.7
14	KM	9mm Luger	Century Arms Canik TP9V2		16.0
16	KM	9mm Luger	FN Herstal FNP-9		6.5
2	KNM	.45 ACP	Springfield XD45	Springfield XD45	3.7
5	KNM	.38 SPL	Smith & Wesson 10-7	Smith & Wesson 10-7	5.7
8	KNM	.357 Magnum	Rossi 971	Rossi 971	1.5
12	KNM	.45 ACP	Springfield XD45	Springfield XD45	-5.9
15	KNM	9mm Luger	Smith & Wesson M&P Shield	Smith & Wesson M&P 9C	15.2
17	KNM	9mm Luger	Kel-Tec P11	Kel-Tec P11	21.0

Table 1: **Test Set Summary.** Note that test set 12 is the strong subclass set. KM: Known Match, KNM: Known NonMatch. The *Inc Slope* column is the Inconclusive Slope which is described in Section 4.1.2. It quantifies the increase in inconclusive percent between each decreased resolution image. For example, the inconclusive rate for Set 10 increases by approximately 10.8% between each of the sequential decreased resolution images. That is, approximately 10% between full and 04, 10% between 04 and 08, and 10% between 08 and 12.

Resolution	Ground Truth	ID	INC-A	INC-B	INC-C	ELIM
00	**KM**	89% (227)	2% (6)	7% (19)	1% (3)	0% (0)
04	**KM**	82% (166)	6% (12)	10% (21)	1% (3)	0% (0)
08	**KM**	77% (153)	12% (24)	9% (18)	2% (3)	0% (0)
12	**KM**	60% (116)	21% (40)	18% (35)	1% (1)	0% (0)
All	**KM**	78% (662)	10% (82)	11% (93)	1% (10)	0% (0)
00	**KNM**	2% (2)	6% (6)	30% (29)	32% (31)	31% (30)
04	**KNM**	1% (1)	14% (11)	39% (30)	22% (17)	22% (17)
08	**KNM**	0% (0)	10% (11)	52% (57)	29% (32)	9% (10)
12	**KNM**	0% (0)	20% (20)	50% (51)	27% (27)	3% (3)
All	**KNM**	1% (3)	12% (48)	43% (167)	28% (107)	16% (60)

Table 2: **Use of 5-Point Scale.** Use of 5-point scale among US and Canadian qualified examiners. Percentages are listed with counts in parentheses. Of the comparisons called as Identification, 662 of 665 (99.5%) were indeed KM. Of the comparisons called as Elimination, 60 of 60 (100%) were indeed KNM. There were a total of 130 Inconclusive As, 260 Inconclusive Bs, and 117 Inconclusive Cs.

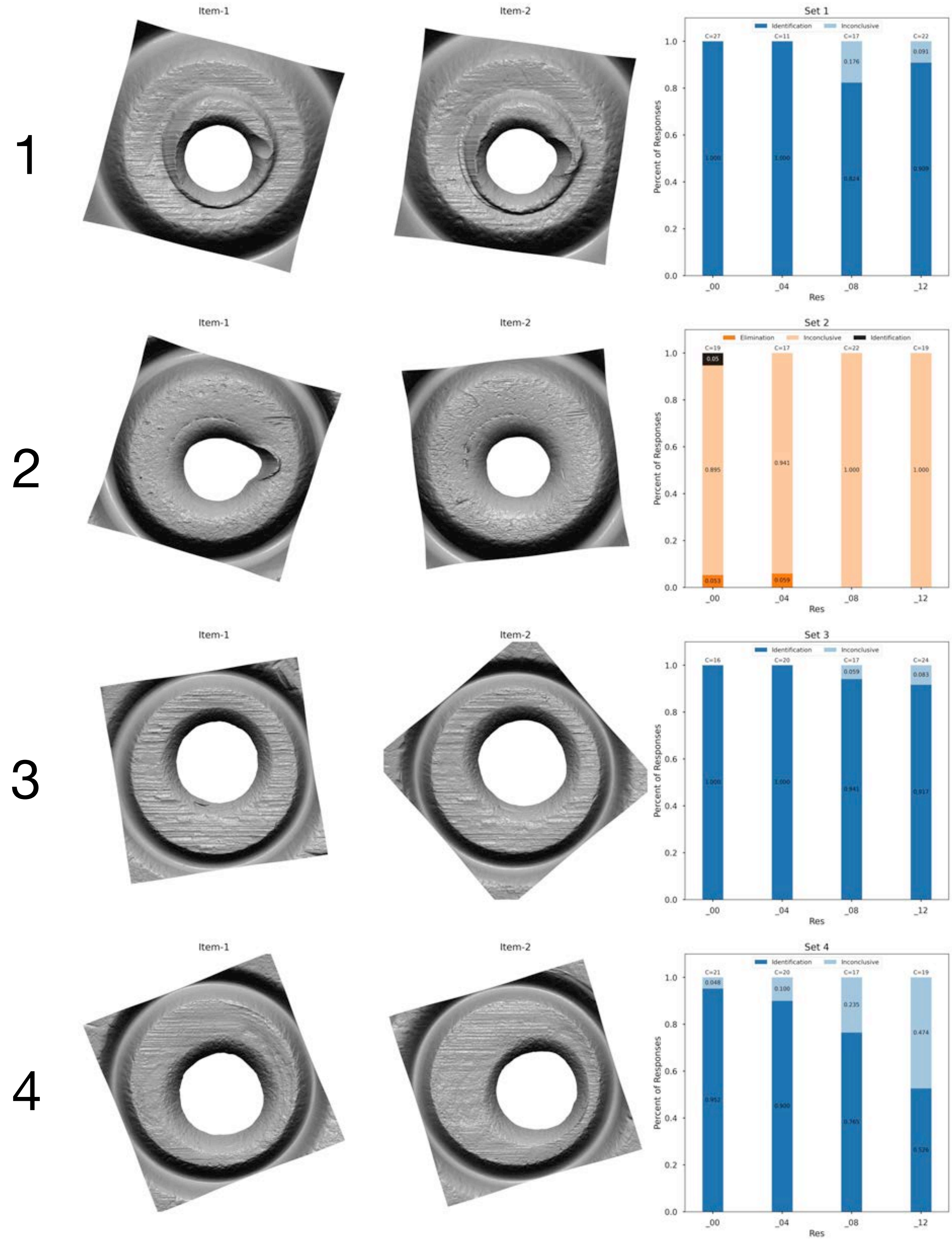


Figure 5: **Results for Test Sets 1-4.** Known Matches are shown with blue bar charts. Known Non-Matches are shown with orange bar charts. Items 1 and 2 are shown in the first two columns. Identifications (dark blue), Inconclusives (light blue and light orange), Eliminations (dark orange), False Positives (black). Participant counts are listed above each bar.

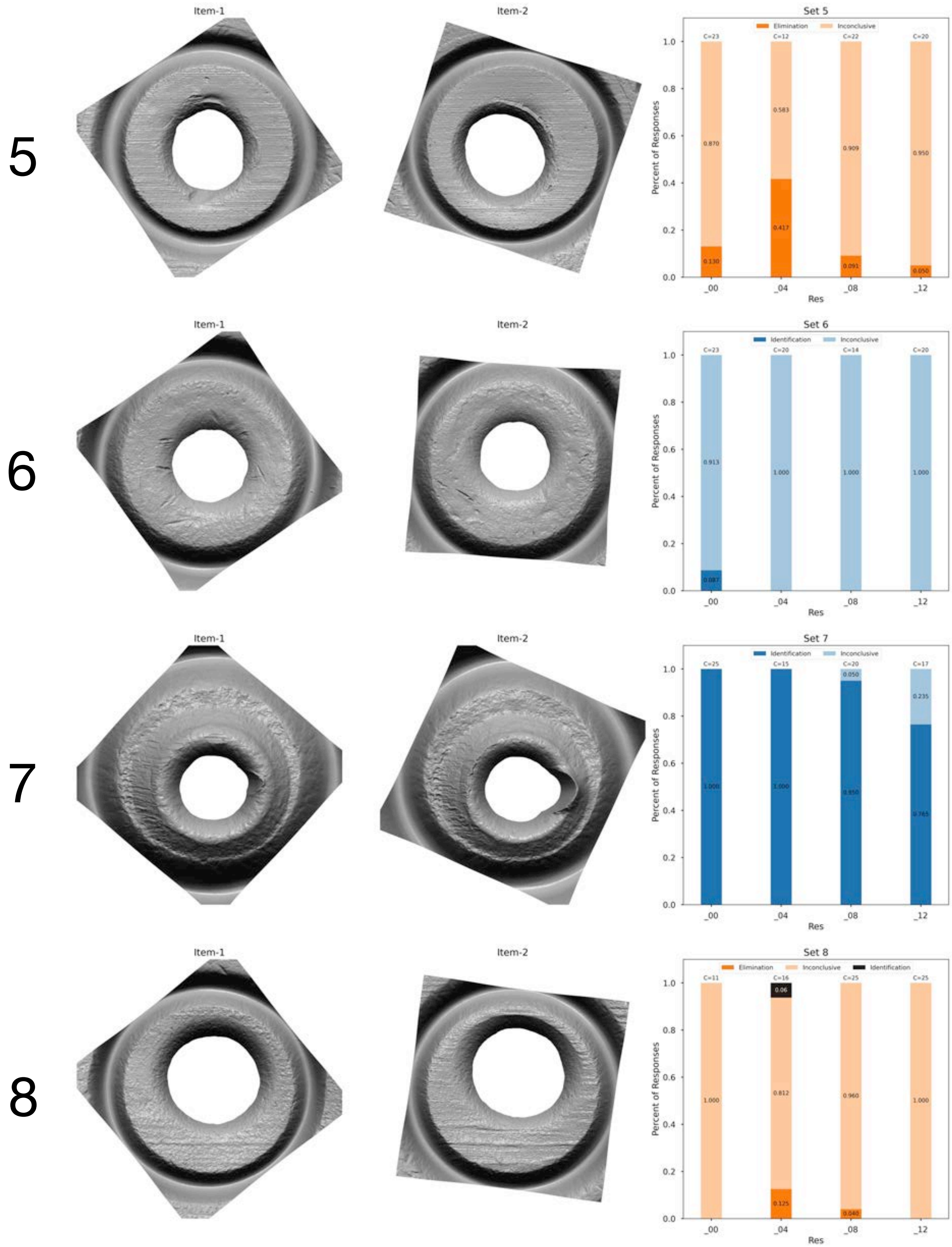


Figure 6: **Results for Test Sets 5-8.** Known Matches are shown with blue bar charts. Known Non-Matches are shown with orange bar charts. Items 1 and 2 are shown in the first two columns. Identifications (dark blue), Inconclusives (light blue and light orange), Eliminations (dark orange), False Positives (black). Participant counts are listed above each bar.

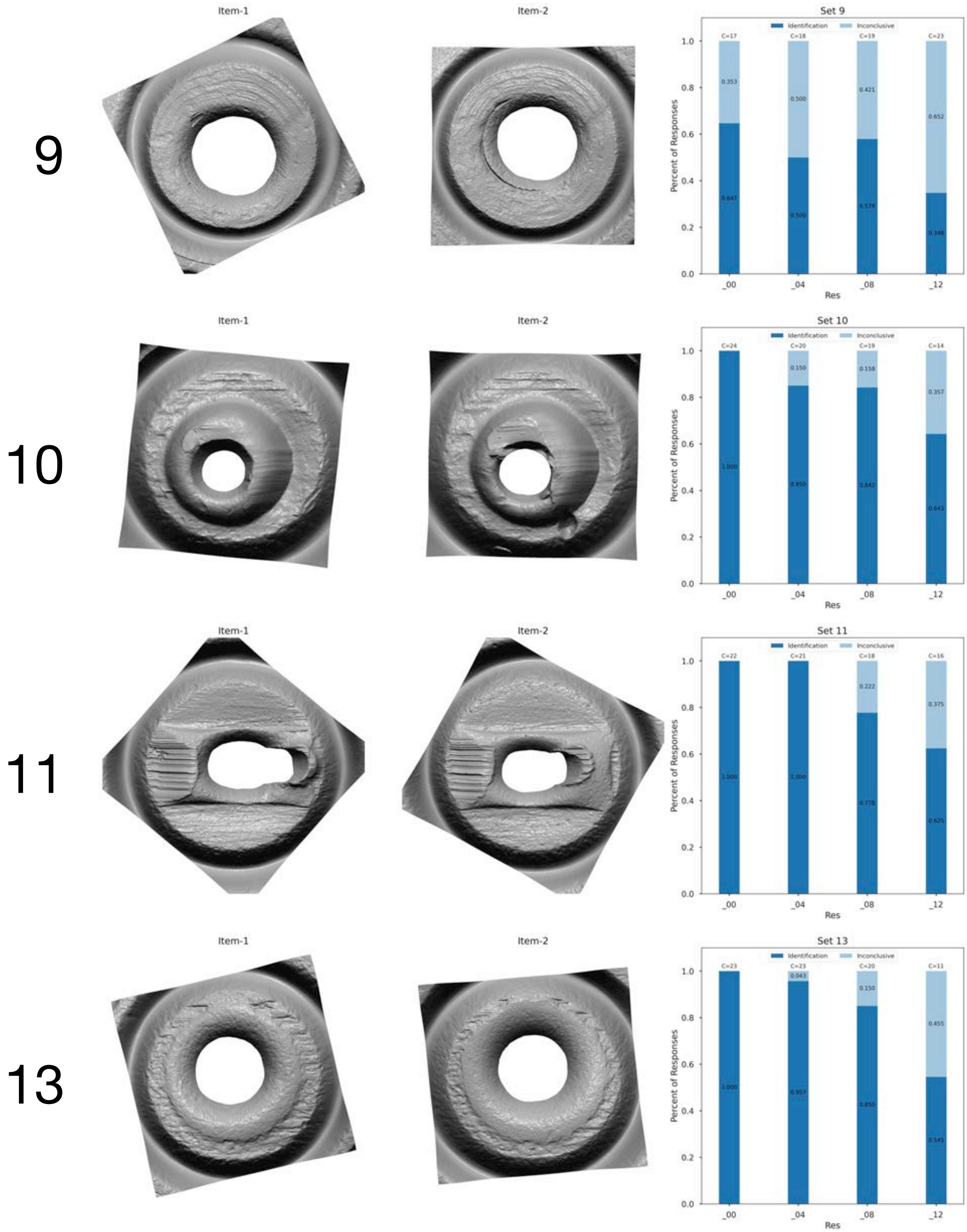


Figure 7: **Results for Test Sets 9-13.** Known Matches are shown with blue bar charts. Known Non-Matches are shown with orange bar charts. Items 1 and 2 are shown in the first two columns. Identifications (dark blue), Inconclusives (light blue). Participant counts are listed above each bar.

