Report of the

Association of Firearm and Tool Mark Examiners

Proficiency Test Review Ad-Hoc Committee

on the Results of the

Collaborative Testing Service (CTS)

Firearms Examination Proficiency Test 23-5262

November 25, 2024

The Proficiency Test Review Ad-Hoc Committee (PTRC) was established by the Association of Firearm and Tool Mark Examiners (AFTE) Board of Directors on February 22, 2024[1]. The mission of the committee is "to research and investigate firearm & tool mark proficiency test results with current focus on the CTS Test 23-5262." The committee has gathered all available information regarding CTS Test 23-5262 and is presenting it in this report.

BACKGROUND

The Collaborative Testing Service (CTS) Firearms Examination Proficiency Test 23-5262 consisted of five items. Participants were told that item 1 contained three known test fired bullets from the suspect's CZ 75 firearm, fired using ammunition consistent with the bullets found at the crime scene. Items 2 through 5 were each one questioned bullet recovered from a crime scene. Participants were asked to compare the items 2 through 5 crime scene bullets to the item 1 bullets and report their findings. The test was released in October 2023 and the due date for submission of results was December 18, 2023.

PMC Bronze 40 S&W 180 grain FMJ ammunition was used for all five items. The item 1 bullets were fired from a CZ 75 pistol with conventional rifling. The items 2, 3, and 5 bullets were all fired from a second CZ 75 pistol with the same general rifling characteristics[2] as the first CZ 75 pistol. The item 4 bullet was fired from a Desert Eagle pistol with different general rifling characteristics than the CZ 75 pistols[3]

CTS issued its summary report on February 12, 2024.[4] Of the 280 responding participants, 219 reported inconclusive results or eliminations for items 2, 3, 4, and 5 when compared to the item 1 bullets. The remaining 61 participants reported at least one identification conclusion. Table 1 contains a summary of responses for each item.

[1] The members of the ad-hoc committee include: James Carroll (Chair), Andy Smith, Rebecca Smith, Ed Wallace, and Todd Weller

[2] General rifling characteristics (GRCs) refers to the caliber of the bullet and the number, direction of twist, and widths of the land and groove impressions present. These are considered class characteristics on fired bullets.

[3] The Desert Eagle pistol had polygonal rifling, whereas the CZ 75 pistols both had conventional rifling.

[4] Collaborative Testing Services, Inc. Firearms Examination Test No 23-5262 Summary Report, February 12, 2024.

Table 1: Summary of all responses for CTS Test 23-5262				
	Item 2	Item 3	Item 4[5]	Item 5
Yes (Identification)	57 (20.4%)	51 (18.2%)	1 (0.4%)	53 (18.9%)
No (Elimination)	88 (31.4%)	92 (32.9%)	275 (98.2%)	92 (32.9%)
Inconclusive	135 (48.2%)	137 (48.9%)	3 (1.1%)	135 (48.2%)

On July 19, 2024, CTS revised its summary report to include a five-page document containing additional information about test preparation, quality control measures, and the interpretation of results.[6]

Due to the unusually high number of false identification responses, the AFTE Board of Directors formed the PTRC to research and investigate these results.

INQUIRY

Collaborative Testing Service Test Preparation and Quality Control Measures

The PTRC engaged in communication with representatives of Collaborative Testing Service (CTS), which is an ISO/IEC 17043:2023 accredited proficiency testing provider. CTS representatives were extremely cooperative and forthcoming with information while simultaneously protecting the privacy interests of their customers.[7]

In developing a firearms examination proficiency test, CTS identifies a host laboratory with adequate facilities to prepare the test materials and adequate expertise to assist in test preparation. Firearms from the host laboratory reference collection are selected for test production. Trials are completed with the firearms and a variety of ammunition prior to determining the final materials to be used in the test. Once selected, all ammunition used is from the same lot in an attempt to reduce variation.

[5] One participant (9GC4WM) did not report a conclusion for item 4 in response to question 1 on the answer sheet, hence the 279 responses represented in Table 1. However, this participant did provide the wording of their conclusions in response to question 2 on the answer sheet, which indicates an elimination conclusion for item 4.
[6] Collaborative Testing Services, Inc. Firearms Examination Test No 23-5262 Summary Report, July 19, 2024. https://cts-forensics.com/reports/23-5262_Web.pdf
[7] The PTRC met virtually with CTS representatives on March 11, 2024.

For test 23-5262, the host laboratory was ANAB[8] accredited and multiple AFTE certified examiners from the host laboratory collaborated with CTS on the test development.[9] The predetermined design of this test included four questioned bullets, items 2 through 5, that were not fired from the same firearm used to fire the item 1 bullets. Three of the questioned bullets were to be fired from one firearm with similar class characteristics to the firearm used to fire the item 1 bullets, and the fourth questioned bullet was to be fired from a firearm with dissimilar class characteristics.

Once the test design was established, CTS determined the number of cartridges that needed to be fired in each firearm. The total quantity was separated into production batches of approximately 120 cartridges each. A CTS representative was present at the host laboratory during production to guide the process. The CTS representative oversaw the following: only one firearm was used at a time, ammunition was tracked, fired components were counted, and verification of the quality of the fired components was performed. The materials from each batch were sealed in labeled packaging prior to the start of the next batch. Once all required cartridges were fired in a firearm, the firearm was removed from the area, counts reconfirmed, and the fired ammunition components sealed in return shipment packaging. Separate return shipment packaging was used for the fired ammunition components from each firearm.

Verification of the quality of the fired ammunition components was performed after each batch was fired. Ten percent of the fired ammunition components from each batch were selected at random and evaluated for consistency, reproducibility, and quality by two AFTE-certified examiners at the host laboratory prior to the start of the subsequent batch. All ammunition components subjected to verification passed the verification process. The verification process does not involve the direct comparison of ammunition components from different batches to evaluate consistency across batches. Rather, it focuses on consistency within the batch and overall quality of the specimens.[10]

Test set assembly was performed at the CTS offices. Only one batch of fired bullets was opened at a time. A final quality control inspection was performed, and any bullets that appeared damaged or otherwise not suitable for test materials were removed. For item 1, three bullets

---

[8] ANAB is the ANSI National Accreditation Board, which offers accreditation in forensic testing. https://anab.ansi.org/accreditation/iso-iec-17025-forensic-testing-laboratory/

[9] While the host laboratory assists CTS with the preparation and evaluation of the test materials, the host laboratory is not aware of the overall test design (i.e., how the materials will be separated or labeled for the final test).

[10] 5-page Attachment to CTS 23-5262 Summary Report (https://cts-forensics.com/reports/23-5262_Web.pdf.) Page 1: "At the host laboratory, CTS staff ensure only one gun is being used at a time, track ammunition use, collect and verify counts of the expected ammunition, confirm completion of verification, and seal all materials from one batch in clearly labeled packaging prior to the start of the next batch"

"Verification is completed after each batch is fired. Ten percent of each batch is selected at random and evaluated for consistency, reproducibility, and quality by the subject-matter experts of the host lab prior to the start of the next batch."

were selected, marked with their item number using a black marker, placed into a jewel box with a matching item number, and the jewel box sealed. For item 4, one bullet was selected, marked with its item number using a black marker, placed into a jewel box with a matching item number, and the jewel box sealed. For items 2, 3, and 5, the batch was divided into three piles, one for each item number. A bullet from each pile was selected, marked with its item number using a black marker, placed into a jewel box with a matching item number, and the jewel box sealed. Two CTS staff members were present during the assembly process, one of whom marked the ammunition components and the other confirmed placement into the correct jewel boxes and sealed the boxes. All packaged items from the same batch were maintained within their batches.

During final test set assembly, one of each packaged item was placed into the test set box. Within each test set, items 2, 3, and 5 were selected from the same batch, ensuring they were fired within a maximum of 120 shots from each other. Because the item 1 jewel boxes were packed using bullets from the same batch, the three item 1 bullets were also fired within a maximum of 120 shots from each other.

Three laboratories were selected for predistribution testing, two of which were accredited. All predistribution participants reported an inconclusive result for items 2 and 3, and all reported elimination for item 4. Item 5 was eliminated by one participant and reported as inconclusive by two participants. Predistribution participants had the following comments:

- Two of the three felt the test was practical, with the third unsure due to increased difficulty level.
- Two of the three felt the quality of the samples met their labs' standard for testing, with the third responding "N/A". One stated that it was the most difficult proficiency test they have ever received, but that it was more similar to casework and further observed that there could be an increase in inconclusive results.
- One stated that there was limited reproduction of individual characteristics within the item 1 bullets, making identification challenging.

When evaluating inconclusive results, CTS considers that some laboratories have a policy against reaching elimination conclusions when class characteristics agree, as is the case with items 1, 2, 3, and 5. Therefore, CTS groups elimination and inconclusive results together under these circumstances and the eight out of nine inconclusive results from the predistribution laboratories for items 2, 3, and 5 were not of concern.

Demographic Data Provided to the PTRC by CTS

CTS met with the committee and provided information that included a summary of test-takers' responses broken into three demographic categories:[11]

- Group A consisted of 191 participants who authorized the release of results to ANAB or A2LA.
- Group B consisted of 38 participants from ISO/IEC 17025 or 17020 accredited laboratories that did not authorize the release of results to ANAB or A2LA, or were accredited by other international accrediting bodies.
- Group C consisted of 50 participants. CTS was not able to determine if they are affiliated with an ISO/IEC 17025 or 17020 accredited laboratory.

The summary of responses from each of these groups is provided in Table 2.

Table 2: Responses for Demographic Groups A, B, and C[12, 13]

Group	Reported Elim for all items	Reported Inc for items 2, 3, and 5, and Elim for item 4	Reported Elim for item 4 and an ID to any other item	Total
Group A	64 (33.5%)	111 (58.1%)	16 (8.4%)	191 (100%)
Group B	16 (42.1%)	9 (23.7%)	13 (34.2%)	38 (100%)
Group C	7 (14.0%)	11 (22.0%)	32 (64.0%)	50 (100%)

As shown in Table 2, the difference in percentage of error between the groups is stark. To make this data less confusing, the committee elected to combine the Elimination and Inconclusive columns since these responses did not contradict the "assigned value" as indicated in the CTS summary report. This combined data is contained in Table 3.

[11] The separation of the data into groups B and C was performed by CTS staff utilizing internet searches to determine client accreditation. It is acknowledged that this was an imprecise method and may not be an entirely accurate representation of participant demographics for these two groups.

[12] One response was inconclusive for all items, including Item 4. This answer is not included in the demographic summary Tables 2, 3, and 4.

[13] For various tables, Elimination is abbreviated "Elim", Inconclusive is abbreviated "Inc", and Identification is abbreviated "ID".

Table 3: Groups A, B, and C by no error vs at least one error			
Group	**No erroneous responses**	**Reported Elim for item 4 and ID to any other item**	**Total**
Group A	175 (91.6%)	16 (8.4%)	191 (100%)
Group B	25 (65.8%)	13 (34.2%)	38 (100%)
Group C	18 (36.0%)	32 (64.0%)	50 (100%)

To test if the differences above are significant, the Chi Square test for independence was used and a significant result was obtained ($p<0.00001$). This strongly supports that differences in responses between groups are statistically significant (and not simply due to chance). Group A consists of labs submitting their responses[14] to ANAB or A2LA accreditation bodies. Group B is a mixture of labs who are ANAB/A2LA accredited (but elected to not submit their responses for accreditation purposes) and internationally accredited laboratories (that are not ANAB or A2LA accredited). Information was not provided about why an accredited laboratory would elect to not submit responses to their accrediting body. However, one possible explanation is that the test may have been used for training purposes (i.e., taken by a trainee). Group C's accreditation status is unknown. This group could consist of non-accredited domestic and international labs, or private practitioners, who have purchased a test and submitted responses for scoring.[15]

Given the differences observed, the PTRC asked CTS to break the above data into domestic vs. international respondents. Table 4 contains this information.

Table 4: Domestic (USA only) versus International			
Group	**No erroneous responses**	**At least one erroneous response**	**Total**
Domestic	177 (90.8%)	18 (9.2%)	195 (100%)
International	41 (48.8%)	43 (51.2%)	84 (100%)

[14] Laboratories must choose to submit their responses to accrediting bodies *before* the ground truth and expected results are revealed. This is done at the time the responses are submitted to CTS.

[15] CTS noted they are not able to distinguish between forensic practitioners vs. non-practitioners and they have no knowledge or evidence of results being submitted from non-practitioners.

The Chi square test for independence strongly supports that the differences between domestic and international responses are significant ($p<0.00001$).

The above demographic data strongly support the conclusion that error was not evenly distributed amongst all participants.[16] As noted by the National Academy of Sciences:

> "The pooling of proficiency-test results across laboratories has been suggested as a means of estimating an "industry-wide" error rate [citation omitted]. But that could penalize the better laboratories; multiple errors on a single test by one laboratory could substantially affect the overall estimated false-match error rate."[17]

Regardless, the rate of error, even among laboratories in the United States, is an outlier from prior tests, and the PTRC sought to better understand the cause(s) of the errors.[18]

CTS's Assistance in Gathering Information from Test Participants

CTS respects the confidentiality of the host laboratory and all test participants. CTS did not identify the host laboratory to the PTRC, but agreed to contact the host laboratory on behalf of the PTRC and ask the host laboratory to contact the PTRC. No communication from the host laboratory has been received.

CTS sent an email on March 20, 2024, to customers who purchased this test, notifying them of the PTRC and the PTRC's email address (PTinvestigate@AFTE.org). CTS also offered to act as an intermediary between test participants and the PTRC and anonymize any information that a participant wished to share with the PTRC. On July 19, 2024, CTS sent another email to customers who purchased this test. Attached to that email was a five-page document containing additional information about test preparation, quality control measures, and the interpretation of results. This same document was incorporated into the summary report referenced above. Also in that email was a reminder that the PTRC was still seeking information from test participants and asking participants to complete a survey (see below), which was hyperlinked. Two test participants contacted the committee anonymously via CTS, and the information they provided will be discussed below.

[16] Note that the Texas Forensic Science Commission found that none of the 18 Texas participating laboratories had reported a false positive error. See page 36 of the report "Final Report on Complaint NO. 21.27; National Innocence Project, University of Colorado Law School (Houston Police Dept Crime Lab; Firearms/Tool Marks) April 26, 2024". https://www.txcourts.gov/media/1459122/complaint-2127-final-report-09042024.pdf

[17] National Research Council "The Evaluation of Forensic DNA Evidence" National Academy Press, Washington DC, 1996. pg 86.

[18] Peterson J, Markham P "Crime Laboratory Proficiency Testing Results, 1978-1991, II: Resolving Questions of Common Origin" Journal of Forensic Sciences 40(6) 1995 pg. 1009-1029. The overall false positive rate for these 14-years of proficiency test data was 1.1%.

Information Received Directly from Test Participants

The AFTE Board of Directors sent a letter to the membership of the organization on February 15, 2024, requesting that anyone with information pertinent to the test contact AFTE via a newly created email address, PTinvestigate@AFTE.org. This email address was also posted on the AFTE forums and distributed by CTS, as indicated in the preceding section of this report. A number of test participants and other concerned parties, including many international examiners, came forward with offers of assistance, general comments regarding their experiences with the test, and/or recommendations for the PTRC to consider.

General comments from participants included the following, in summary:

- The condition of the bullets was poor in some instances, to include abrasions on the ogive and bearing surface, suspected to be water tank damage
- Few individual characteristics/poorly marked
- Poor reproducibility of marks amongst the item 1 bullets

Recommendations for the PTRC to consider included the following, in summary:

- Participant demographics
- Number of submitted results versus number of tests purchased
- Test design and production
- Laboratory policies that may have influenced results

Two test participants who had reported identification conclusions came forward and spoke with the PTRC under the condition that their identities would remain confidential. They will be referred to here as Examiner #1 and Examiner #2. Two additional test participants who had reported identification conclusions came forward anonymously through CTS and provided statements and copies of documentation. They will be referred to here as Examiner #3 and Examiner #4. The reported results from these four examiners are shown in Table 5.

Table 5: Test Answers from the Four Examiners who Provided Additional Information

	Item 2	Item 3	Item 4	Item 5
“Examiner #1”	No (ELIM)	No (ELIM)	No (ELIM)	Yes (ID)
“Examiner #2”	Yes (ID)	Yes (ID)	No (ELIM)	Yes (ID)
“Examiner #3”	Yes (ID)	Yes (ID)	No (ELIM)	Yes (ID)
“Examiner #4”	Yes (ID)	Yes (ID)	No (ELIM)	Yes (ID)

EXAMINER #1

Examiner #1 (E1) is based in the United States and began training in the mid-1990s. E1's training was modeled after the AFTE Training Program and E1 was authorized for independent comparison casework in 1999. E1 has participated in ongoing professional development since that time, and has participated in nearly every available validation study during E1's career. E1 uses a modern light comparison microscope and takes photographs of comparisons. E1 uses traditional pattern matching and prefers to observe agreement in multiple land impressions in order to reach an identification conclusion. Elimination conclusions based on disagreement of individual characteristics (when class characteristics agree) are permissible in E1's laboratory.

E1 has been taking CTS proficiency tests since 1998 and has never before reported an incorrect result, nor has E1 ever reported an inconclusive result. Inconclusive results are considered "perfectly acceptable" in E1's laboratory, both on proficiency tests and in casework.

E1 treats proficiency tests like actual casework. Just as E1 would do in casework, E1 first compared the item 1 bullets to one another and was able to index them. The item 1 bullets were sub labeled items 1A, 1B, and 1C. Item 4 was immediately eliminated based on disagreement of class characteristics. Item 1A was compared to item 2, then to item 3, and finally to item 5. The comparisons were challenging and no conclusions were reached at that point. Item 2 was then compared to item 3 and an identification conclusion was reached. E1 stated that a lot of agreement was observed and it was not difficult to reach a conclusion. At this point, items 2 and 3 were eliminated from item 1 because the level of agreement observed on items 2 and 3 was significantly greater than item 1A. Items 2 and 3 were compared to item 5, but E1 was unable to index them to one another. Because of the high level of agreement between items 2 and 3, E1 concluded that item 5 was fired from a different firearm. Item 5 was then compared to items 1A, 1B, and 1C. E1 believed sufficient agreement was observed, leading to a conclusion of identification. E1's conclusions were verified independently, but not blinded, by another qualified examiner from the same laboratory and the verifier agreed with E1's conclusions without any further discussion.

E1 stated that this test was the most difficult CTS test E1 has ever taken. E1 spent nearly 10 hours on this test, which was significantly longer than other CTS tests. E1 did not have any assumptions about the expected results prior to taking the test, and a test with all different-source bullets when compared to the test bullets would not be surprising. E1 stated that, while inconclusive results are acceptable, E1 felt that E1 "should be able" to reach definitive conclusions on a proficiency test. At the time E1 submitted the results, E1 was confident that they were correct. Now knowing the ground truth, E1 accepts that the submitted results were not correct.

E1 provided the four photographs taken during the comparison of item 1A to item 5 which resulted in an identification conclusion. See Appendix A, Figures A1 through A4.

EXAMINER #2

Examiner #2 (E2) is based in the United States and began training in 2019. E2's training mirrored the AFTE Training Program and E2 was authorized for independent bullet and cartridge case comparisons in 2021. E2 has participated in ongoing professional development since that time, but not specifically comparison training. E2 uses a modern light comparison microscope and takes photographs of comparisons. E2 uses traditional pattern matching. Elimination conclusions based on disagreement of individual characteristics (when class characteristics agree) are permissible in E2's laboratory.

E2 has been taking CTS proficiency tests for the past few years and has never before reported an incorrect result. Three days before the due date for this proficiency test, while E2 was in progress taking the test, the quality assurance manual in E2's laboratory was amended. New language was added to the effect that ANAB considers inconclusive results on proficiency tests to be "unexpected results" and this must be reported to ANAB. E2's interpretation of this was that definitive conclusions must be reached in order to pass the test.

Just as E2 would do in casework, E2 first compared the item 1 bullets to one another to assess them. Item 4 was eliminated based on disagreement of class characteristics. The item 1 bullets were compared to items 2, 3, and 5 for 32 hours over a 45-day period of time. The bullets were not well marked. E2 was planning to report inconclusive results for all three items, however, once E2 became aware of the quality assurance manual revisions (discussed above), E2 decided to report the conclusion that E2 was nearest to, which was identification. This was based, in part, on the fact that E2 was unable to reach an elimination conclusion. E2's conclusions were verified independently, but not blinded, by another qualified examiner from the same laboratory. There was discussion between E2 and the verifier, both of whom were hesitant to report identification conclusions, however, both were aware of the recent quality assurance manual amendment and felt that a definitive conclusion was required.

E2 provided the three photographs taken during the comparison. See Appendix A, Figures A5 through A7.

EXAMINER #3

Examiner #3 (E3) submitted a statement through CTS, which was anonymized. The relevant portions of that statement are summarized as follows:

E3 had approximately six months of independent casework experience prior to receiving this test and had not taken a CTS proficiency test previously. E3 did not begin the test immediately upon receiving it due to casework obligations within the laboratory. After a few days working on the test, E3 felt pressured to make a decision because the due date was quickly approaching. Unlike casework, proficiency tests were not subjected to any form of technical or administrative review prior to the submission of results to the test provider. E3 was under the belief that inconclusive

results were not an option because a test has only a right or wrong answer. Because E3 observed some areas of agreement, E3 decided to report identification conclusions for items 2, 3, and 5.

It appears from E3's documentation that item 4 was eliminated based on disagreement of class characteristics.

E3 provided the photographs taken during the comparison. See Appendix A, Figures A8 through A17.

EXAMINER #4

Examiner #4 (E4) submitted a statement through CTS, which was anonymized. The relevant portions of that statement are summarized as follows:

E4 stated that two of the item 1 bullets could be "matched" to one another, but the third bullet was poorly marked and could not be "matched". The item 1 bullets were identified to items 2, 3, and 5 based on marks in the land impressions. Items 2, 3, and 5 "match each other well" and they "match" the item 1 bullets, but not as well because the item 1 bullets were not marked well. If this were actual casework, additional bullets would have been test fired. Item 4 was different in appearance and could not have been fired from the same pistol as the item 1 bullets.

Based on the phraseology and terminology used by E4, it appears that E4 is not a native English speaker. Therefore, some details may be lost in language translation.

E4 provided the photographs taken during the comparison. See Appendix A, Figures A18 through A27.

Participant Survey

In order to gather information from the largest number of test takers possible, the PTRC developed an online survey intended for all test participants. Dr. Pate Skene of the University of Colorado Boulder, who is also the Forensic Science Standards Board liaison to the OSAC[19] Human Factors Task Group, provided invaluable assistance to the committee with the creation of the survey. In developing the survey, the PTRC attempted to balance a desire to gather as much information as possible with the time required to complete the survey and the impact the time commitment might have on participation.

The survey was first announced on May 27, 2024, during a presentation at the annual AFTE conference in Anchorage, Alaska. An email to the membership was sent on May 29, 2024, and an AFTE forum post was made on that same day. Additionally, on June 12, 2024, and again on July 19, 2024, CTS sent an email to customers who purchased the test.

[19] Organization of Scientific Area Committees for Forensic Science: https://www.nist.gov/organization-scientific-area-committees-forensic-science

There were 127 responses to the survey. However, seven respondents did not indicate their conclusions, making their responses less useful, and one appeared to be a duplicate submission.[20] For the discussion that follows, only the remaining 119 survey responses will be considered.

15 of the survey respondents indicated at least one false identification conclusion involving items 2, 3, and 5.[21] One of the 15 also indicated a false identification conclusion involving item 4.[22] These 15 respondents include Examiners #1 and #2, described in the preceding section.[23] See Table 6 for a summary of the indicated conclusions for these 15 respondents.

Table 6: Summary of survey answers for 15 respondents who indicated at least one false identification

	Item 2	**Item 3**	**Item 5**
Identification	13	12	13
Inconclusive	1	1	2
Elimination	1	2	0

These 15 respondents range in experience from less than one year to over 20 years. Eight are located in the United States. They were trained in a variety of ways including on-the-job training, formal courses in their respective laboratories, joint training with other agencies, and outside training. 11 indicated that their conclusions are verified on proficiency tests. All are employed in a crime laboratory, 11 of which are accredited. One of the 15 respondents indicated they were not (yet) fully trained and authorized to conduct independent comparisons (i.e. were still in training). 12 of the respondents indicated that the item 1 bullets were of value for class and individual characteristics comparison but were poorly marked. Two indicated that the item 1 bullets were marked as well as bullets typically encountered in casework, and one indicated that the item 1 bullets were of value for class characteristic comparison only.

[20] It appears that survey respondent #75 is a duplication of survey respondent #74 because all answers are identical, including narrative responses, and they were submitted within seconds of one another. Because of this, the responses from survey 75 were discounted. Surveys 74 and 75 are surveys in which at least one false identification was reported.

[21] Ibid

[22] This respondent reported identification conclusions for all four questioned bullets. Additionally, this respondent is the only one to indicate on the survey that this CTS test was "easier than some proficiency tests". ALL other respondents indicated that this CTS Test was more difficult than others.

[23] The survey was intended for ALL test participants, and Examiners #1 and #2 confirmed to the PTRC that they submitted a response to the survey. The identities of Examiners #3 and #4 are not known to the PTRC and, therefore, it is not known if they participated in the survey.

Of the 15 participants, many reported they "observed some agreement but used other reasoning to conclude identification was the best choice." If this option was selected, the respondents were given the following options to explain their "other reasoning":

- "I assumed identifications were present (i.e. Item 4 was an elimination therefore Items 2, 3 and 5 were likely "identifications")"
- "I did not think inconclusive was appropriate to use even if it was technically allowed"
- "Inconclusive was not allowed, therefore I took my best guess."

If "other reasoning" was not selected, the respondents stated they had "observed sufficient and significant agreement of individual characteristics". The proportions of these responses for each item are summarized in Table 7.

Table 7: Survey responses to "What was the basis for your identification" for respondents who reported having made at least one false identification

	Item 2	**Item 3**	**Item 5**
Total # errors	13	12	13
Used other reasoning	8 (62%)	6 (50%)	6 (46%)
● **Assumed ID was present**	5	5	4
● **Inc not allowed, took best guess**	1	1	1
● **Felt Inc not appropriate, even if technically allowed**	2	0	1
Observed sufficient agreement	5 (38%)	6 (50%)	7 (54%)

For each false identification conclusion, the respondents were asked to indicate their level of confidence in their conclusion using the following range, which was later assigned the numerical value indicated in parenthesis for statistical purposes: not at all confident (1), slightly confident (2), somewhat confident (3), moderately confident (4), and extremely confident (5). The average and median levels of confidence range from slightly confident to somewhat confident. See Table 8 for a summary of confidence reported by examiners who reported false identifications.

Table 8: Summary of confidence for reported false identifications						
	Item 2 Verbal Confidence	**Item 2 numerical equivalent**	**Item 3 Verbal Confidence**	**Item 3 numerical equivalent**	**Item5 Verbal Confidence**	**Item 5 numerical equivalent**
Average	Somewhat	3	Somewhat confident	3	Somewhat confident	3
Median	Slightly confident	2	Slightly confident	2	Somewhat confident	3
N	13		12		13	

Of the 15 respondents who reported at least one false identification conclusion, three (20%) indicated that definitive conclusions were expected on proficiency tests. Of the 104 respondents who did not report a false identification conclusion, 8 (8%) indicated that definitive conclusions were expected on proficiency tests. Fisher's Exact test was used to determine whether there was a significant relationship between reporting an error on this test and definitive conclusions being expected on proficiency tests. A significant relationship was not found (p=0.1435), indicating the difference in percentages above can be explained by chance. However, the committee encourages readers of this report to use caution in interpreting this result. As shown in Table 7, and in the interviews, the committee has evidence that *some* errors were likely caused by examiners believing their laboratory policy required definitive conclusions, but that was not the *only* factor to cause error.