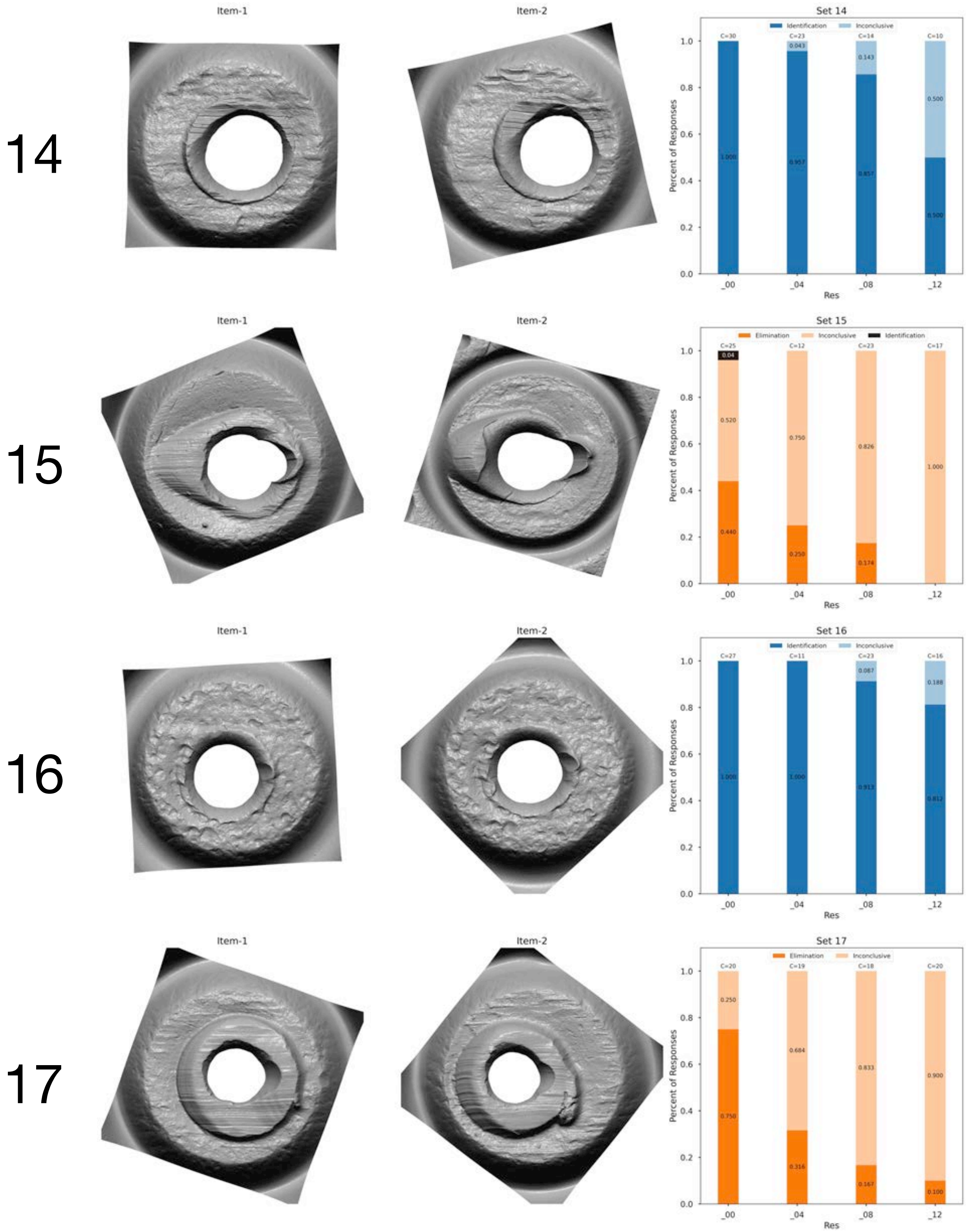


Figure 8: **Results for Test Sets 14-17.** Known Matches are shown with blue bar charts. Known Non-Matches are shown with orange bar charts. Items 1 and 2 are shown in the first two columns. Identifications (dark blue), Inconclusives (light blue and light orange), Eliminations (dark orange), False Positives (black). Participant counts are listed above each bar.

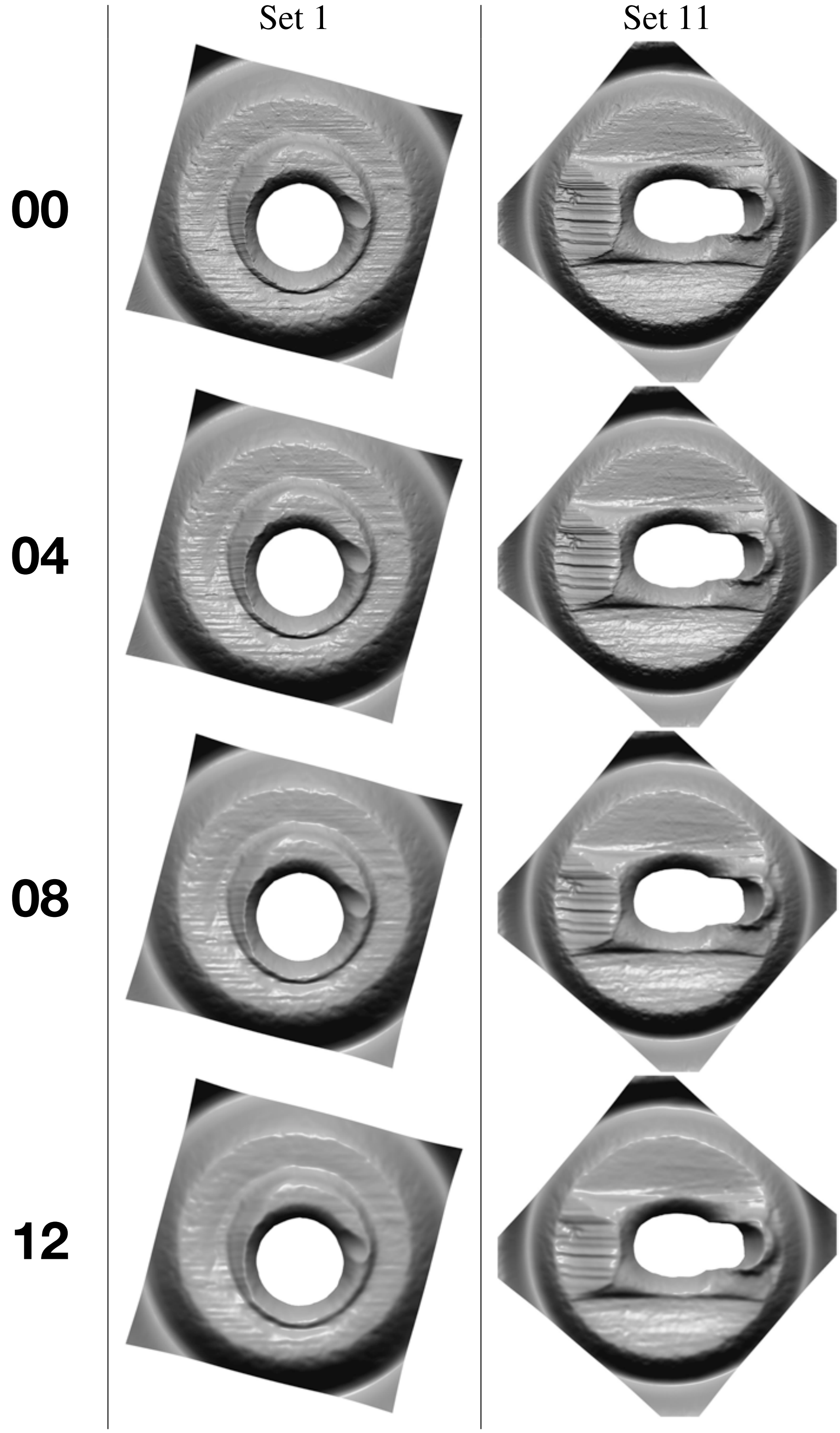


Figure 9: **Multiple Resolution Surfaces of Test Sets 1 and 11.** Full resolution at top with decreasing resolution in subsequent rows. Note that fine details observable in high resolution surfaces are not visible at lower resolution surfaces. The test fires of Set 1 were not compared to those of Set 11, the surfaces are combined in this figure to save space.

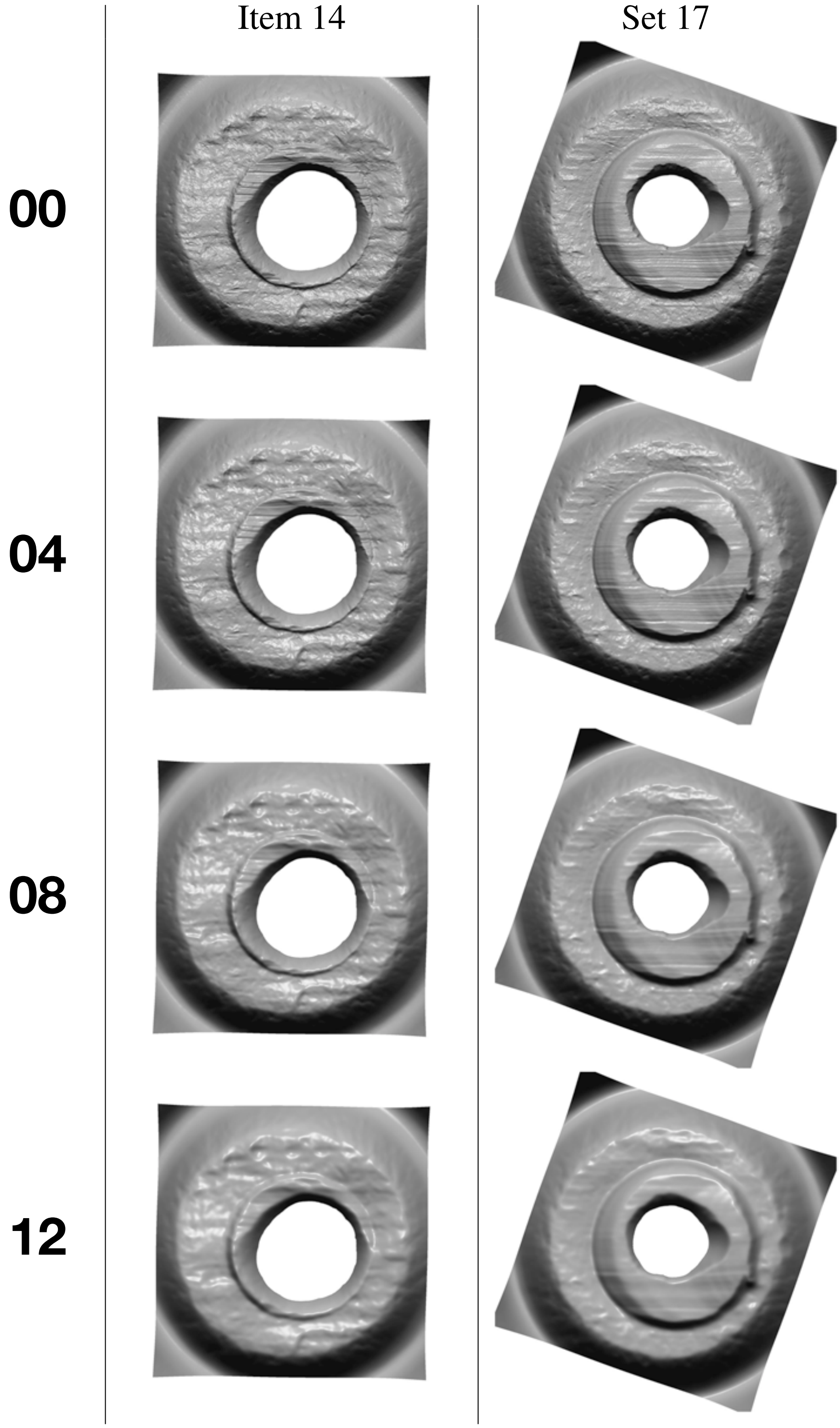


Figure 10: **Multiple Resolution Surfaces of Test Sets 14 and 17.** Full resolution at top with decreasing resolution in subsequent rows. Note that fine details observable in high resolution surfaces are not visible at lower resolution surfaces. The test fires of Set 14 were not compared to those of Set 17, the surfaces are combined in this figure to save space.