The survey results are contained in an Excel data file, which can be sorted using the Excel software. Because of the increased utility of being able to sort the data, the file containing the entire survey results has been submitted to AFTE for publication on the AFTE website alongside this report.[24]

Review of Test Sets

Multiple test sets were made available to, and were microscopically reviewed by, members of the PTRC. This review confirmed reports from examiners that were either received directly by the PTRC, were included in the final CTS report, or were available by other means such as on the AFTE forums, that there was a lack of consistency in the samples included in the test. Low reproducibility of individual characteristics was often observed amongst the item 1 bullets, despite knowing they originated from the same source. In some instances the item 2, item 3, and item 5 bullets demonstrated a higher level of reproducibility which could allow same-source

[24] www.afte.org

attribution between these three items, but this again was inconsistent and varied from test set to test set.[25] As far as could be determined, and where the labeling persisted, all of the bullets were labeled correctly and there were no indications that any of the samples had been packaged incorrectly leading to erroneous results.[26]

Evaluation of Similarity Scores Using 3-Dimensional Technology

The committee was provided 10 different CTS test sets, identified in this report as Test Set 1 through Test Set 10. Six of these test sets had not been sent to laboratories and CTS donated these to the committee for evaluation. Additionally, four other laboratories provided complete test sets after their own evaluation. A committee member had access to an instrument[27] that can measure microscopic marks and then use a computer algorithm to intercompare the surface measurements.[28, 29]

The committee elected to use this instrument to measure and compare the test sets. This was done to determine if there were measurable differences between different sets. The instrument provides a similarity score from 0 to 1.0, with a higher score representing greater geometric surface similarity. Several of the results are found in the figures below, and all the results are found in Appendix B. As a reminder about the ground truth of these samples: item 1 contained three bullets, which examiners were told had been fired from the same firearm (meant to represent "test fires" collected from a recovered firearm). CTS has reported that items 2, 3, and 5, all meant to represent "evidence" bullets, were fired from the same firearm, but different from item 1. Therefore, when comparing item 1, there are three same-source comparisons. When comparing items 2, 3, and 5 with each other, three additional same-source comparisons are performed. Finally, when comparing item 1 (3 bullets) to items 2, 3, and 5, a total of 9 different-source comparisons are possible. Those comparisons are represented in Figures 1, 2, and 3 in blue (item 1 intra-compared with itself), green (items 2, 3, and 5 with each other) and orange (item 1 vs items 2, 3, and 5).

[25] One member of the PTRC purchased two test sets for their laboratory. The bullets in one test set were adequately marked such that it was determined definitive conclusions were appropriate. In the second test set the bullets were not adequately marked and conclusive results could not be supported (i.e., inconclusive results were the most appropriate response).

[26] It should be noted that the PTRC received no information from examiners who came forward suggesting that their samples had been swapped resulting in incorrect conclusions being reported to CTS.

[27] In this case the instrument brand is Evofinder. The committee used this instrument out of convenience since it was in-house. The use of Evofinder for the purposes of this report is not an endorsement of a particular instrument provider.

[28] There are numerous research articles that discuss this type of technology. For some background information see: Senin N. et al "Three-Dimensional Surface Topography Acquisition and Analysis for Firearm Identification" J Forensic Sci, 51(2), 2016 doi: 0.1111/j.1556-4029.2006.00048.x

[29] See also: Vorburger TV, Petraco N "Topography Measurements and Applications in Ballistic and Tool Mark Identifications" Surf. Topogr: Metrol. Prop. 4, 2016 doi: 10.1088/2051-672X/4/1/013002

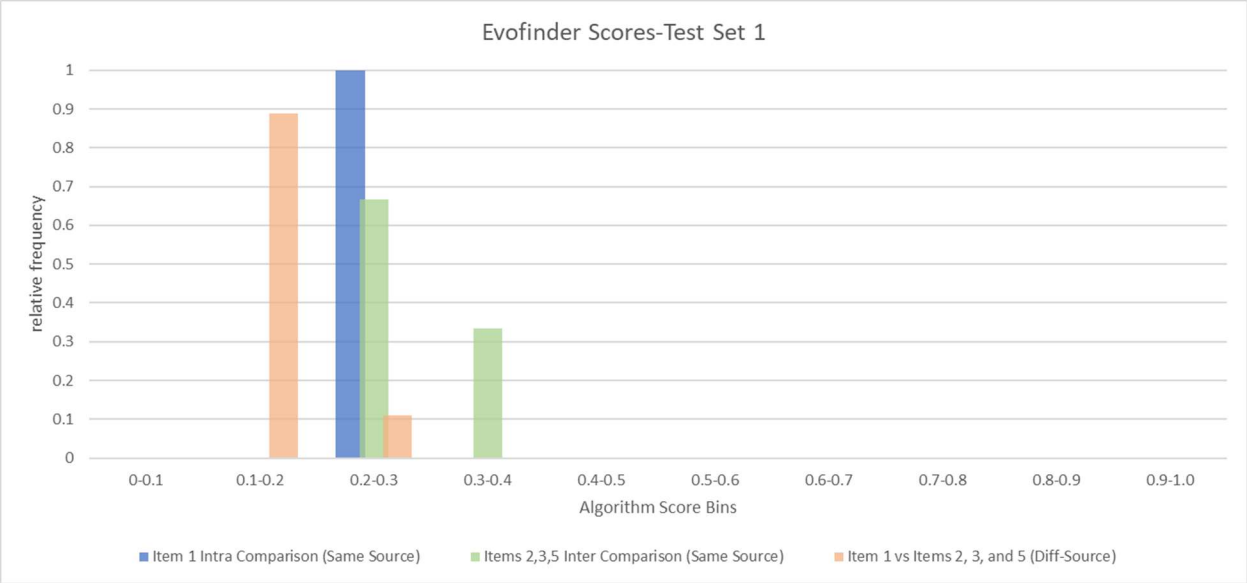


Figure 1: Relative Frequency of Evofinder Algorithm Scores from Test Set 1. Blue represents intra-comparison of item 1 (three bullets all fired from the same firearm). Green represents inter-comparison of items 2,3, and 5 (three bullets all from the same firearm, but different from item 1). Orange represents comparisons of bullets from item 1 to items 2, 3, and 5 (all non-matching, or different-source comparisons).

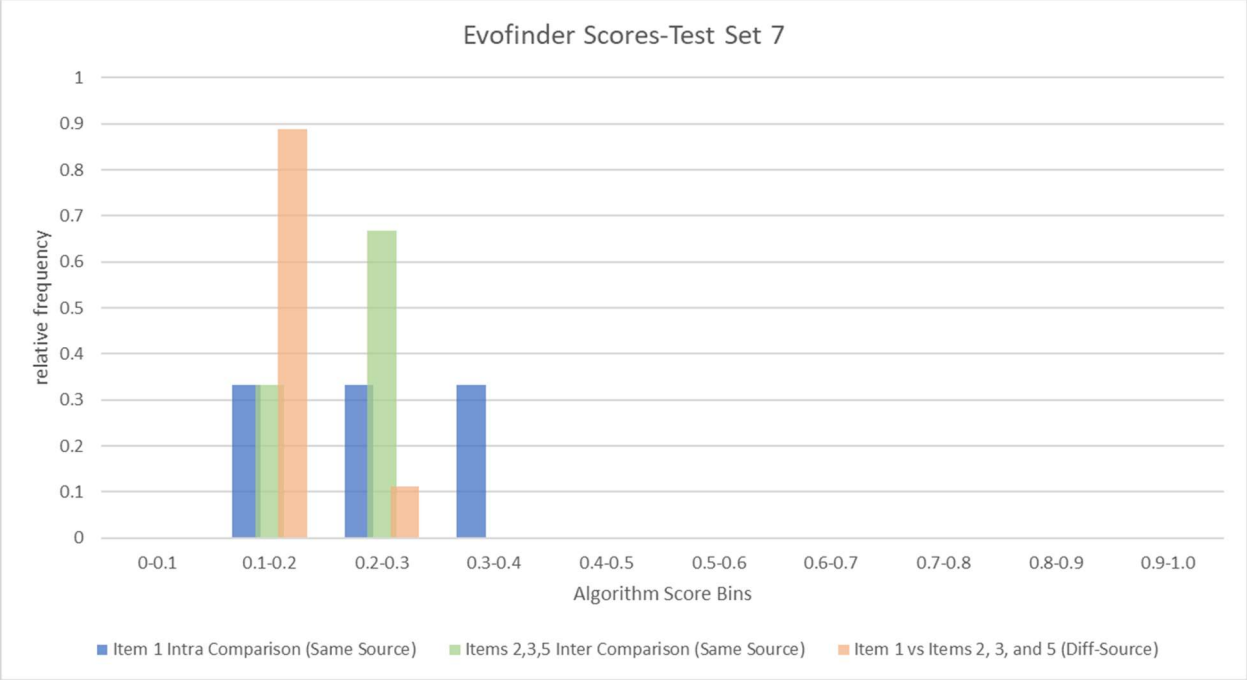


Figure 2: Relative Frequency of Evofinder Algorithm Scores from Test Set 7. Blue represents intra-comparison of item 1 (three bullets all fired from the same firearm). Green represents inter-comparison of items 2,3, and 5 (three bullets all from the same firearm, but different from item 1). Orange represents comparisons of bullets from item 1 to items 2, 3, and 5 (all non-matching, or different-source comparisons).

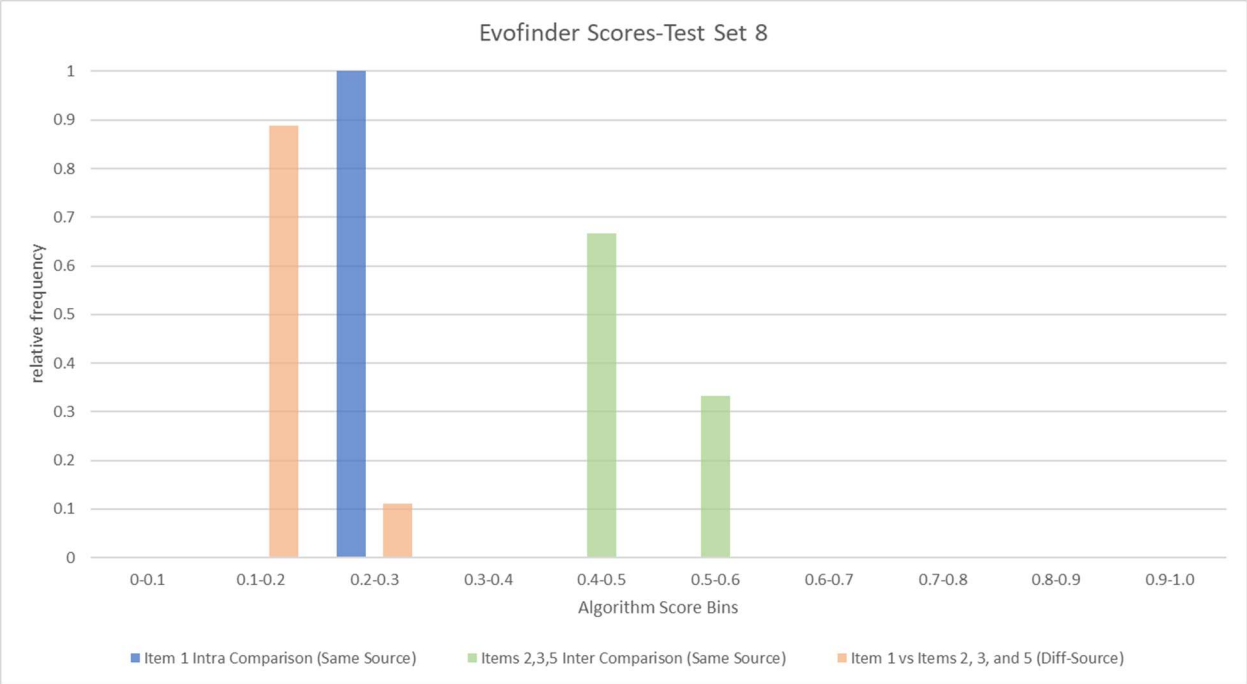


Figure 3: Relative Frequency of Evofinder Algorithm Scores from Test Set 8. Blue represents intra-comparison of item 1 (three bullets all fired from the same firearm). Green represents inter-comparison of items 2,3, and 5 (three bullets all from the same firearm, but different from item 1). Orange represents comparisons of bullets from item 1 to items 2, 3, and 5 (all non-matching, or different-source comparisons).

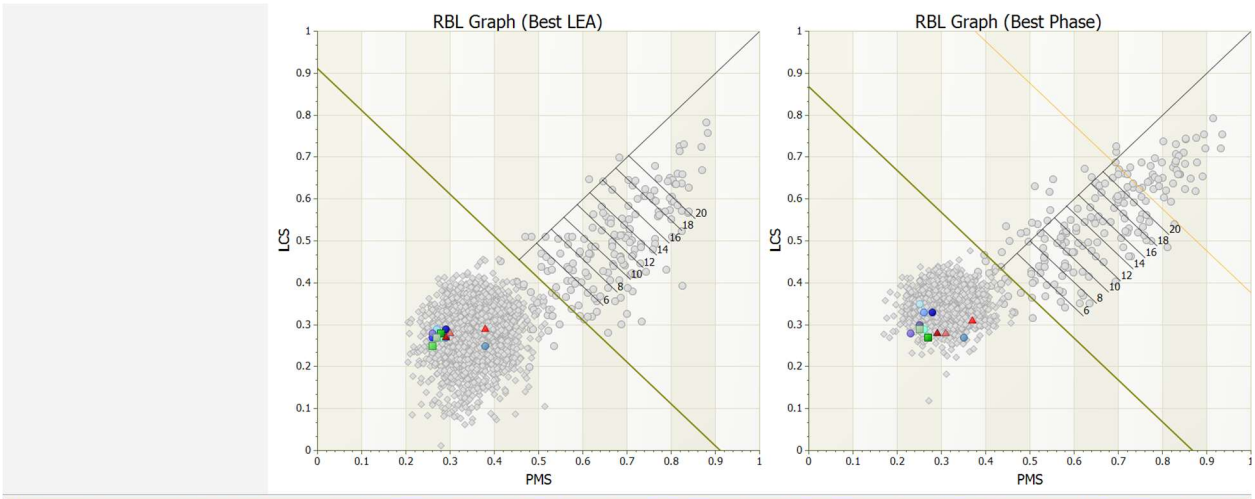
The three figures provide data to support several conclusions. The first is that there was set-to-set variation. Test Set 1 and Test Set 7 had algorithm scores that were all relatively low and clustered together for the different source comparisons and same source comparisons. These results are different from Test Set 8, where the items 2, 3, and 5 comparisons were algorithmically separated from the other comparisons. However, the item 1 intra-comparisons still resulted in relatively low scores that were not algorithmically separated from the different-source comparisons. This data provides an important lesson to all stakeholders: the use of test consensus to judge the "correctness" of an individual's responses may be inappropriate. Because current proficiency tests involve shipping individually fired ammunition components, there may be significant test-to-test variance.[30]

[30] Even if there were no test-to-test variance, the use of consensus to judge correctness is still problematic. For example, see Arkes and Koehler "Inconclusives and error rates in forensic science: a signal detection theory approach" Law, Probability and Risk 20 (2021). pg 163: "A third reason to be wary of a 'wisdom of the crowd' gold standard is that some individuals may be more skilled than the crowd. Whereas the crowd might deem a particular comparison to be inconclusive, a particularly skilled examiner might pick up on one or more subtle cues that indicate that a particular paired sample came from different sources. Under the wisdom of the crowd standard, that individual's conclusion would be scored as an error even when it is objectively correct (see Biederman and Kotsoglou 2021, p 5 and Weller and Morris 202, p. 701 for a similar point)."

The second observation from the Evofinder data is that in each of the test sets, the item 1 intra-comparisons resulted in relatively low scores. These figures support observations reported to the committee, and directly observed by committee members, that the item 1 bullets were poorly marked and examiners had difficulty finding areas of reproducibility despite knowing they originated from the same firearm.

Finally, the 3D results provide no affirmative evidence of sample mislabeling or mispackaging. While these results support that many test sets had poorly-marked bullets, the committee found no evidence that same-source items were accidentally swapped for items that were supposed to be from different sources. The above conclusion does not represent a large sampling of all test sets, but it is a possibility the committee considered and found no evidence to support.

The Houston Forensic Science Center (HFSC) has a different 3D measurement system.[31] HFSC personnel scanned several test sets and shared their results with the PTRC. The results from one test set is in Figure 4 and the remaining data can be found in Appendix C.



Acquisition Name 1	Acquisition Name 2	Best LEA PMS	Best LEA LCS	Best LEA FMR	Best Phase PMS	Best Phase LCS	Best Phase FMR
2023PT2	2023PT1C	0.38	0.25	0.00	0.35	0.27	0.00
2023PT3	2023PT3	0.38	0.29	0.30	0.37	0.31	0.32
2023PT3	2023PT5	0.30	0.28	0.00	0.31	0.28	0.00
2023PT5	2023PT1B	0.29	0.27	0.00	0.26	0.33	0.00
2023PT3	2023PT1A	0.29	0.28	0.00	0.25	0.30	0.00
2023PT3	2023PT1C	0.29	0.29	0.00	0.28	0.33	0.00
2023PT2	2023PT5	0.29	0.27	0.00	0.29	0.28	0.00
2023PT2	2023PT1B	0.28	0.27	0.00	0.25	0.35	0.00
2023PT1A	2023PT1B	0.28	0.28	0.00	0.27	0.27	0.00
2023PT5	2023PT1C	0.28	0.27	0.00	0.26	0.29	0.00
2023PT3	2023PT1B	0.27	0.29	0.00	0.26	0.29	0.00
2023PT1B	2023PT1C	0.27	0.27	0.00	0.25	0.29	0.00
2023PT5	2023PT1A	0.26	0.28	0.00	0.23	0.28	0.00
2023PT2	2023PT1A	0.26	0.27	0.00	0.25	0.29	0.00
2023PT1A	2023PT1C	0.26	0.25	0.00	0.25	0.29	0.00

Figure 4: Graphs from a Houston Forensic Science Center CTS test set. The colored dots represent a comparison of two items within one CTS test set. The gray dots are the underlying database population used by the algorithm. The colored dots fall below the diagonal line, indicating that for all comparisons (same source and difference source) within this test set, the instrument found no support for a same source conclusion.

[31] The system was the Ultra Forensic Technology Quantum 3D Microscope (Q3M). For more information on the algorithm and scoring function, see the following report: Roberg D, Beauchamp A, Levesque S "Objective Identification of Bullets Based on 3D Pattern Matching and Line Counting scores" International Journal of Pattern Recognition and Artificial Intelligence 33(11) 2019 DOI:10.1142/S0218001419400214

The graphs from the HFSC instrument show all comparisons falling below the diagonal line, indicating the instrument and algorithm found no substantial correspondence of toolmarks on the bullets. Similar results were found for all 4 sets scanned by the HFSC.

In closing, at the time of this report the committee is not aware of any studies comparing algorithms such as described in this small study to examiner performance. The committee is not recommending that laboratories immediately deploy 3D-based algorithms for use in casework without complete and thorough validation. However, the committee does recommend research to determine if these types of instruments could be used as a possible quality control measure for both proficiency test providers as well as forensic laboratories.[32]

Non-Elimination Conclusions for Item 4

Information from CTS Summary Report

As shown in Table 1, one identification and three inconclusive results were reported for item 4. Item 4 was fired from a firearm with polygonal rifling, whereas items 1, 2, 3, and 5 were all fired from firearms with conventional rifling.

For the participant (Web Code WD8HVT) who reported an identification conclusion for Item 4, the narrative conclusion contained in Table 2 of the CTS report indicates that Item 4 "... was fired by a second firearm." This suggests that an error was made by this participant when entering the result for Item 4 into the answer sheet.[33]

For the participants (Web Codes DF4HAH, KPN4DT, and H8972G) who reported inconclusive results for Item 4, one (DF4HAH) indicated on the CTS answer sheet that "Item 4 has extremely shallow rifling" and another (KPN4DT) indicated that "Land borders are not so clear". No other information was provided to CTS that further explains the inconclusive results.

Information from Participant Survey

One survey respondent indicated that they reached an identification conclusion for Item 4 based on the observation of sufficient and significant agreement of individual characteristics. However, this contradicts any information contained in the CTS report and appears to be an error when completing the survey.

Five survey respondents indicated they reached inconclusive results for Item 4, which is more than is contained in the CTS report, calling into question the accuracy of these responses. However, one of the respondents, who reported inconclusive for all unknown

[32] For a discussion about different levels of algorithm deployment in forensic pattern matching disciplines see: Swofford H, Champod C "Implementation of algorithms in pattern & impression evidence: A responsible and practical roadmap" Forensic Science International: Synergy 3 (2021) https://doi.org/10.1016/j.fsisyn.2021.100142

[33] It should be noted that this participant's narrative conclusions in table 2 of the CTS report indicate an identification conclusion for items 2, 3, and 5.

> bullets, included the following comment: "My item 4 was also of such low quality I could not conclude anything other than inconclusive (B) due to the shallowness of the rifling and lack of any characteristics. My verifier determined the same and because it was a CTS we had a third person verify all of the findings who also reached the same conclusions."

Assuming the bullets in these test kits exhibited *discernible* class characteristic differences (i.e., polygonal rifling vs. conventional rifling), then these inconclusive responses would not be appropriate.[34] Discernable class characteristic differences provide compelling evidence of two different firearms.

SUMMARY AND DISCUSSION

No Evidence of Sample Swap

The committee found no evidence to support that CTS had inadvertently swapped samples, causing examiners to compare same-source samples when CTS intended for them to compare different-source samples. CTS described a robust quality control system which, if strictly followed, should effectively prevent sample mix-ups. It is also important to note that none of the examiners who came forward suggested this as a reason for their errors.

Cause of Examiner Error

The committee found no singular cause for those examiners who mis-identified items 2, 3, and/or 5 to Item 1. However, based on the interviews and survey responses, some important lessons and themes were revealed by individual situations.

First, multiple examiners reported allowing their identification conclusions to be influenced by factors outside of their analysis of the evidence. For example, Examiner #1 stated they had never reported inconclusive on any prior proficiency test over 20+ years, and thus may have put themselves under pressure to do the same for this test. Examiner #2 stated they believed inconclusive results were no longer allowed after a change to their quality assurance manual. Examiner #3 also felt that an inconclusive result was not allowed since they were taking a test. From the survey results, when concentrating on misidentifications of item 2: 8 of the 13 respondents (62%) reported not basing their conclusion solely on the microscopic analysis but instead they used "other reasoning", such as assuming an identification had to be present, or stating inconclusive was not allowed so they took their best guess. Examiners should be aware this type of reasoning is not a sound basis for an identification conclusion, and in this case it led to erroneous conclusions.

[34] The committee was not able to examine any of the item 4 bullets from the test sets in which item 4 was not eliminated and, therefore, is not able to directly comment on the discernibility of their class characteristics.

It is deeply concerning that at least one individual may have interpreted ANAB's requirement to report to the accrediting body any inconclusive results when the expected result is a definitive conclusion to mean that inconclusive results are not permissible.[35] ANAB has stated through their "Heads Up" communications that, while an inconclusive result may be unexpected, ANAB is not immediately judging that result to be inappropriate. Rather, ANAB expects the laboratory to evaluate the cause of the unexpected result and determine if any further action is warranted.[36] Additionally, for a consensus-based proficiency test, ANAB defines the consensus result as the expected result.[37] This, in and of itself, is problematic when the test materials exhibit a high degree of variation from test set to test set, as seen with the test under discussion here. However, for a consensus-based test, the consensus result cannot even be determined until after test participants submit their results and the results are tabulated. In the case of CTS Test 23-5262, "CTS determined that the assigned value for Items 2, 3, and 5 includes both elimination and inconclusive."[38]

The committee wants to be clear that inconclusive results should not always be accepted without question. Laboratories must evaluate the inconclusive results and determine whether or not they are appropriate given the condition of the test materials and in consideration of laboratory protocols. If determined not to be appropriate, corrective action should be taken.

Second, the committee noted that multiple examiners reported requiring unusually long amounts of time to complete this test. Examiner #1 stated this test took significantly longer than any other CTS tests they had taken. Examiner #2 spent over 32 hours, spanning 45 days, on this test. Examiner #3 spent several days working on the test, and with the deadline for submission approaching, felt time pressure to finish. A comparison that is taking significantly longer than normal can be a strong indicator that the data is limited and caution is, therefore, warranted. There is no set amount of time for how long a comparison may take, as that is dependent on an examiner's pace and the evidence before them. However, the time taken by these examiners could have served as a warning that the data before them was limited. At least for some (e.g., Examiners #2 and #3), this may have been exacerbated by their understanding of their laboratory policy regarding inconclusive results.

---

[35] ANAB's document AR3125, Accreditation Requirements for Forensic Testing and Calibration (Effective 02/01/2023), section 7.7.5: "The process for monitoring the performance of the laboratory and personnel shall: … f) require notification to ANAB within 30 days when the expected results is not attained during any monitoring activity". Note 2 f) "When an identification or exclusion is the expected result, an outcome of inconclusive is considered an unexpected result." https://anab.qualtraxcloud.com/ShowDocument.aspx?ID=12371 See p.13

[36] ANAB Heads UP Communication 2303 (https://anab.qualtraxcloud.com/ShowDocument.aspx?ID=29808) and 2304 (https://anab.qualtraxcloud.com/ShowDocument.aspx?ID=30248)

[37] ANAB AR3125, section 7.7.5, Note 2 f) "For a consensus-based proficiency test, the consensus result is the expected result."

[38] Collaborate Testing Services, Inc. Firearms Examination Test No 23-5262 Summary Report, July 19, 2024, p.2. https://cts-forensics.com/reports/23-5262_Web.pdf

Finally, it appears that some examiners erred when they conflated lack of *disagreement* with significant/sufficient *agreement*. The specimens examined by the committee, as well as the documentation provided, suggest that examiners may have mistaken relatively few striated marks in agreement as being significant because there were few striated marks to begin with. However, in actuality, this low level of agreement was simply coincidental.

The committee recognizes that examiners made errors when taking this test, and nothing stated above should be taken as justification for these errors. However, the committee believes that the factors described above likely contributed to the errors.