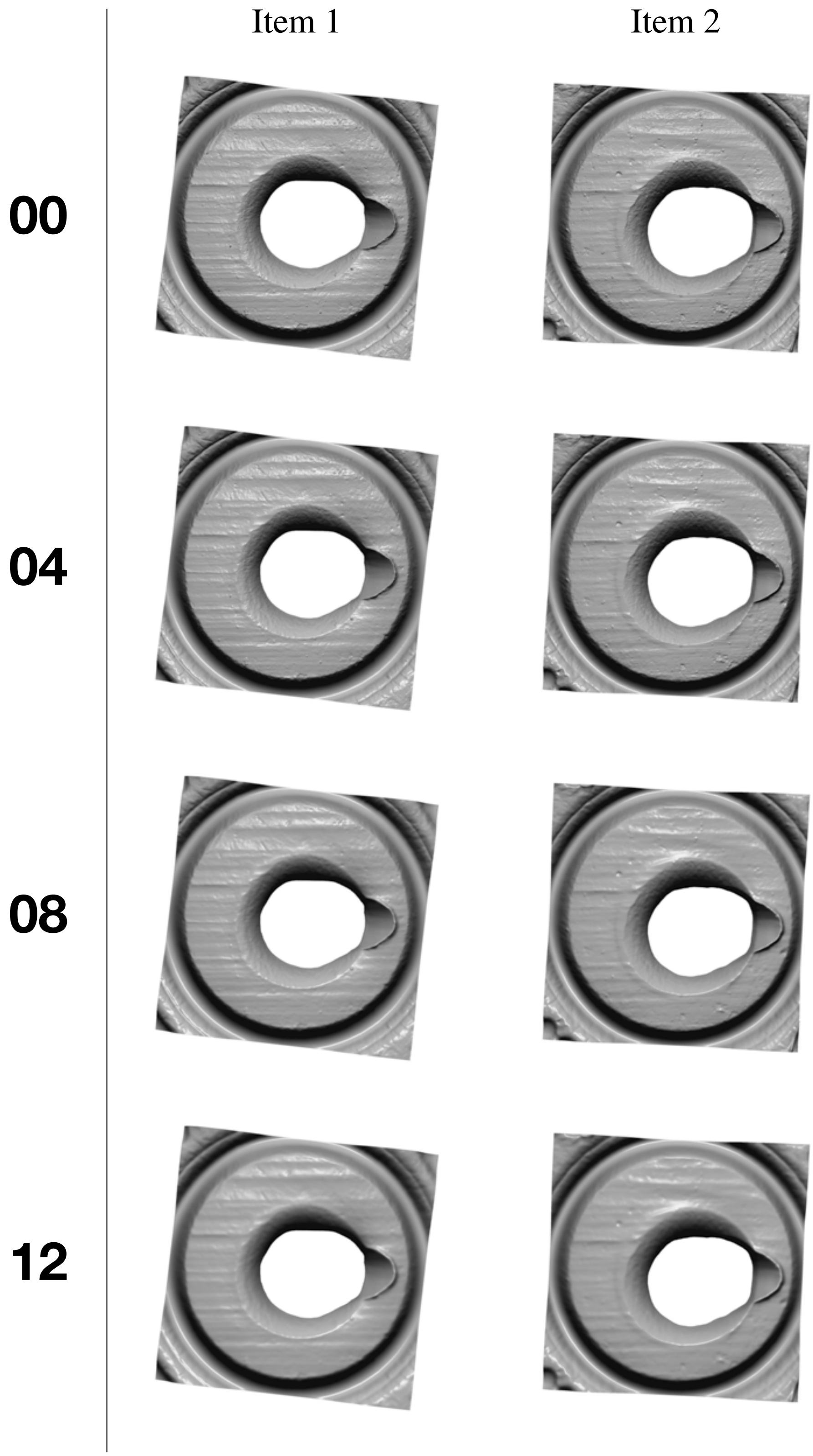


Figure 11: **Test Set 12 (Subclass).** Full resolution at top with decreasing resolution in subsequent rows.

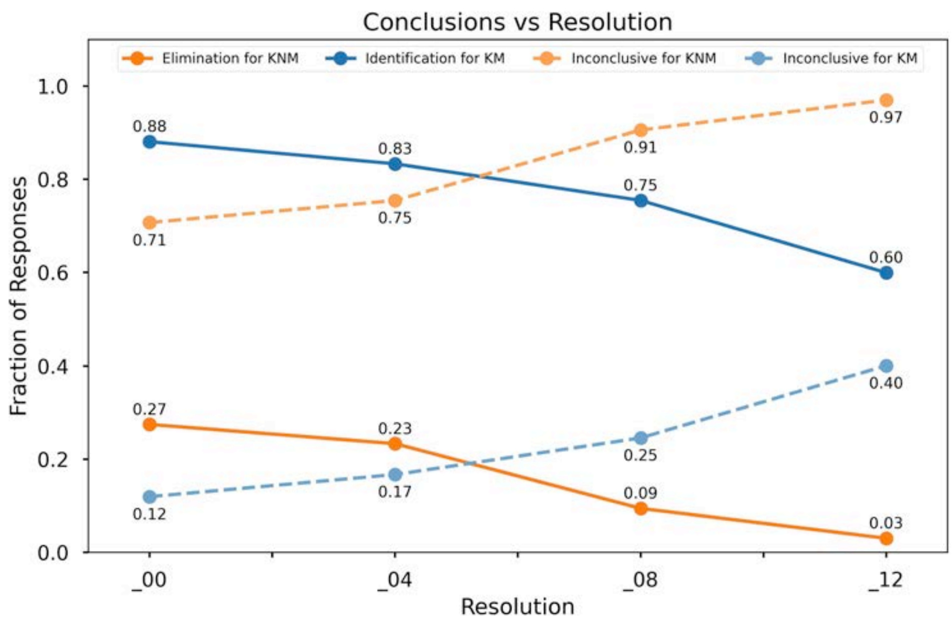


Figure 12: **Overall Inconclusive Rates.** The use of inconclusives increases as the resolution decreases for both KM and KNM. This plot includes all sixteen core test sets and does not include the subclass set.

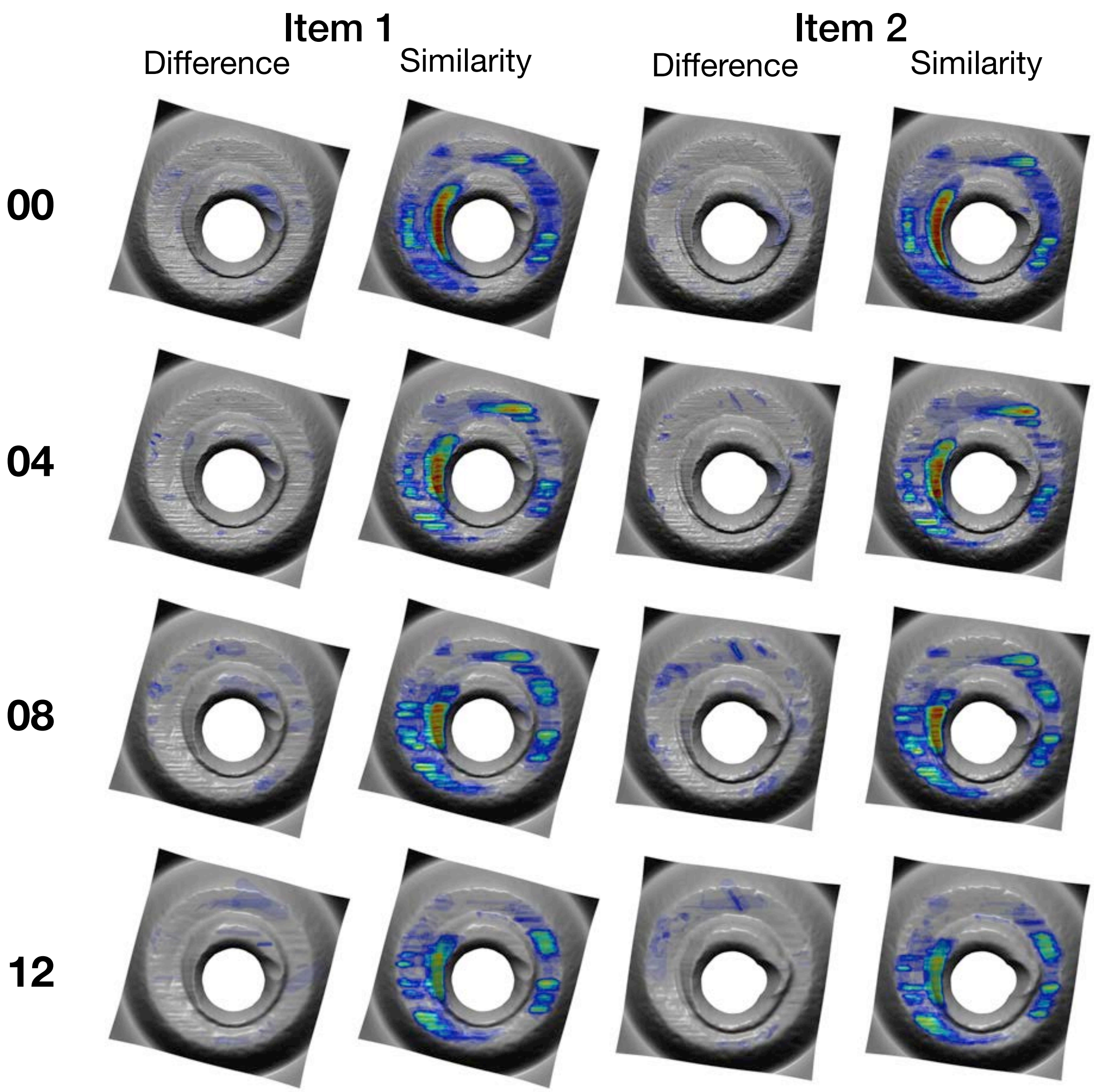


Figure 13: **Annotation Maps for Test Set 1. Known Match.** Columns from Left to Right: Item 1 Difference, Item 1 Similarity, Item 2 Difference, Item 2 Similarity. Rows from Top to Bottom: Full resolution, Resolution 04, Resolution 08, Resolution 12.

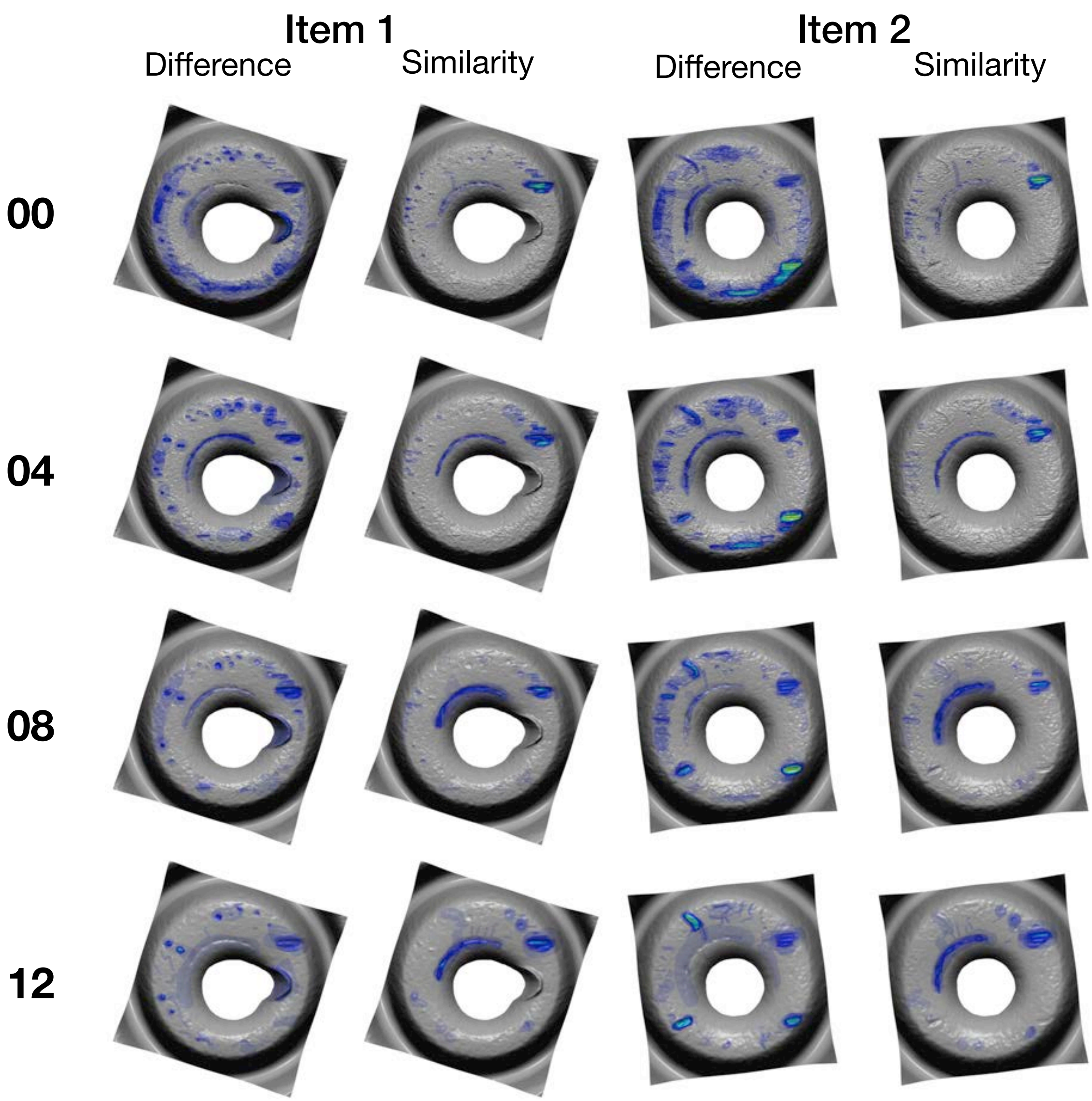


Figure 14: **Annotation Maps for Test Set 2. Known Non-Match.** Columns from Left to Right: Item 1 Difference, Item 1 Similarity, Item 2 Difference, Item 2 Similarity. Rows from Top to Bottom: Full resolution, Resolution 04, Resolution 08, Resolution 12.