Use of CTS Results as an Industry-Wide Error Rate

As demonstrated by the differences in error rates between different demographic groups (see Tables 2 and 3), the PTRC strongly recommends against calculating an overall false positive error rate from this test as being reflective of the discipline as a whole or reflective of any individual examiner. As noted by the National Academy of Sciences (NAS): "The risk of error is properly considered case by case, taking into account the record of the laboratory performing the tests, the extent of redundancy, and the overall quality of the results."[39] Most laboratories (~80%) did not report false identification conclusions for this test. Additionally, laboratories that did report false identification conclusions were notified of the result and should undergo root cause analysis and take corrective action in order to mitigate the chance of this occurring again. The NAS recognizes the value of forensic laboratory quality assurance and why using historical proficiency test results is inappropriate:

> "Estimating rates at which nonmatching (sic) samples are declared to match from historical performance on proficiency tests is almost certain to yield wrong values. When errors are discovered, they are investigated thoroughly so that corrections can be made. A laboratory is not likely to make the same error again, so the error probability is correspondingly reduced"[40]

RECOMMENDATIONS

1) Laboratories that reported one or more false identifications on this test should conduct a thorough root cause analysis. This will likely result in additional training for their examiners. However, this should be seen as a valuable learning opportunity. The awareness gained from this experience can greatly reduce the chance of similar occurrences in the future, resulting in better trained examiners who are less likely to err.[41]

[39] National Research Council "The Evaluation of Forensic DNA Evidence" National Academy Press, Washington DC, 1996. pg 87.
[40] Ibid, pg 86.
[41] Ibid, pg 86.

2) Laboratories should clearly indicate to their examiners, either via policy or some other means, that the entire range of conclusions used in casework is available when taking a proficiency test. Specifically, inconclusive is an appropriate response when the observations do not support a definitive conclusion, even though the ground truth is known to the test preparer. The three-dimensional scan data (in x3p file format) of the test sets evaluated by the committee should be made available to the community for further evaluation and research. AFTE may consider hosting the data on the AFTE website or reaching an agreement with a third-party to host the data.
3) Laboratories should ensure that comparison sets included in new examiner training programs contain sufficient variety and complexity of specimens. This should include specimens that have a low quantity of striated marks, similar to the bullets in this proficiency test. These same types of specimens should also be a part of ongoing training and professional development for experienced examiners.
4) CTS should reevaluate their internal quality control procedures for the production of firearm examination proficiency tests. Specifically, the process for verifying the quality and consistency of fired ammunition components during test material production should be more structured and robust. Additionally, when predistribution test responses include inconclusive results when the ground truth is different-source, CTS should more thoroughly investigate the reasons and take appropriate remedial action.[42] The committee is not suggesting that proficiency tests should be "easier". On the contrary, proficiency tests should be challenging, but because they are interlaboratory comparisons, the test materials need to be as consistent as possible from test set to test set.

LIMITATIONS

This report contains the opinions of the committee and is based on the information available as of the date of the report. If additional information becomes available these opinions may be subject to revision.

ACKNOWLEDGEMENTS

The committee would like to thank Dr. Pate Skene of the University of Colorado Boulder for his assistance with developing the survey. The committee would also like to thank Cathy Brown and Nicole Shields of Collaborative Testing Service (CTS) for their cooperation and willingness to provide information relevant to the PTRC's investigation.

[42] In the 5-page Attachment to CTS 23-5262 Summary Report (https://cts-forensics.com/reports/23-5262_Web.pdf.), page 5, CTS has already committed to implementing a more robust quality control process.

And finally, the committee would like to thank those test participants who came forward to discuss their responses to the test. They provided invaluable insight and were the brave few who stepped forward when most did not. The profession owes them thanks and gratitude.

Appendix A

Comparison Images from Examiners 1 through 4

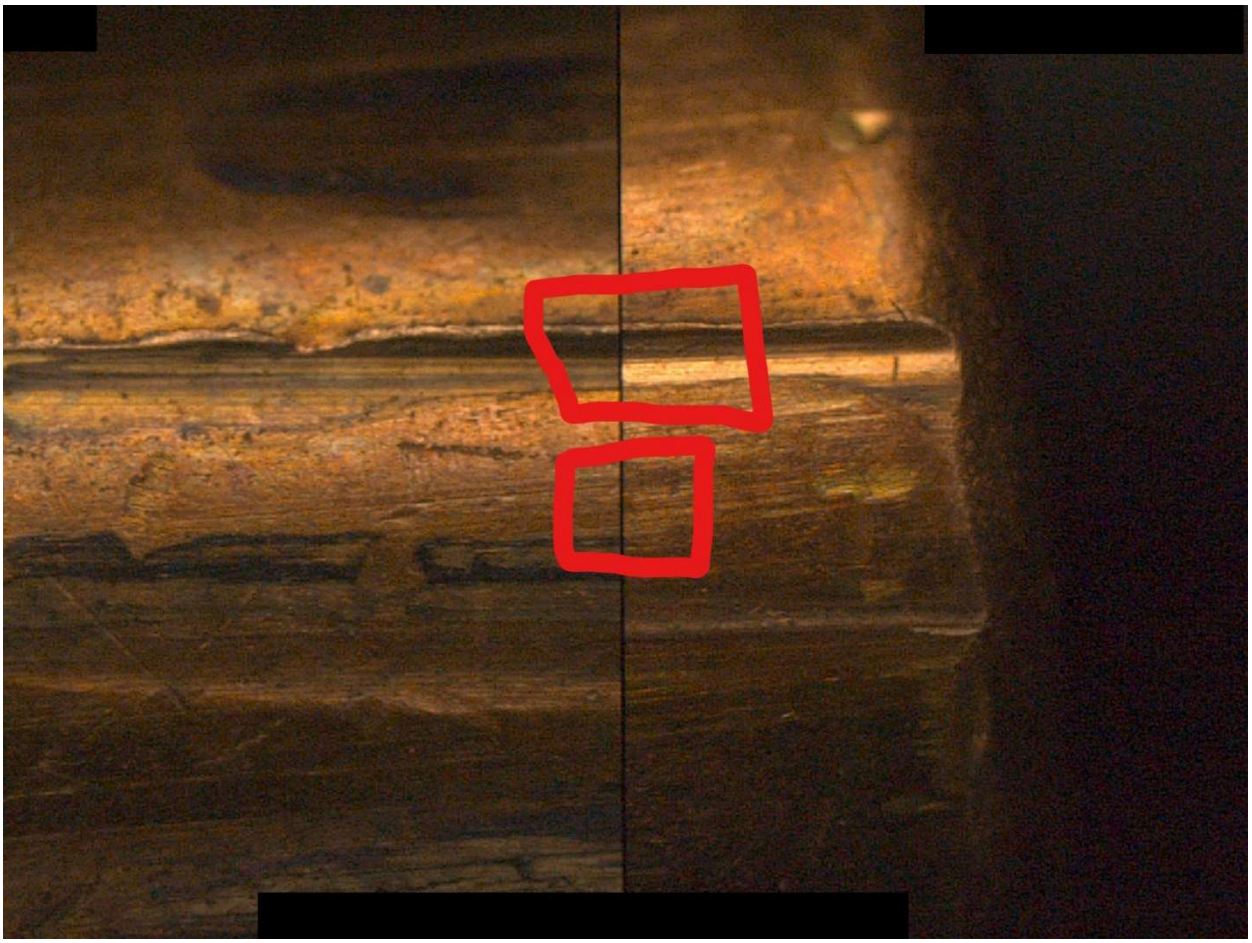


Figure A1: Image from Examiner #1. Item 1 (left) vs. Item 5 (right). Identifiers were redacted, but the red annotations placed by the examiner were left in place.

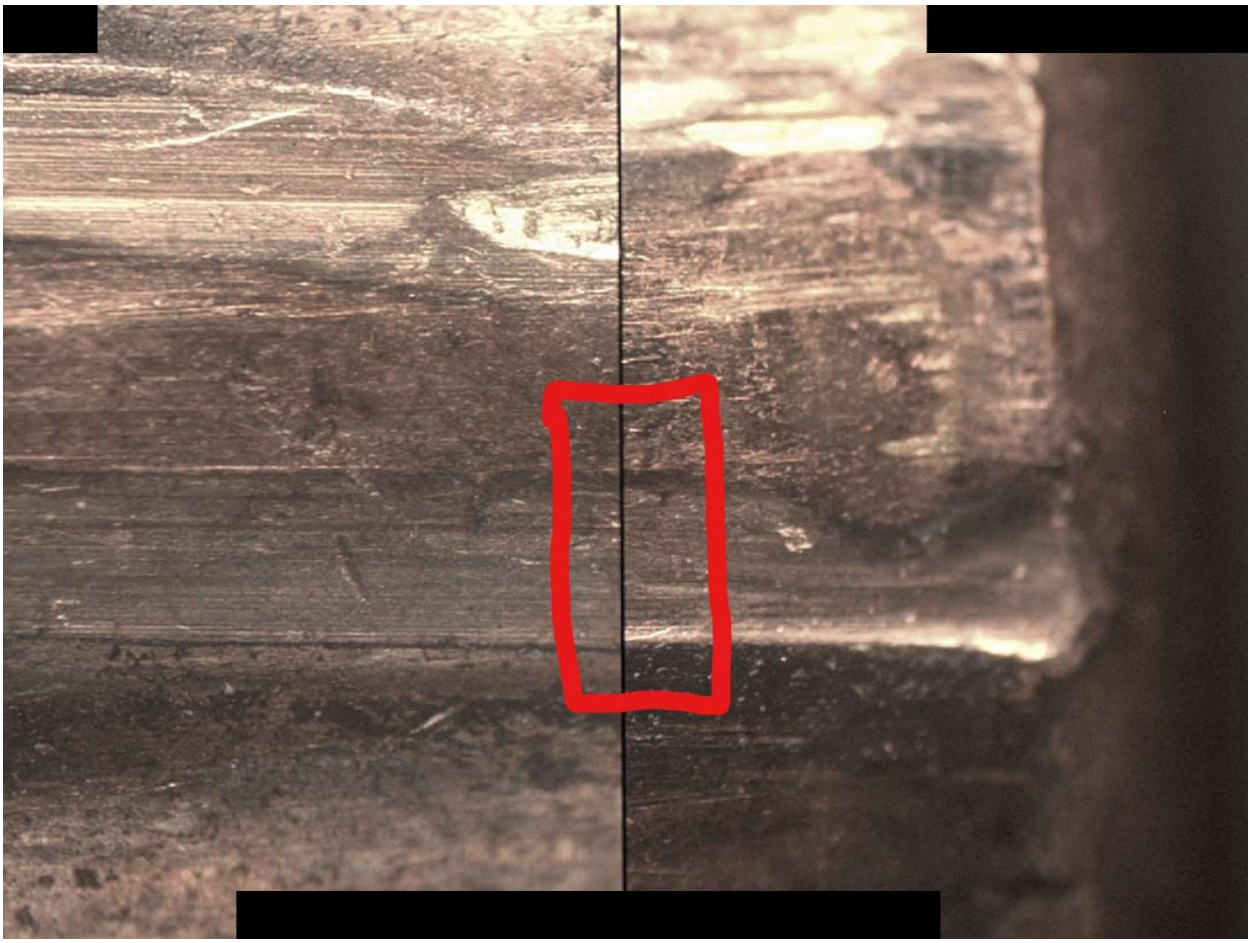


Figure A2: Image from Examiner #1. Item 1 (left) vs. Item 5 (right). Identifiers were redacted, but the red annotations placed by the examiner were left in place.

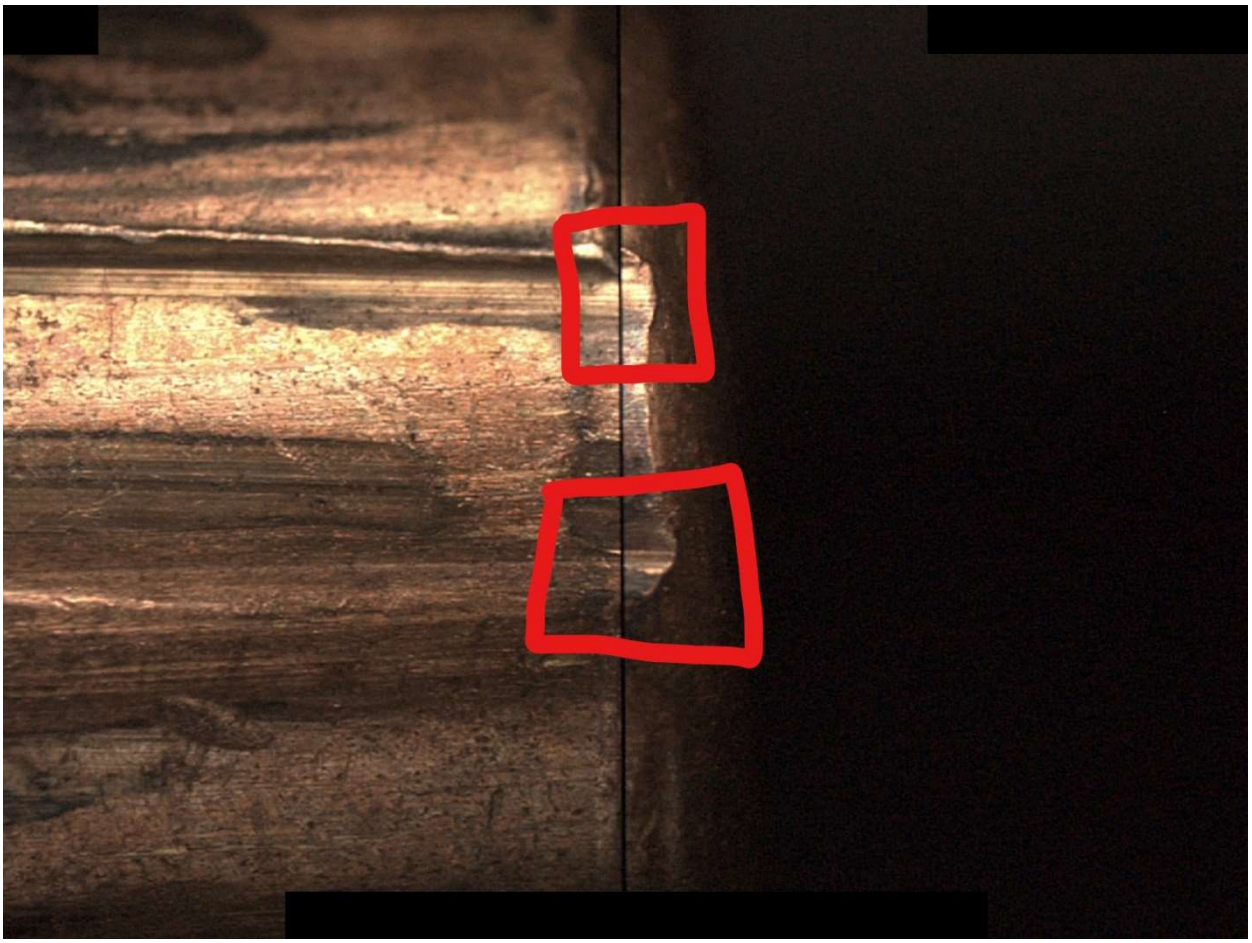


Figure A3: Image from Examiner #1. Item 1 (left) vs. Item 5 (right). Identifiers were redacted, but the red annotations placed by the examiner were left in place.

Figure A4: Image from Examiner #1. Item 1 (left) vs. Item 5 (right). Identifiers were redacted, but the red annotations placed by the examiner were left in place.

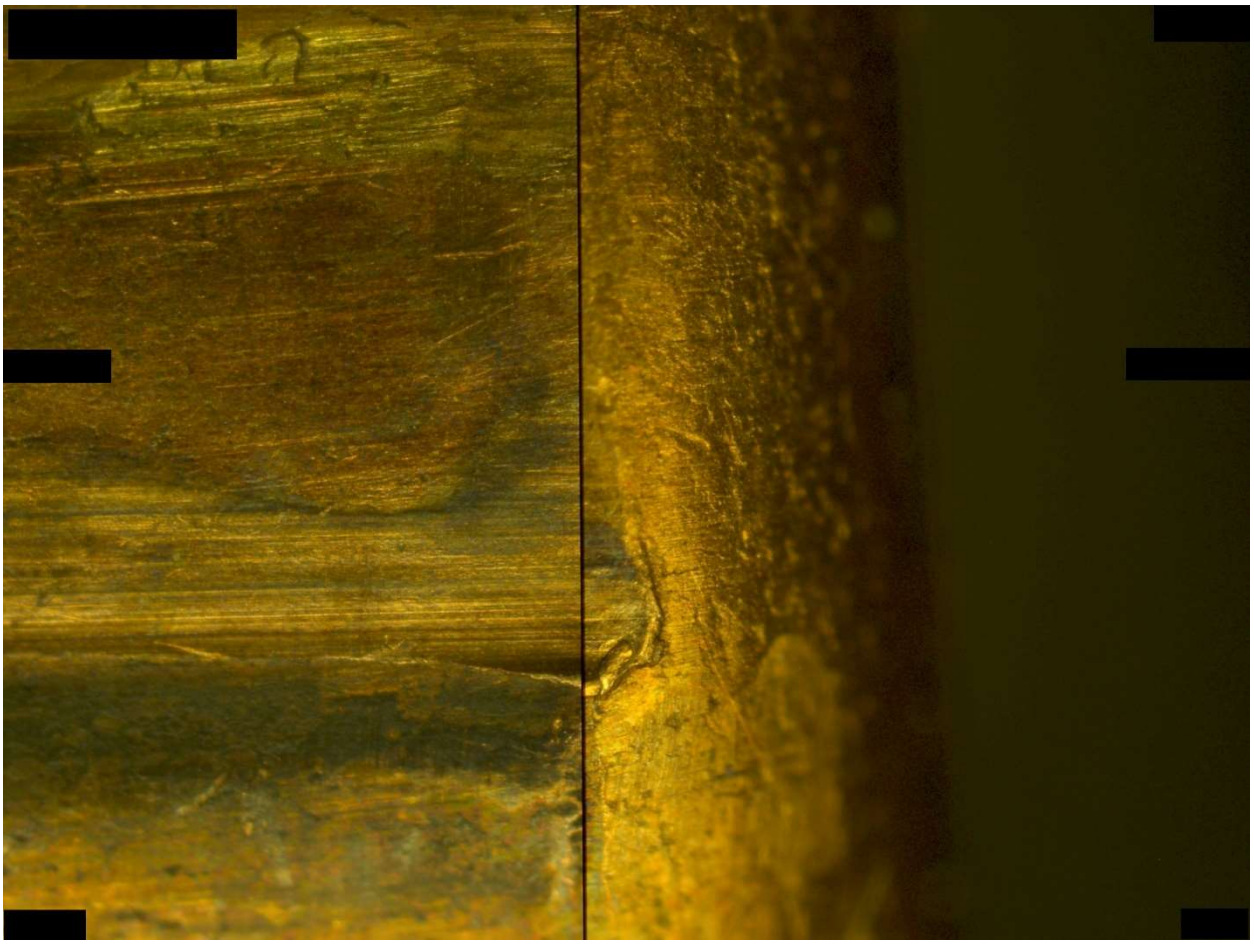


Figure A5: Image from Examiner #2. Item 1 (left) vs. Item 5 (right). Identifiers were redacted.

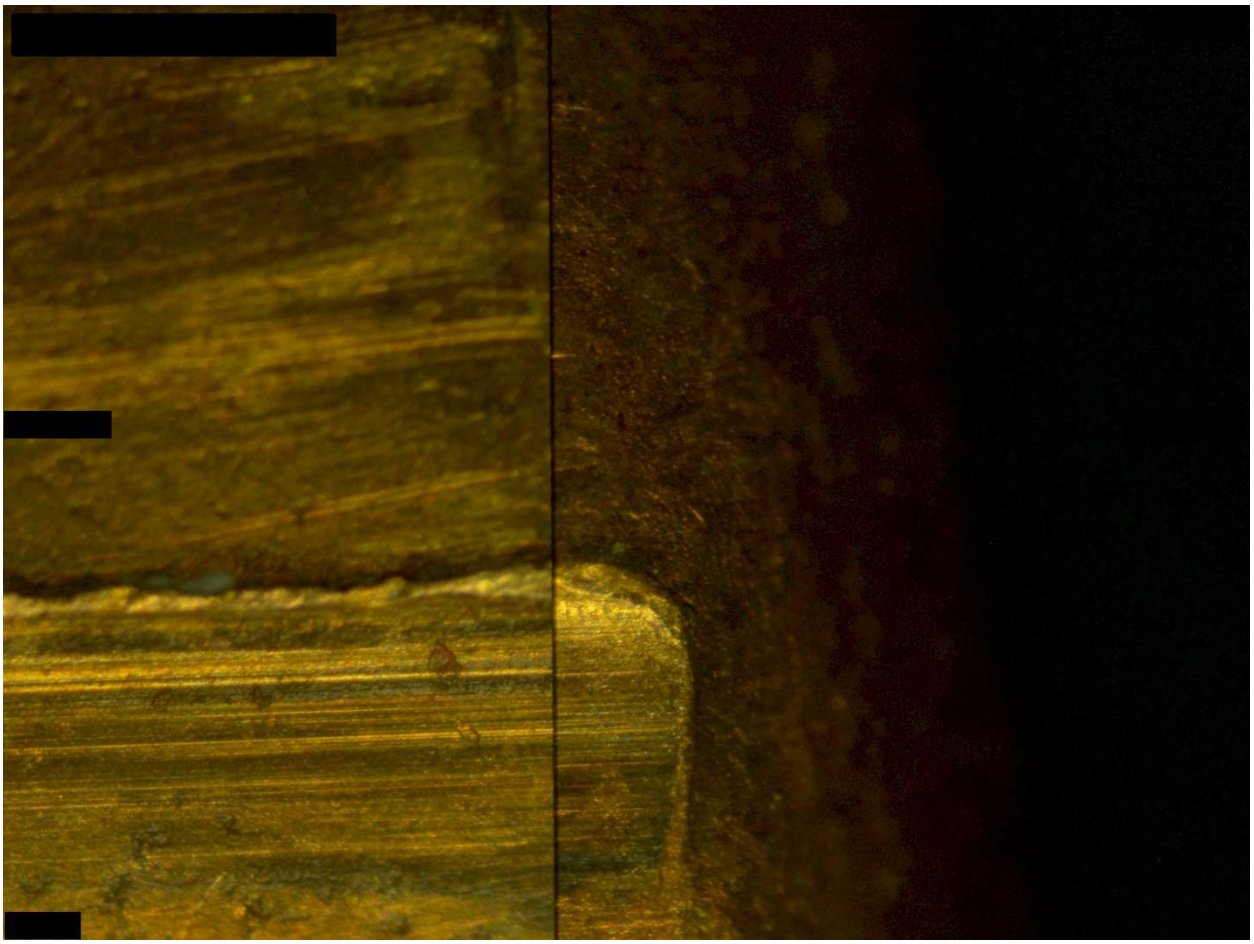


Figure A6: Image from Examiner #2. Item 1 (left) vs. Item 2 (right). Identifiers were redacted.

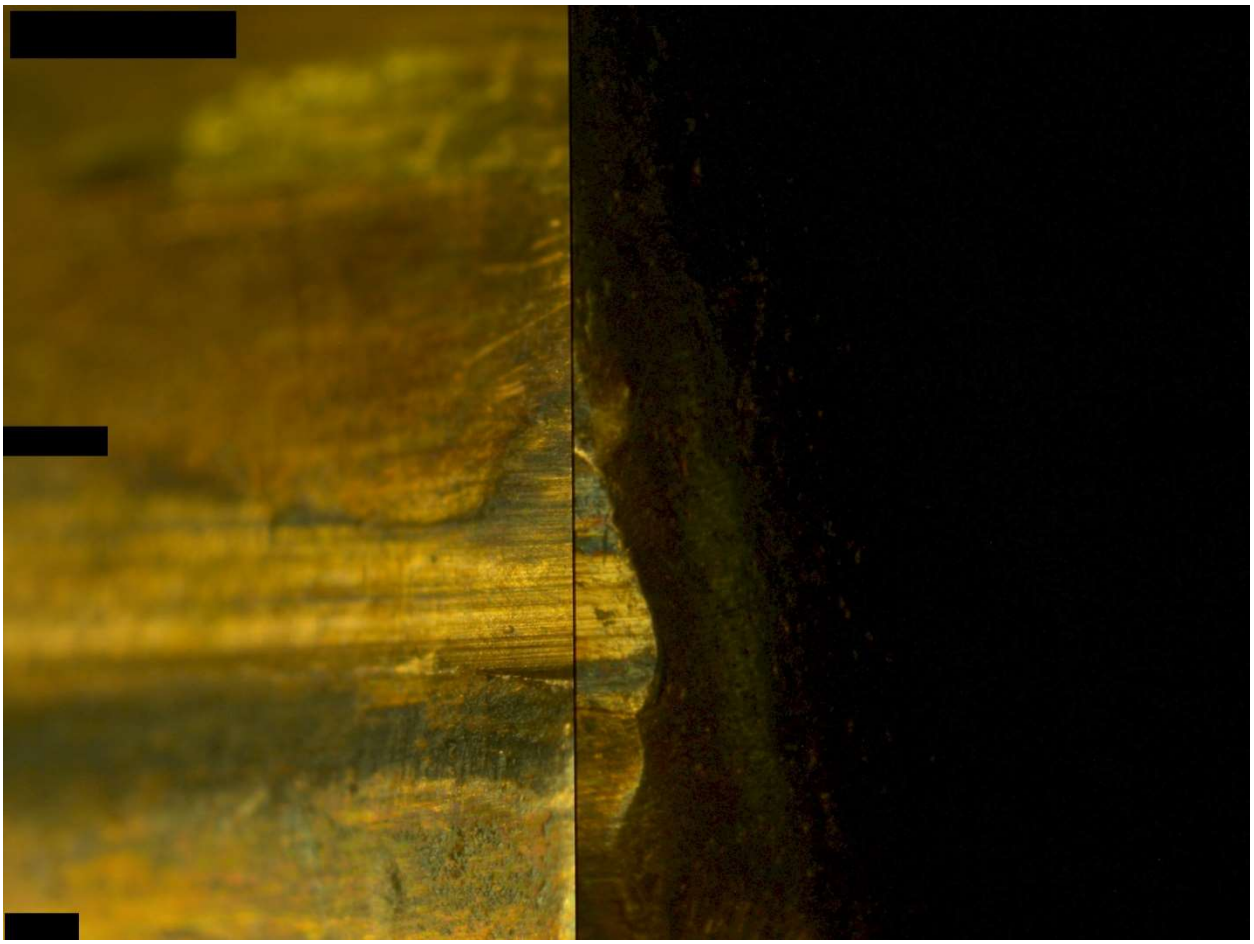


Figure A7: Image from Examiner #2. Item 1 (left) vs. Item 3 (right). Identifiers were redacted.