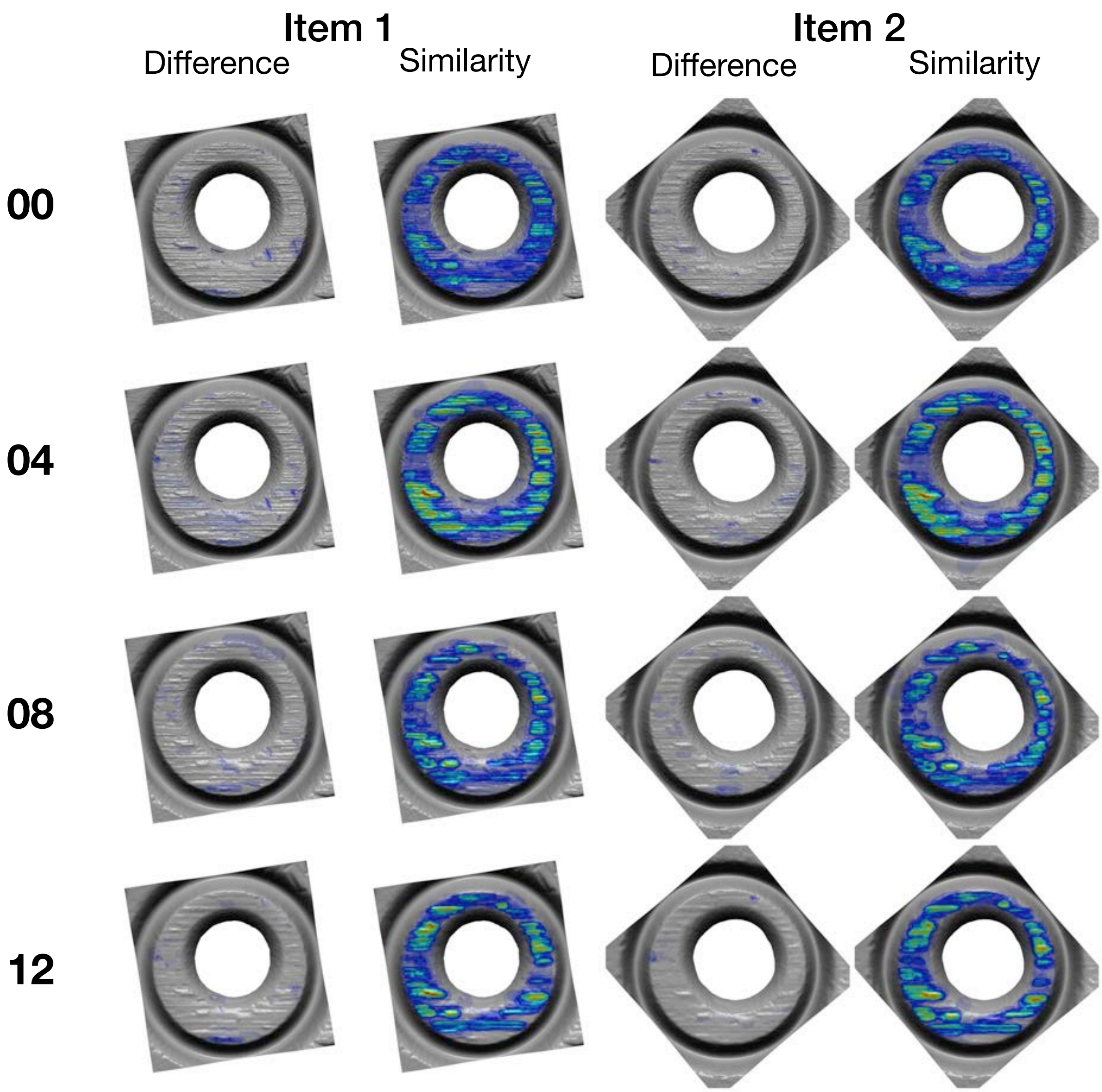


Figure 15: **Annotation Maps for Test Set 3. Known Match.** Columns from Left to Right: Item 1 Difference, Item 1 Similarity, Item 2 Difference, Item 2 Similarity. Rows from Top to Bottom: Full resolution, Resolution 04, Resolution 08, Resolution 12.

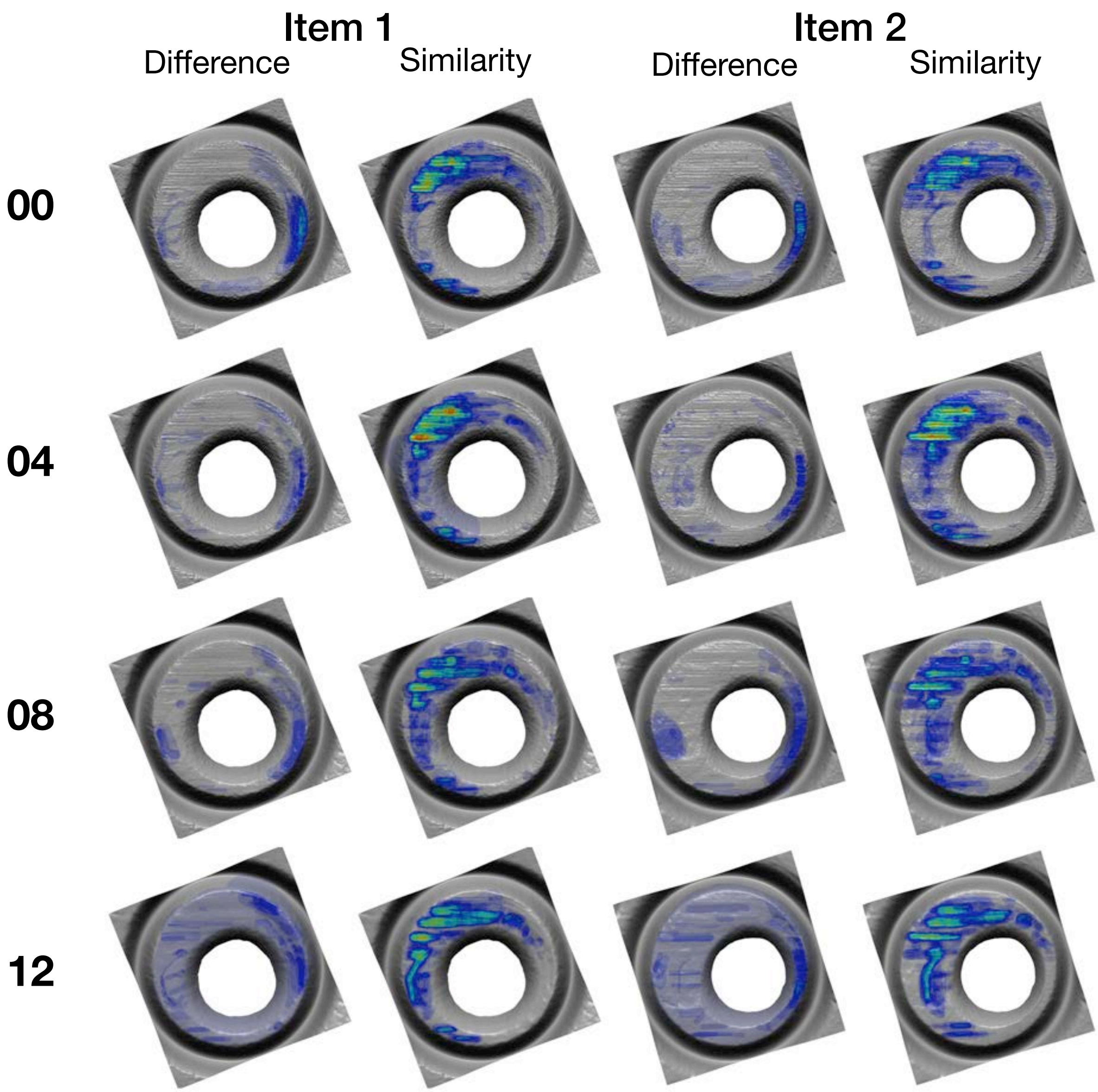


Figure 16: **Annotation Maps for Test Set 4. Known Match.** Columns from Left to Right: Item 1 Difference, Item 1 Similarity, Item 2 Difference, Item 2 Similarity. Rows from Top to Bottom: Full resolution, Resolution 04, Resolution 08, Resolution 12.

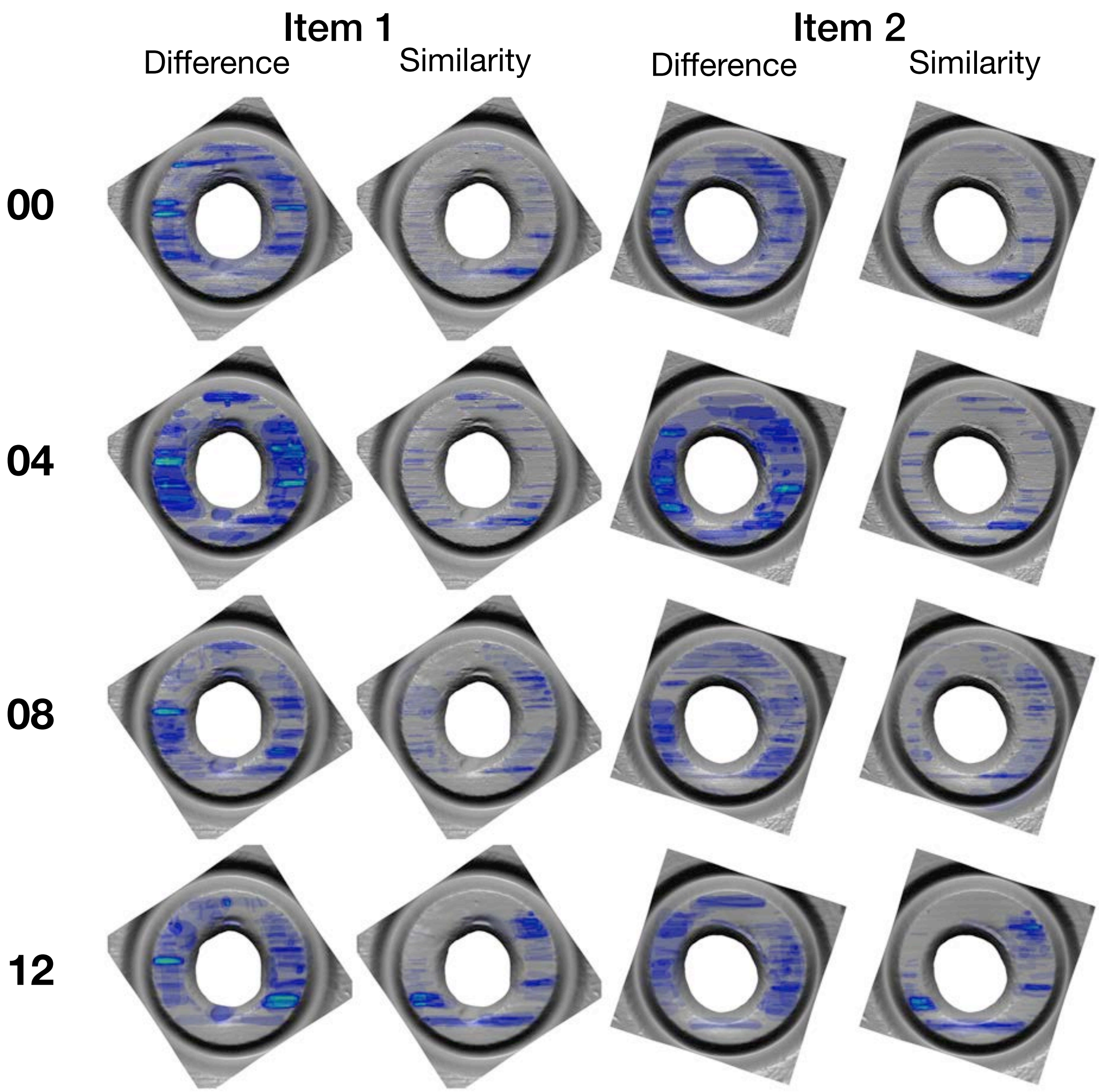


Figure 17: **Annotation Maps for Test Set 5. Known Non-Match.** Columns from Left to Right: Item 1 Difference, Item 1 Similarity, Item 2 Difference, Item 2 Similarity. Rows from Top to Bottom: Full resolution, Resolution 04, Resolution 08, Resolution 12.