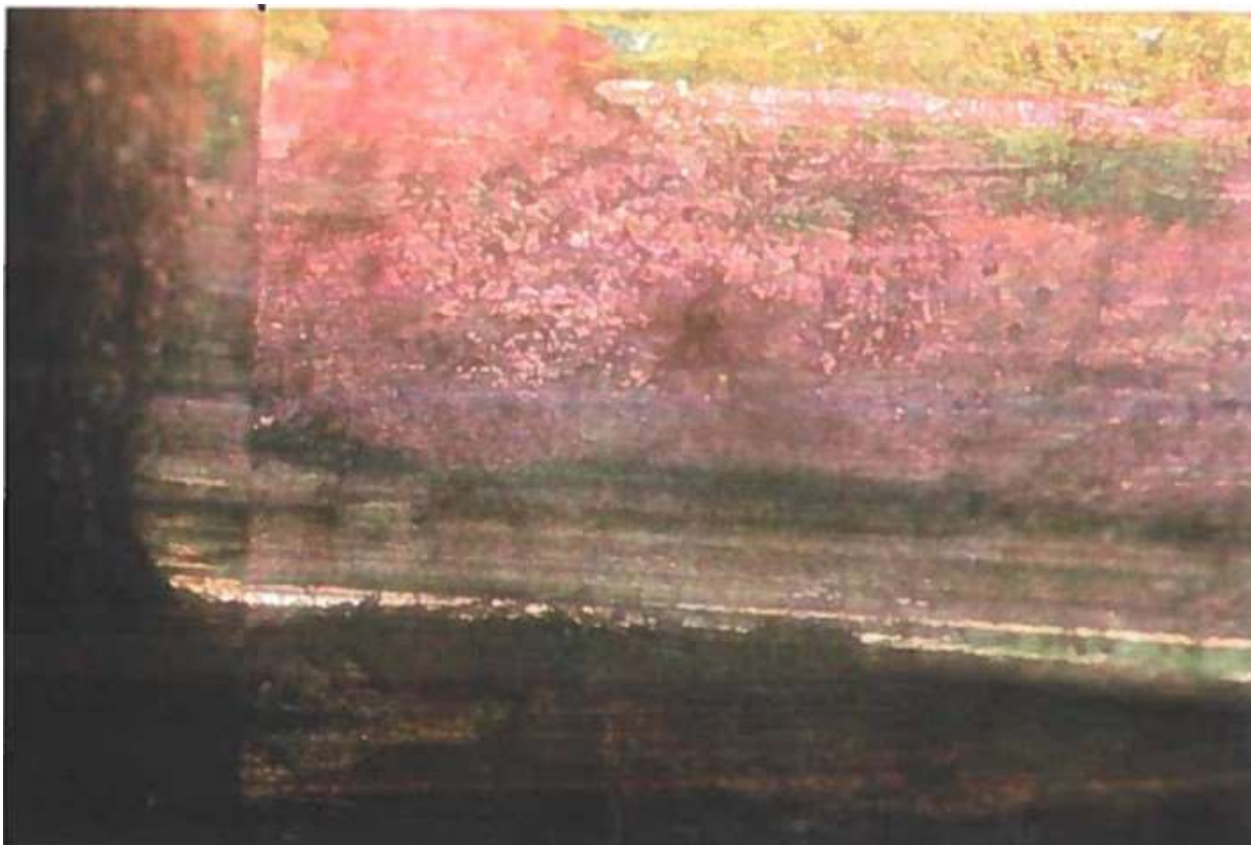


Figure A8: Image from Examiner #3. Item 3 (left) vs. Item 1, land impression 1 (right). 15x magnification.

Figure A9: Image from Examiner #3. Item 3 (left) vs. Item 1, land impression 1 (right). 20x magnification.

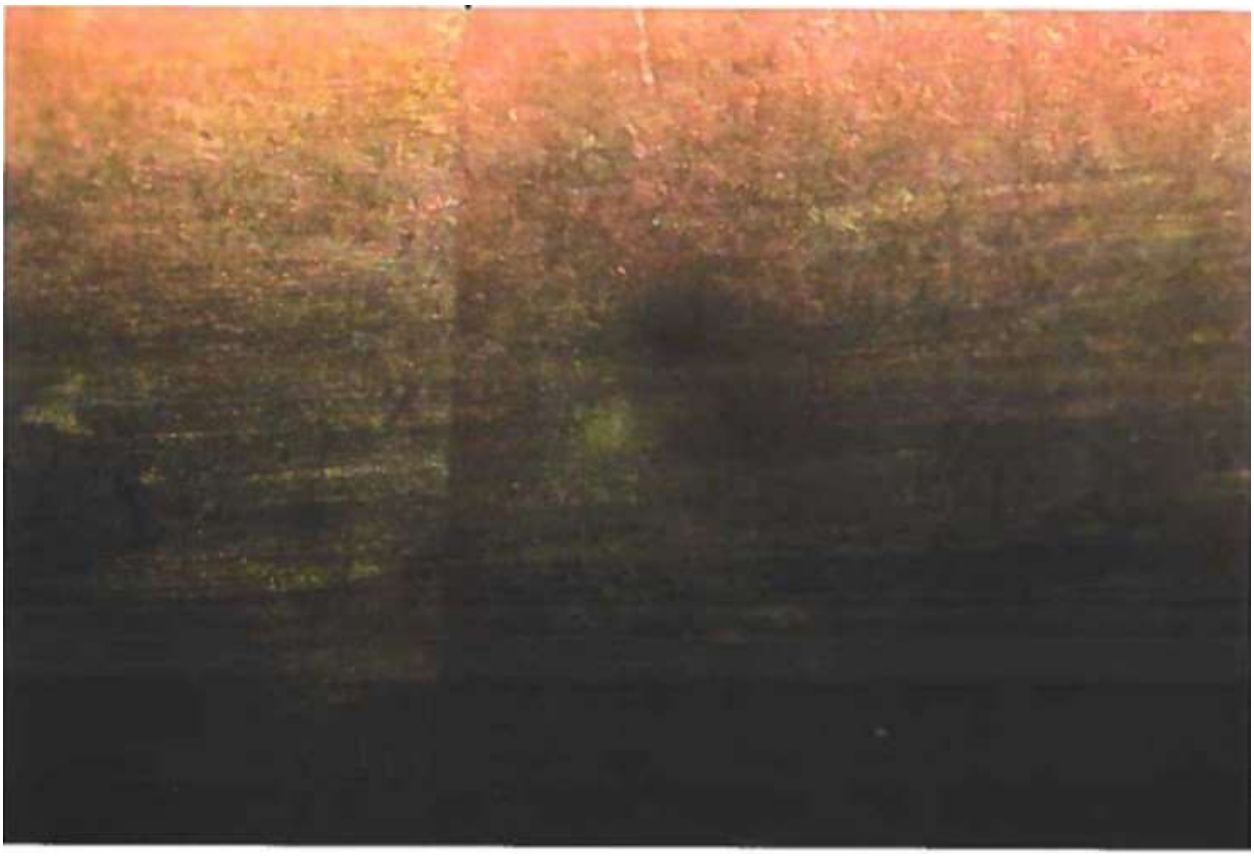


Figure A10: Image from Examiner #3. Item 3 (left) vs. Item 1, groove impression 1 (right). 20x magnification.

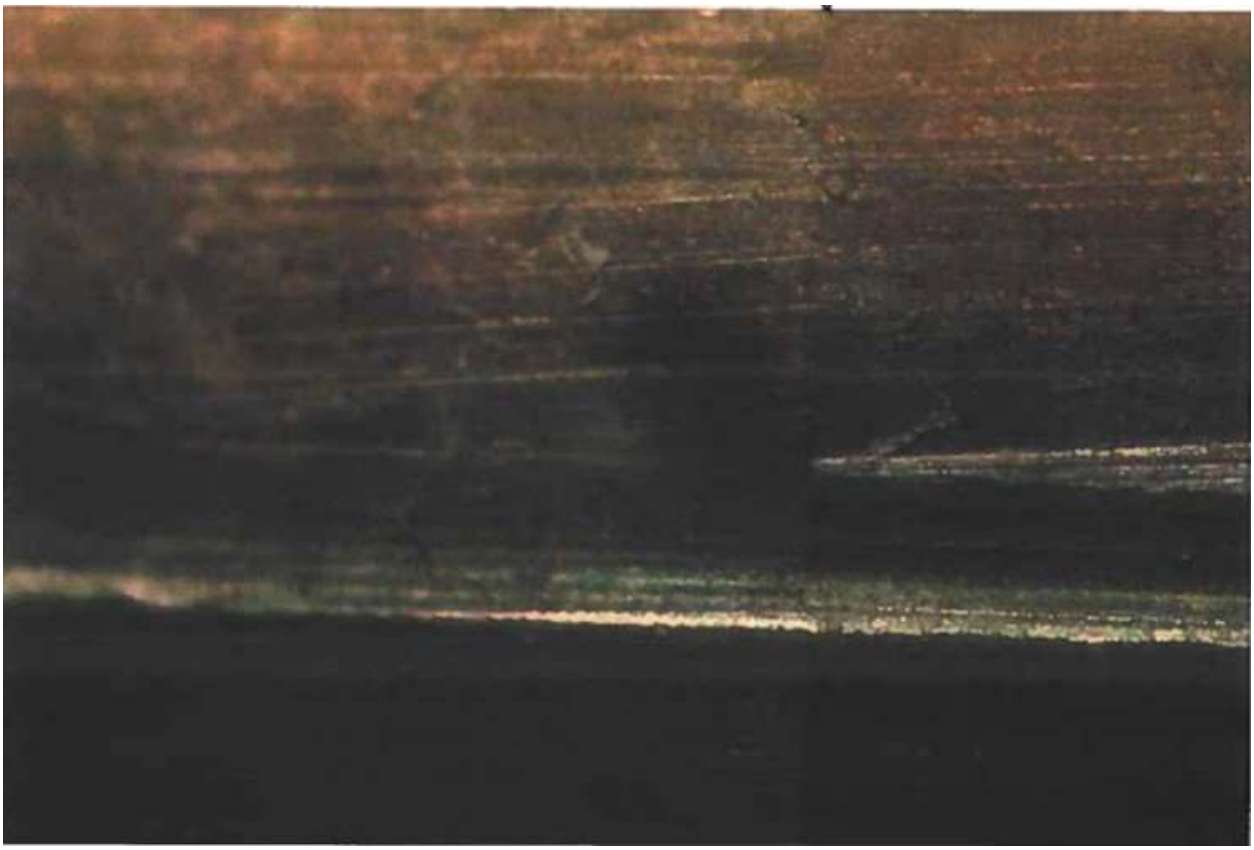


Figure A11: Image from Examiner #3. Item 3 (left) vs. Item 1, land impression 2 (right). 20x magnification.

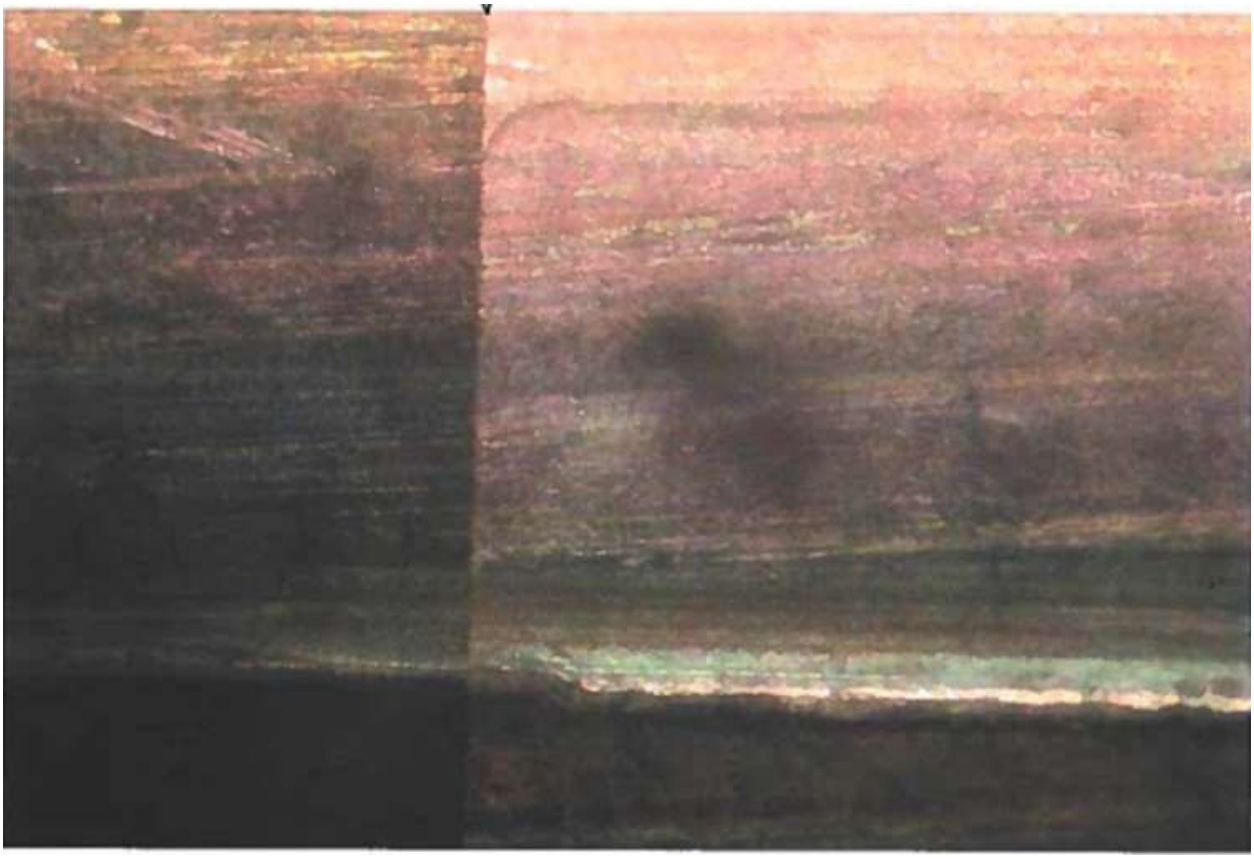


Figure A12: Image from Examiner #3. Item 2 (left) vs. Item 1, land impression 1 (right). 20x magnification.

Figure A13: Image from Examiner #3. Item 2 (left) vs. Item 1, groove impression 1 (right). 20x magnification.