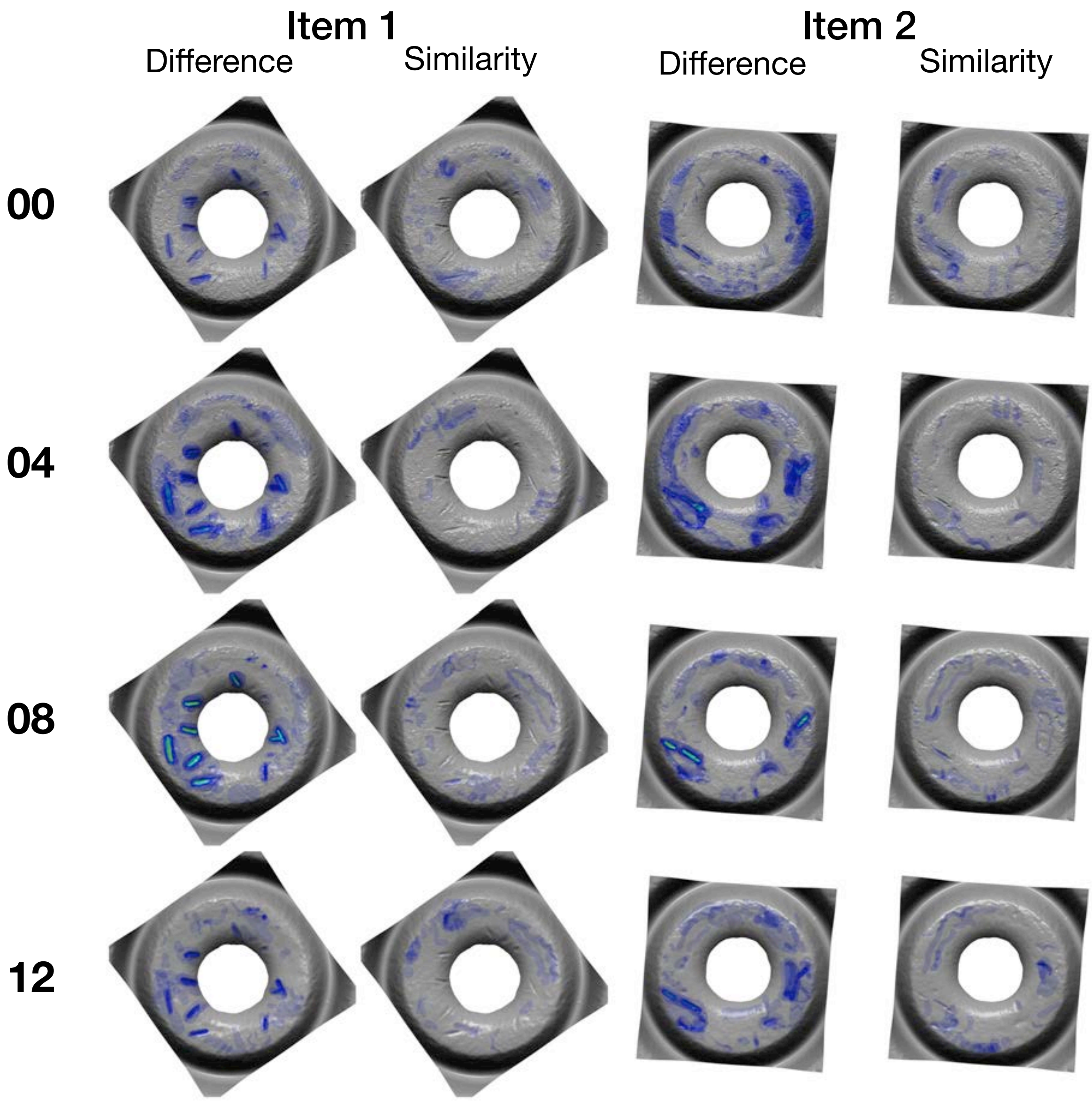


Figure 18: **Annotation Maps for Test Set 6. Known Match.** Columns from Left to Right: Item 1 Difference, Item 1 Similarity, Item 2 Difference, Item 2 Similarity. Rows from Top to Bottom: Full resolution, Resolution 04, Resolution 08, Resolution 12.

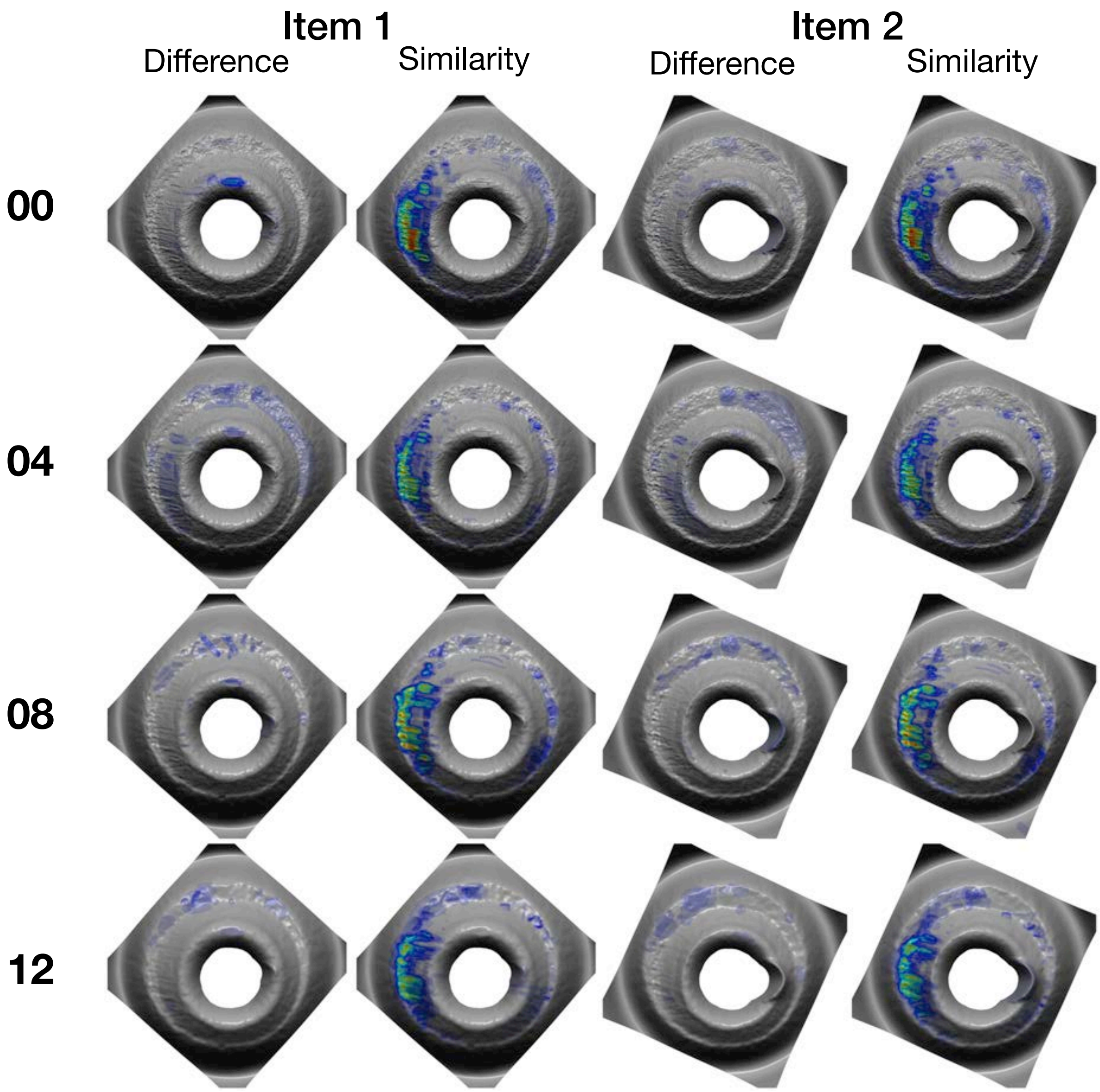


Figure 19: **Annotation Maps for Test Set 7. Known Match.** Columns from Left to Right: Item 1 Difference, Item 1 Similarity, Item 2 Difference, Item 2 Similarity. Rows from Top to Bottom: Full resolution, Resolution 04, Resolution 08, Resolution 12.

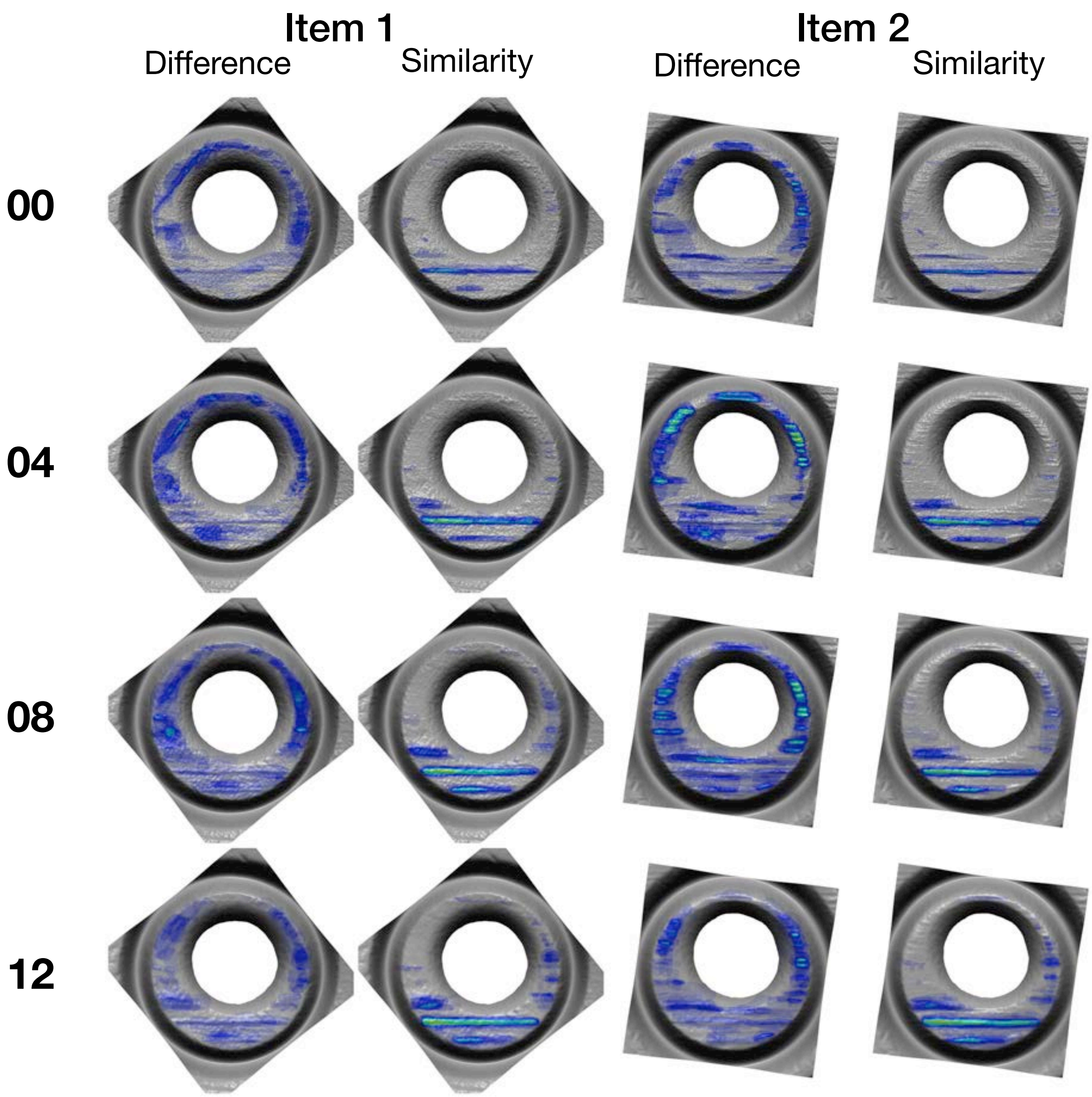


Figure 20: **Annotation Maps for Test Set 8. Known Non-Match.** Columns from Left to Right: Item 1 Difference, Item 1 Similarity, Item 2 Difference, Item 2 Similarity. Rows from Top to Bottom: Full resolution, Resolution 04, Resolution 08, Resolution 12.

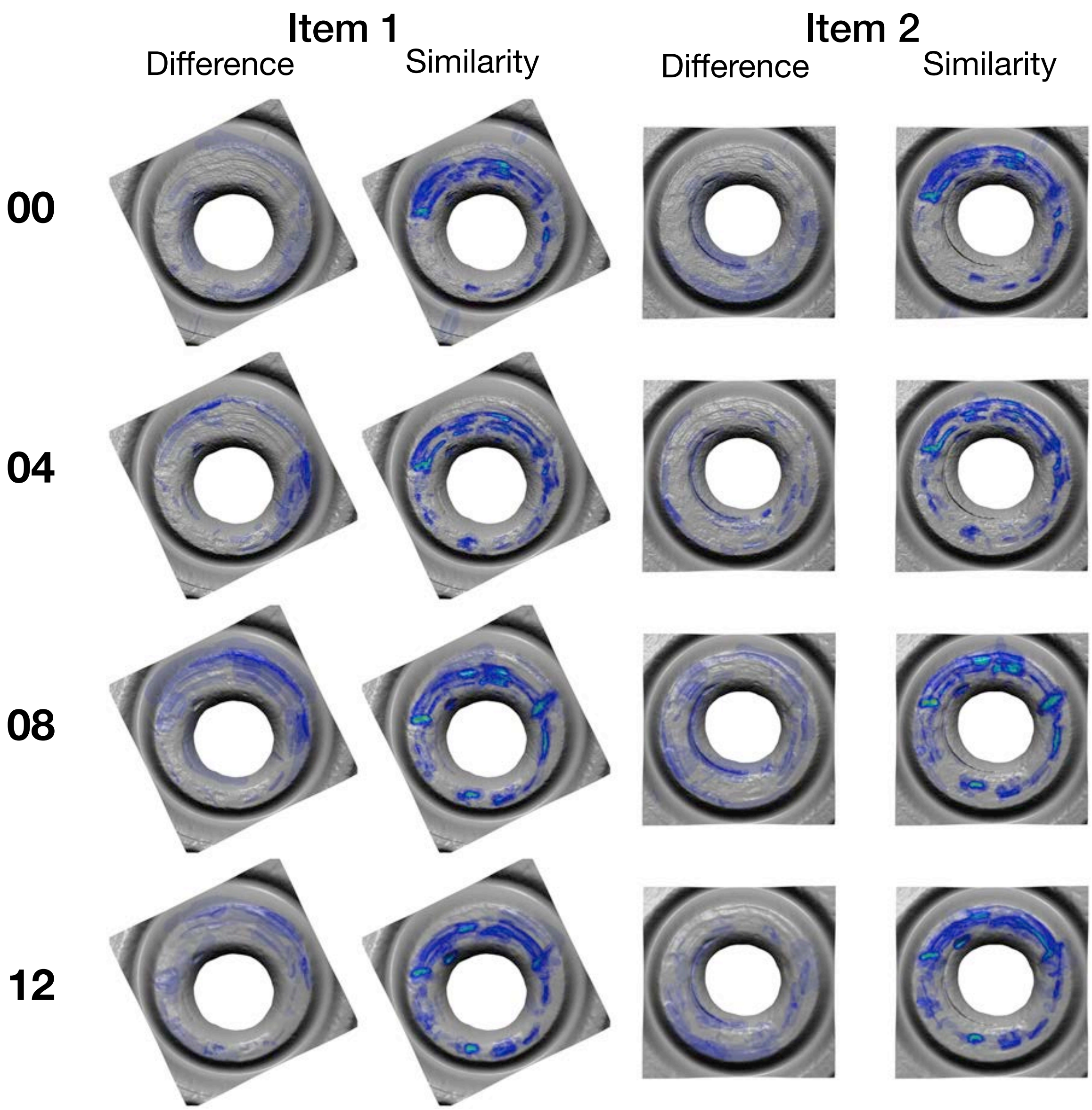


Figure 21: **Annotation Maps for Test Set 9. Known Match.** Columns from Left to Right: Item 1 Difference, Item 1 Similarity, Item 2 Difference, Item 2 Similarity. Rows from Top to Bottom: Full resolution, Resolution 04, Resolution 08, Resolution 12.

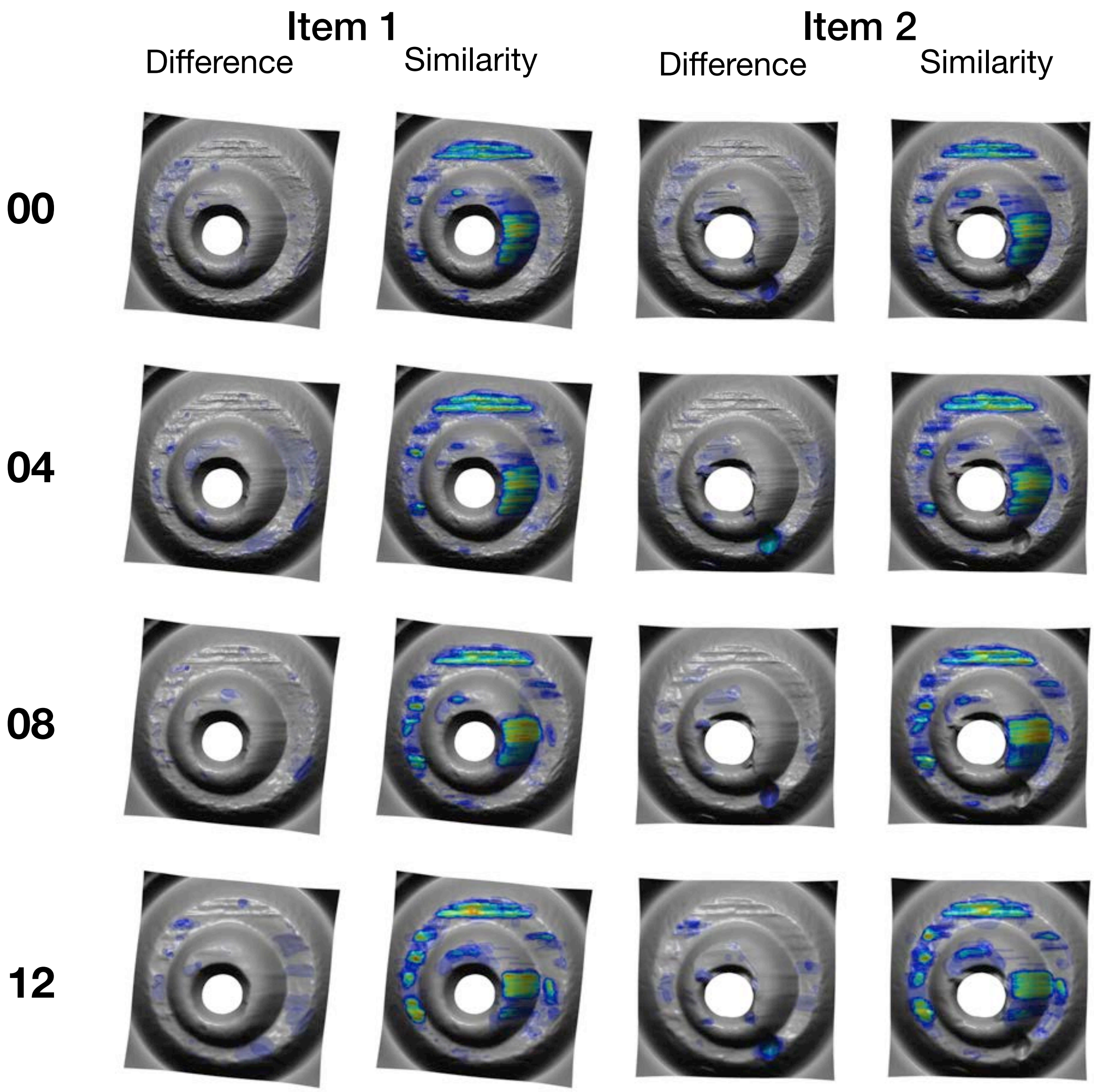


Figure 22: **Annotation Maps for Test Set 10. Known Match.** Columns from Left to Right: Item 1 Difference, Item 1 Similarity, Item 2 Difference, Item 2 Similarity. Rows from Top to Bottom: Full resolution, Resolution 04, Resolution 08, Resolution 12.

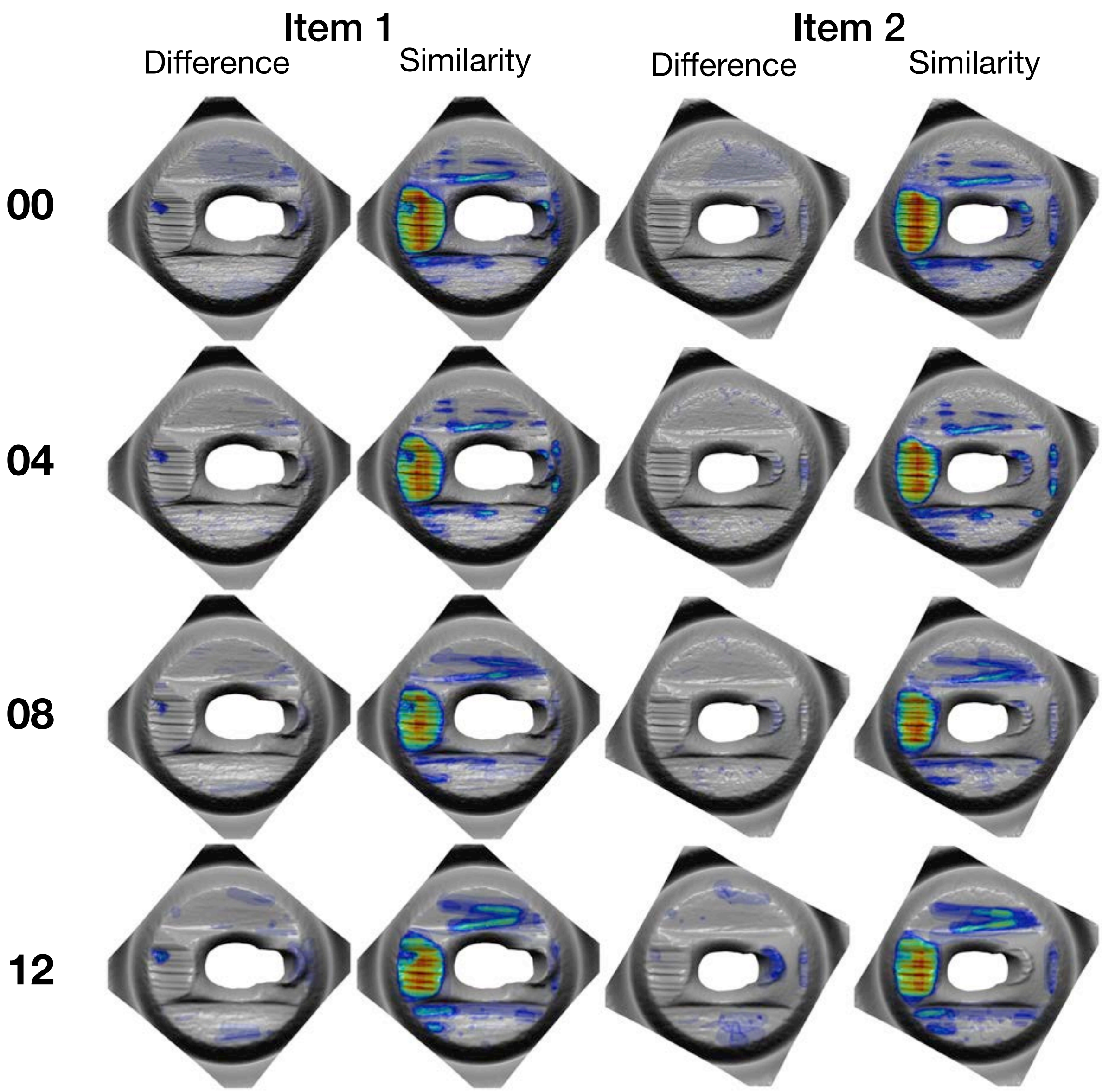


Figure 23: **Annotation Maps for Test Set 11. Known Match.** Columns from Left to Right: Item 1 Difference, Item 1 Similarity, Item 2 Difference, Item 2 Similarity. Rows from Top to Bottom: Full resolution, Resolution 04, Resolution 08, Resolution 12.

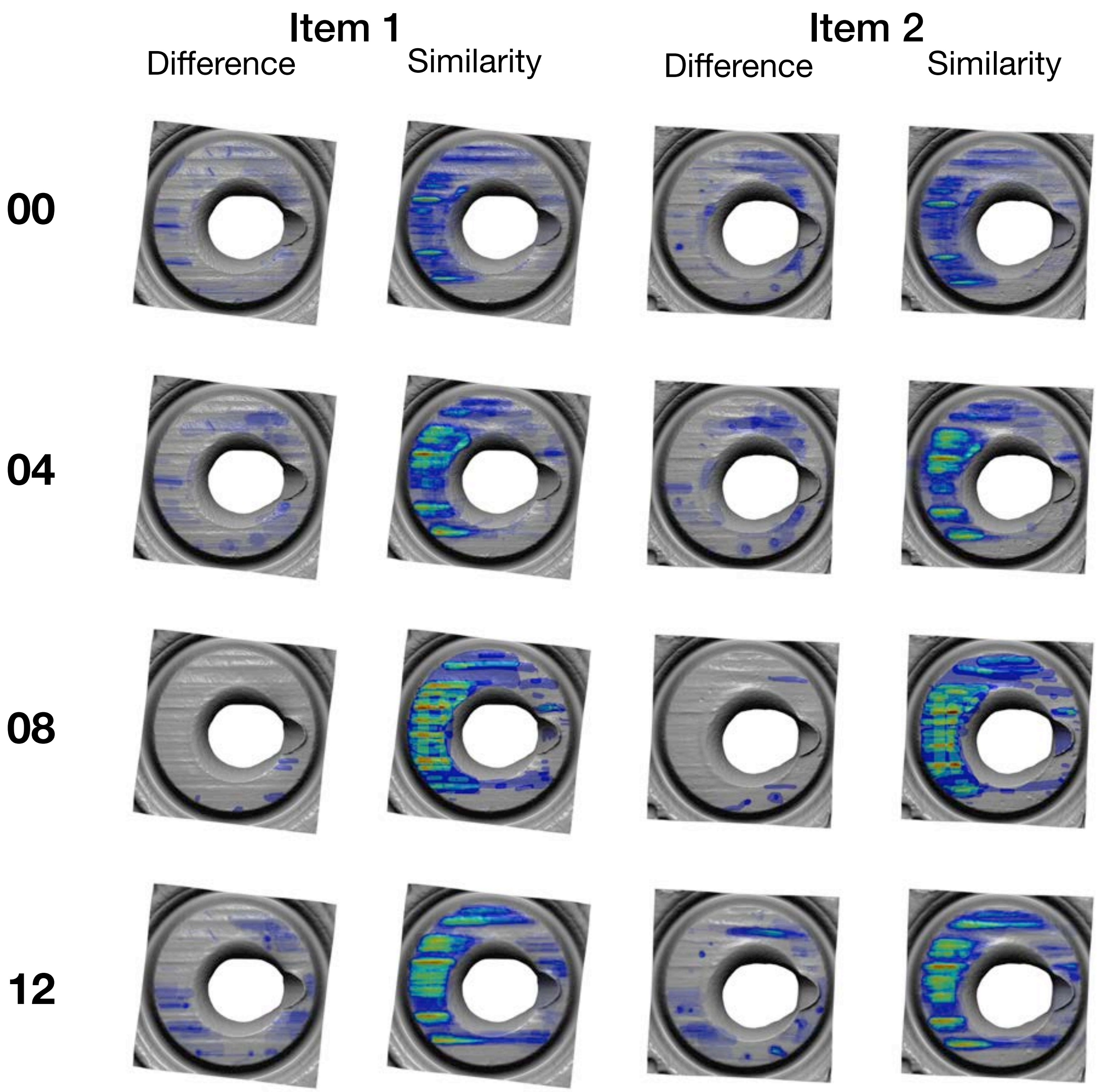


Figure 24: **Annotation Maps for Test Set 12. Known Non-Match.** Columns from Left to Right: Item 1 Difference, Item 1 Similarity, Item 2 Difference, Item 2 Similarity. Rows from Top to Bottom: Full resolution, Resolution 04, Resolution 08, Resolution 12.

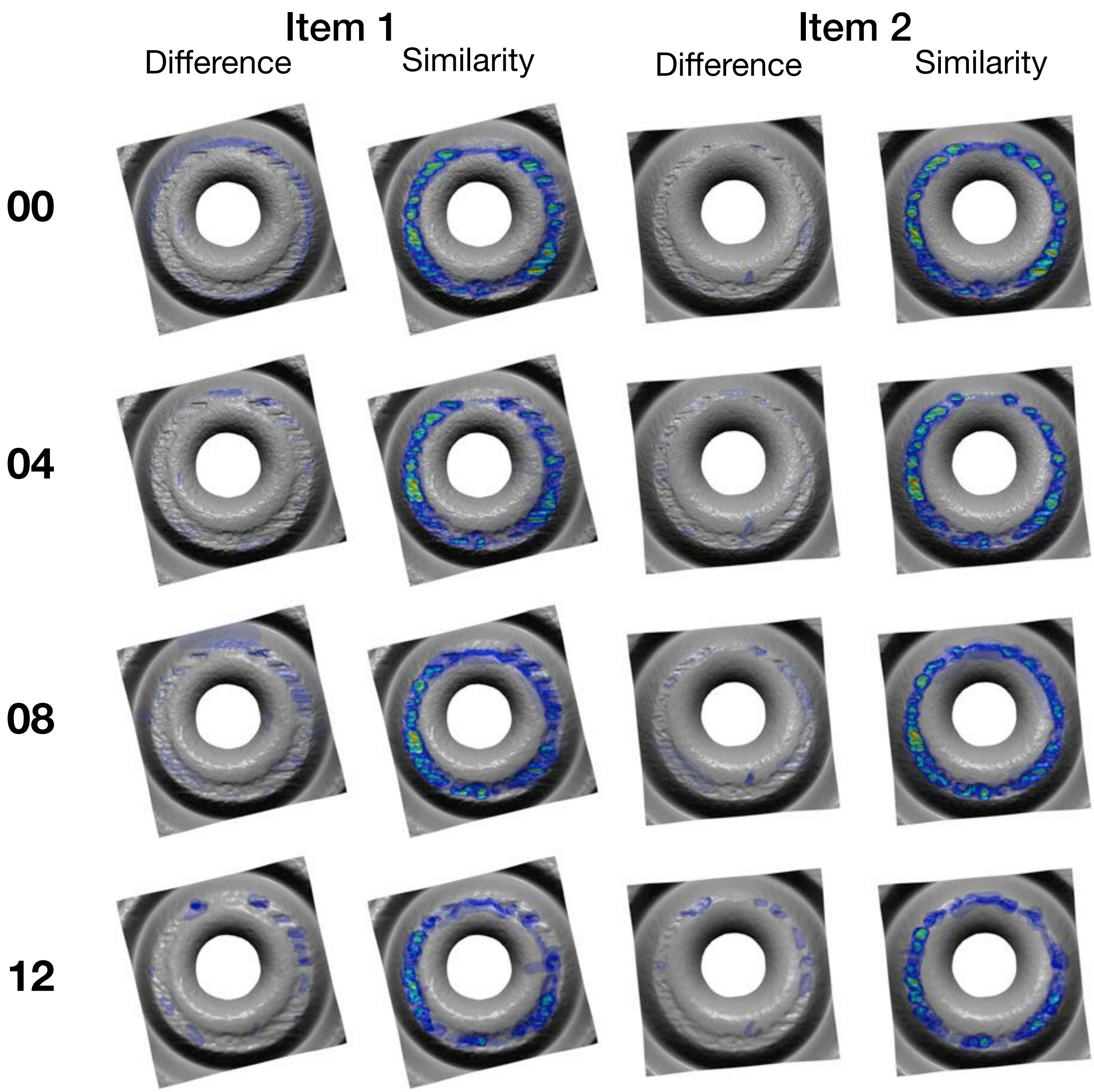


Figure 25: **Annotation Maps for Test Set 13. Known Match.** Columns from Left to Right: Item 1 Difference, Item 1 Similarity, Item 2 Difference, Item 2 Similarity. Rows from Top to Bottom: Full resolution, Resolution 04, Resolution 08, Resolution 12.

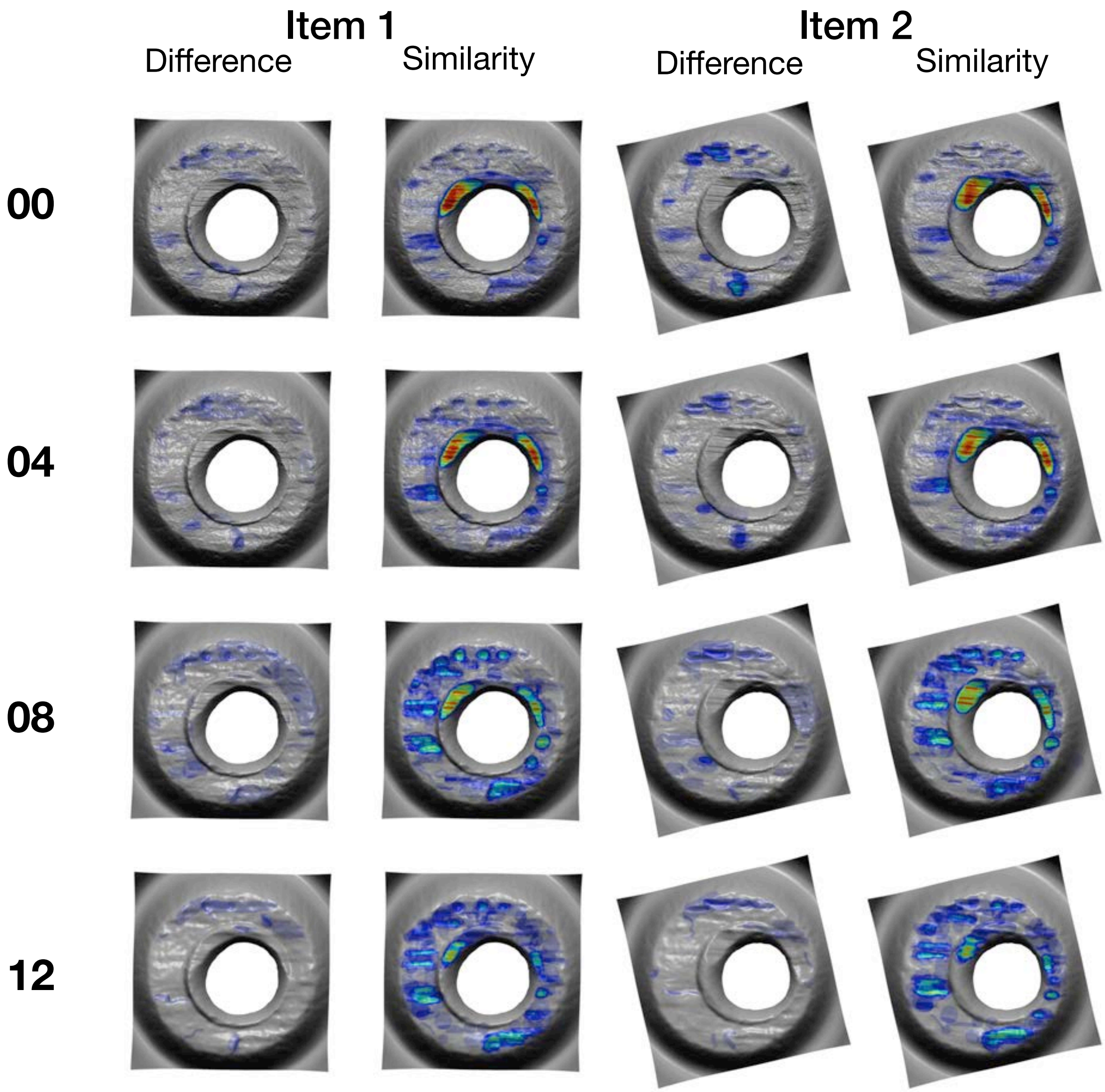


Figure 26: **Annotation Maps for Test Set 14. Known Match.** Columns from Left to Right: Item 1 Difference, Item 1 Similarity, Item 2 Difference, Item 2 Similarity. Rows from Top to Bottom: Full resolution, Resolution 04, Resolution 08, Resolution 12.

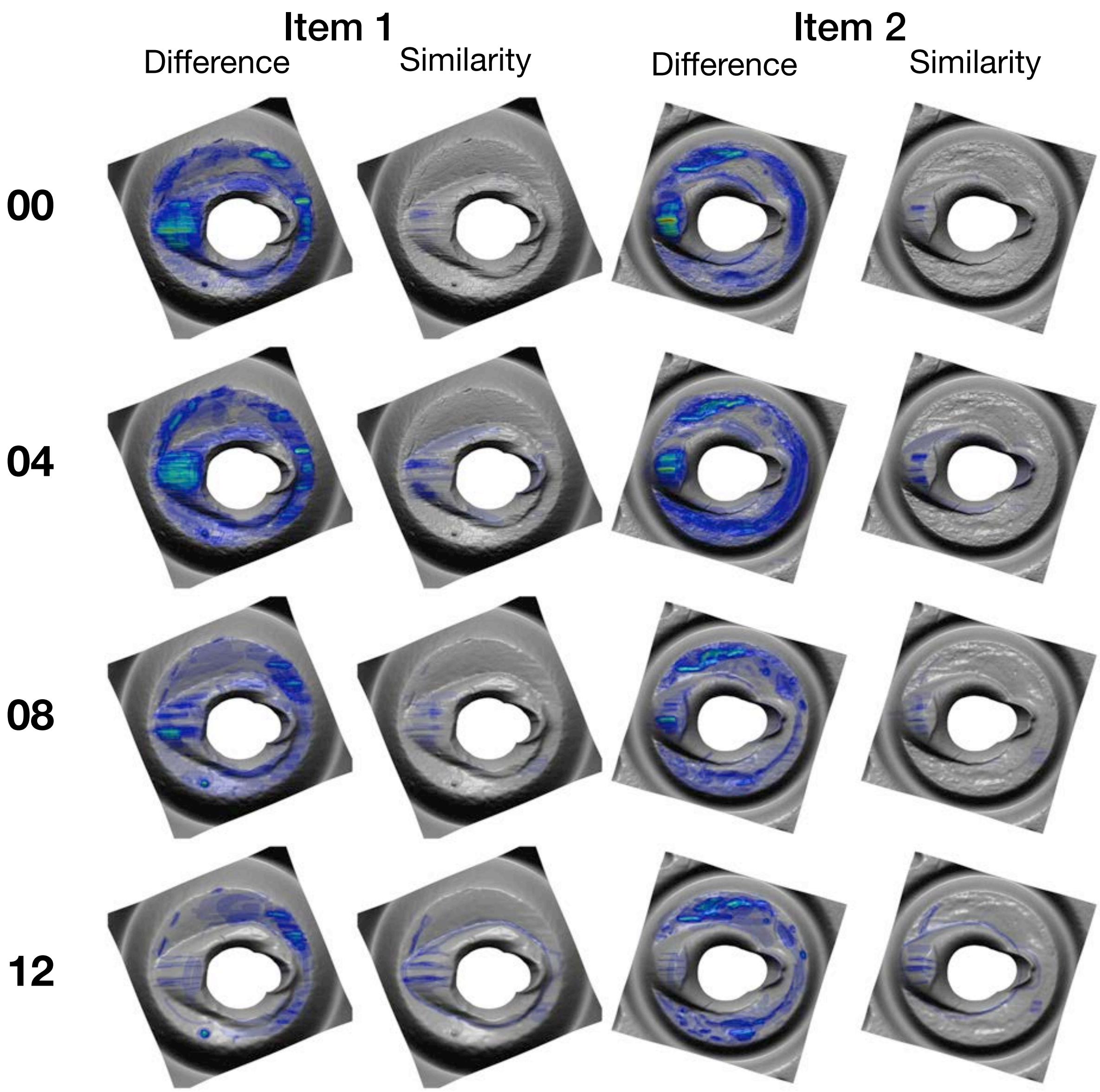


Figure 27: **Annotation Maps for Test Set 15. Known Non-Match.** Columns from Left to Right: Item 1 Difference, Item 1 Similarity, Item 2 Difference, Item 2 Similarity. Rows from Top to Bottom: Full resolution, Resolution 04, Resolution 08, Resolution 12.

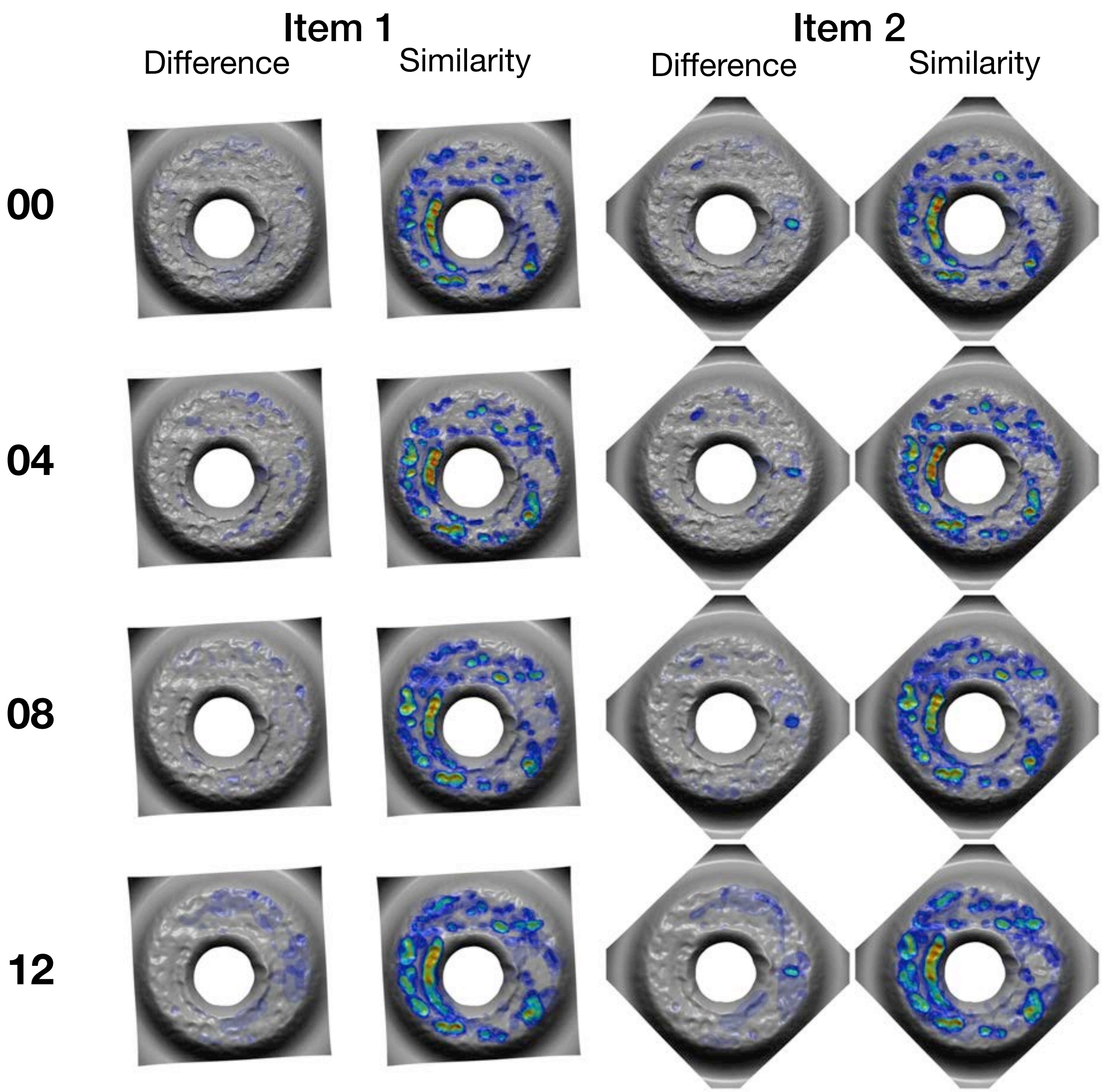


Figure 28: **Annotation Maps for Test Set 16. Known Match.** Columns from Left to Right: Item 1 Difference, Item 1 Similarity, Item 2 Difference, Item 2 Similarity. Rows from Top to Bottom: Full resolution, Resolution 04, Resolution 08, Resolution 12.

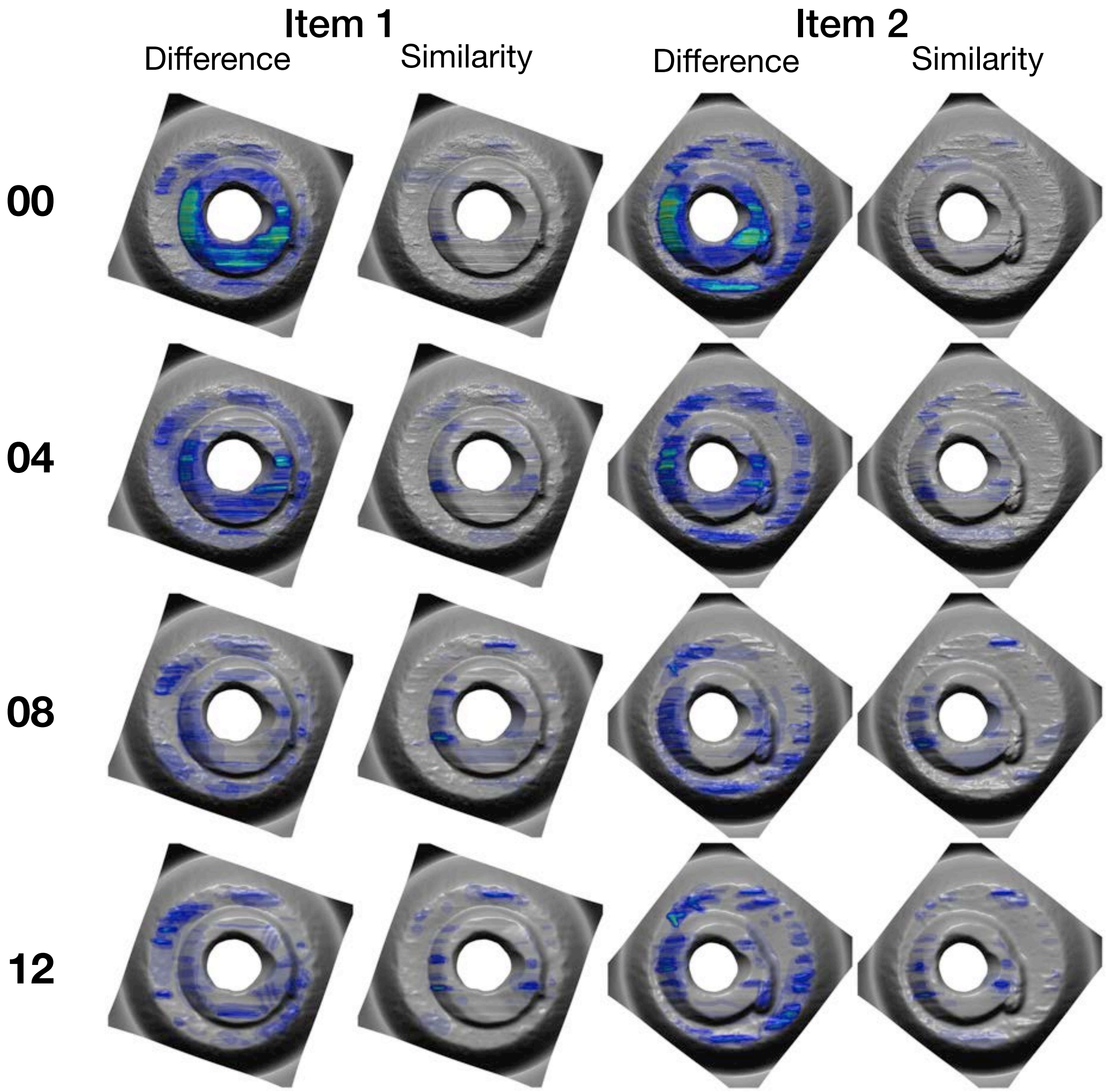


Figure 29: **Annotation Maps for Test Set 17. Known Non-Match.** Columns from Left to Right: Item 1 Difference, Item 1 Similarity, Item 2 Difference, Item 2 Similarity. Rows from Top to Bottom: Full resolution, Resolution 04, Resolution 08, Resolution 12.

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

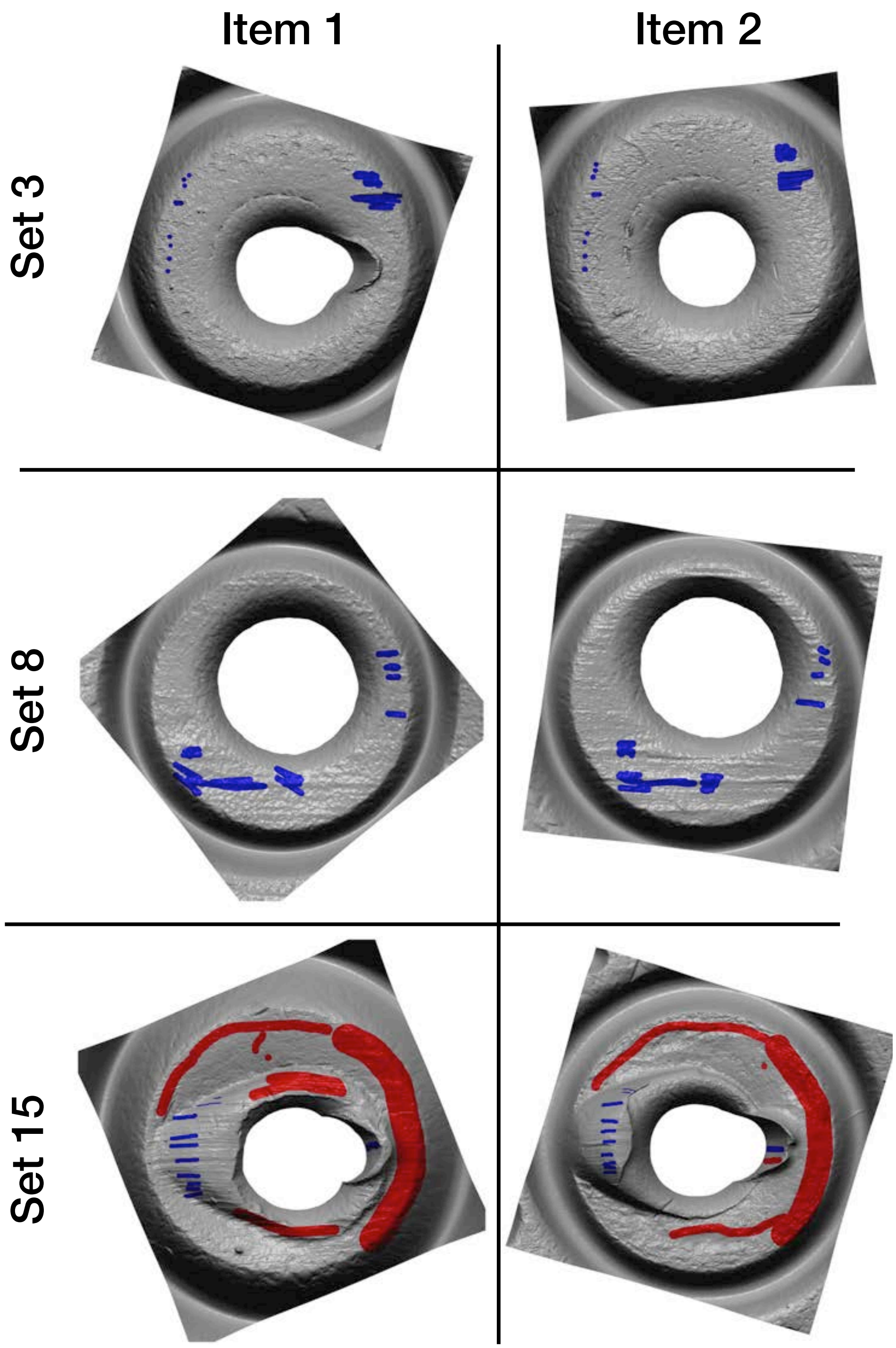
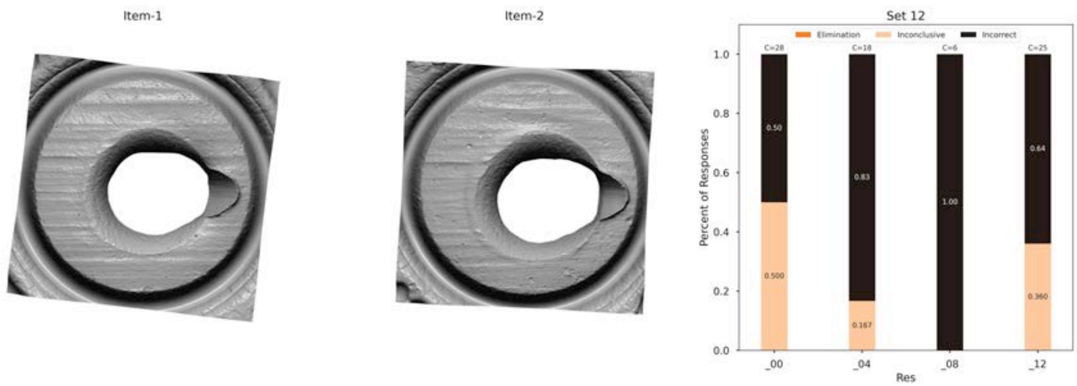


Figure 31: **False Positive Pairs.** The three false positives among the core test sets. Each test set pair is colored to show the areas of the surface marked as similar (blue) and different (red). Note that this color map is different from the annotation maps in the other figures. These figures are colored as annotated by a single participant.

12



Resolution	ID	Inc A	Inc B	Inc C	ELIM
00	14	6	8	0	0
04	15	1	2	0	0
08	6	0	0	0	0
12	16	9	0	0	0

Figure 32: **Results for Test Set 12.** This test set contains a KNM exhibiting subclass marks. As expected, no participant Eliminated the two scans. (Top) The fewest errors were seen in the full resolution scans. Inconclusives (light orange), False Positives (black). Participant counts are located above each bar. (Bottom) A breakdown of conclusions by resolution. Counts are shown. Note that all inconclusives at resolution 12 were Inconclusive A, whereas at full resolution there was a split between Inconclusive A and B.

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] C Chapnick, E. Meschke, P. Duez, T. Weller, J. Marshall, and R. Lilien. Results of 3d virtual comparison microscopy error rate study (VCMERS) for firearm forensics. *(submitted)*, (under review), 2020.

[5] 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.

[6] 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*, 63:1069–84, 2018.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

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

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

[16] 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.

[17] 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.

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

[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.