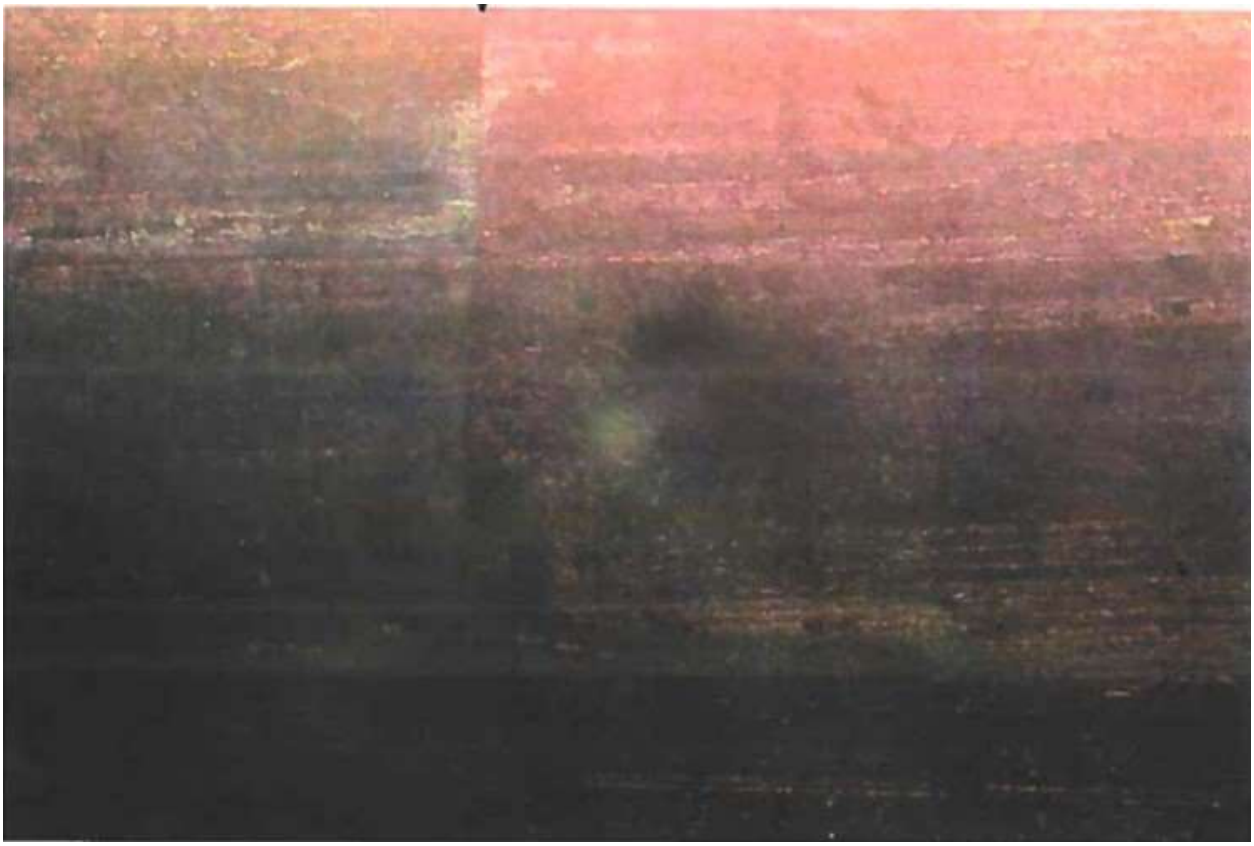


Figure A14: Image from Examiner #3. Item 2 (left) vs. Item 1, groove impression 2 (right). 20x magnification.

Figure A15: Image from Examiner #3. Item 5 (left) vs. Item 1, groove impression 1 (right). 20x magnification.

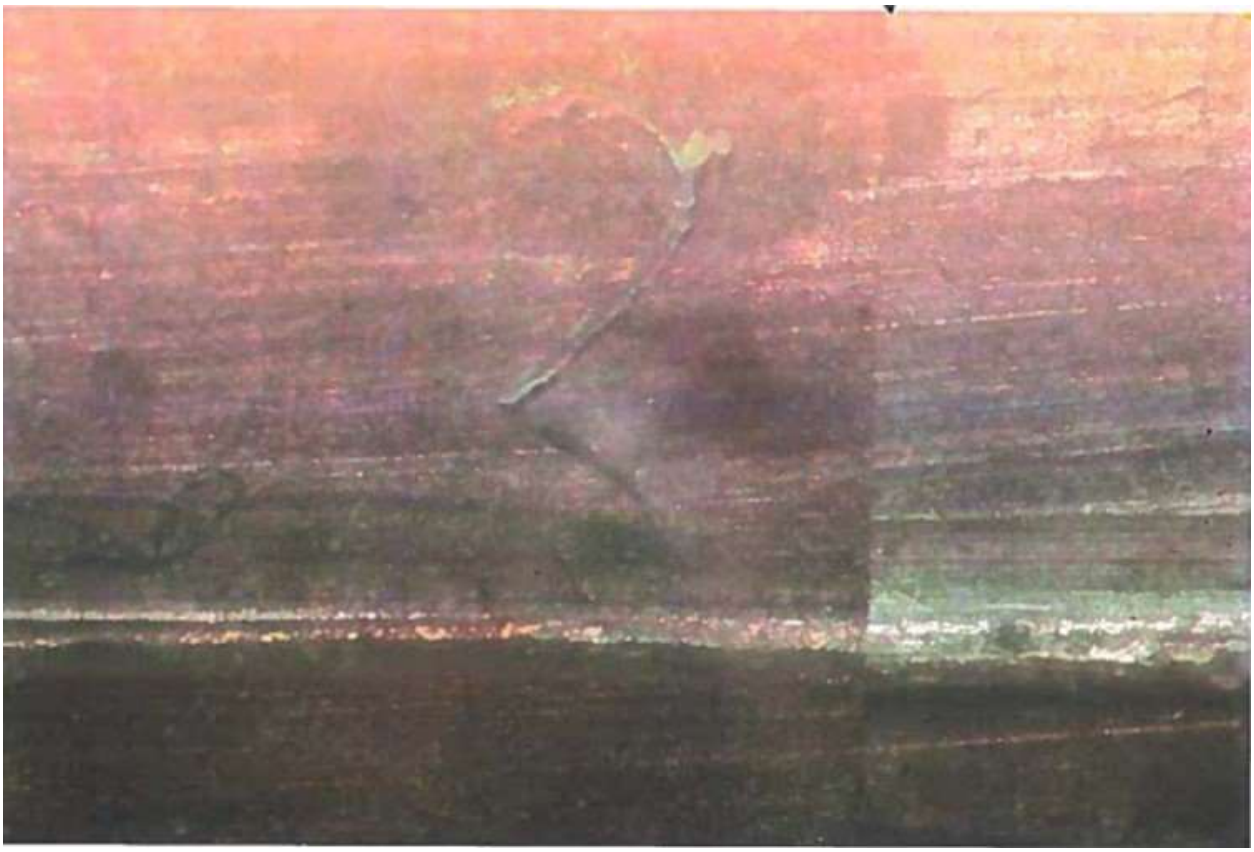


Figure A16: Image from Examiner #3. Item 5 (left) vs. Item 1, land impression 1 (right). 20x magnification.

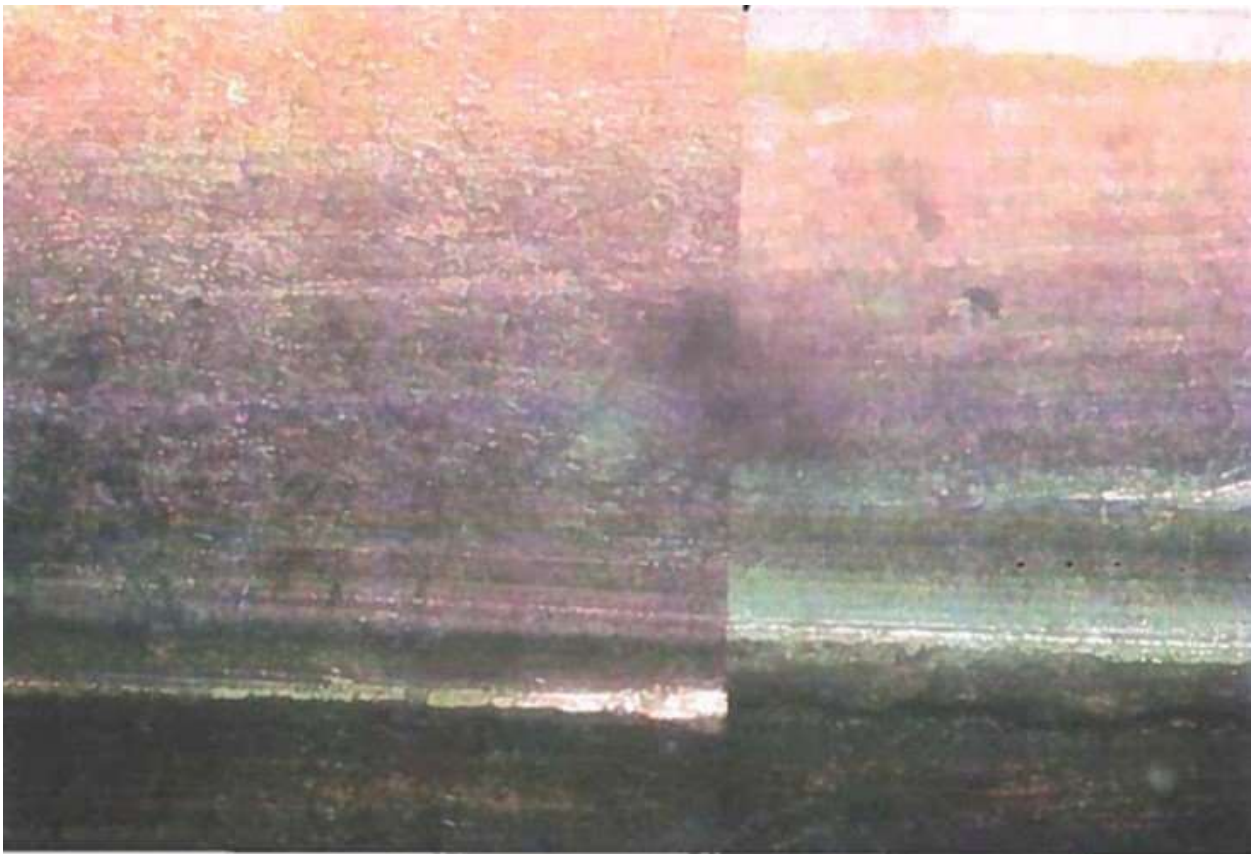


Figure A17: Image from Examiner #3. Item 5 (left) vs. Item 1, land impression 2 (right). 15x magnification.

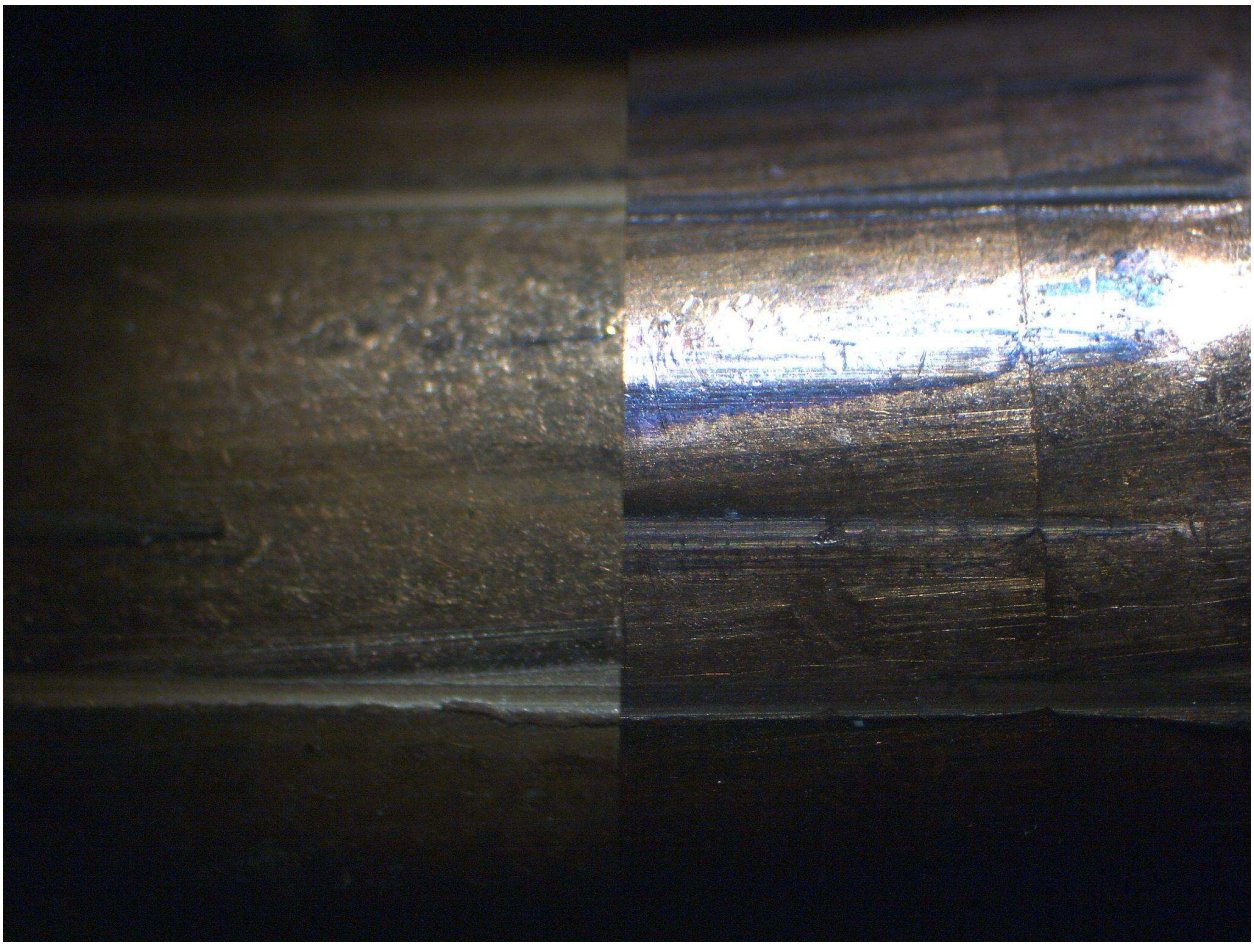


Figure A18: Image from Examiner #4. Item 1 (left) vs. Item 5 (right).

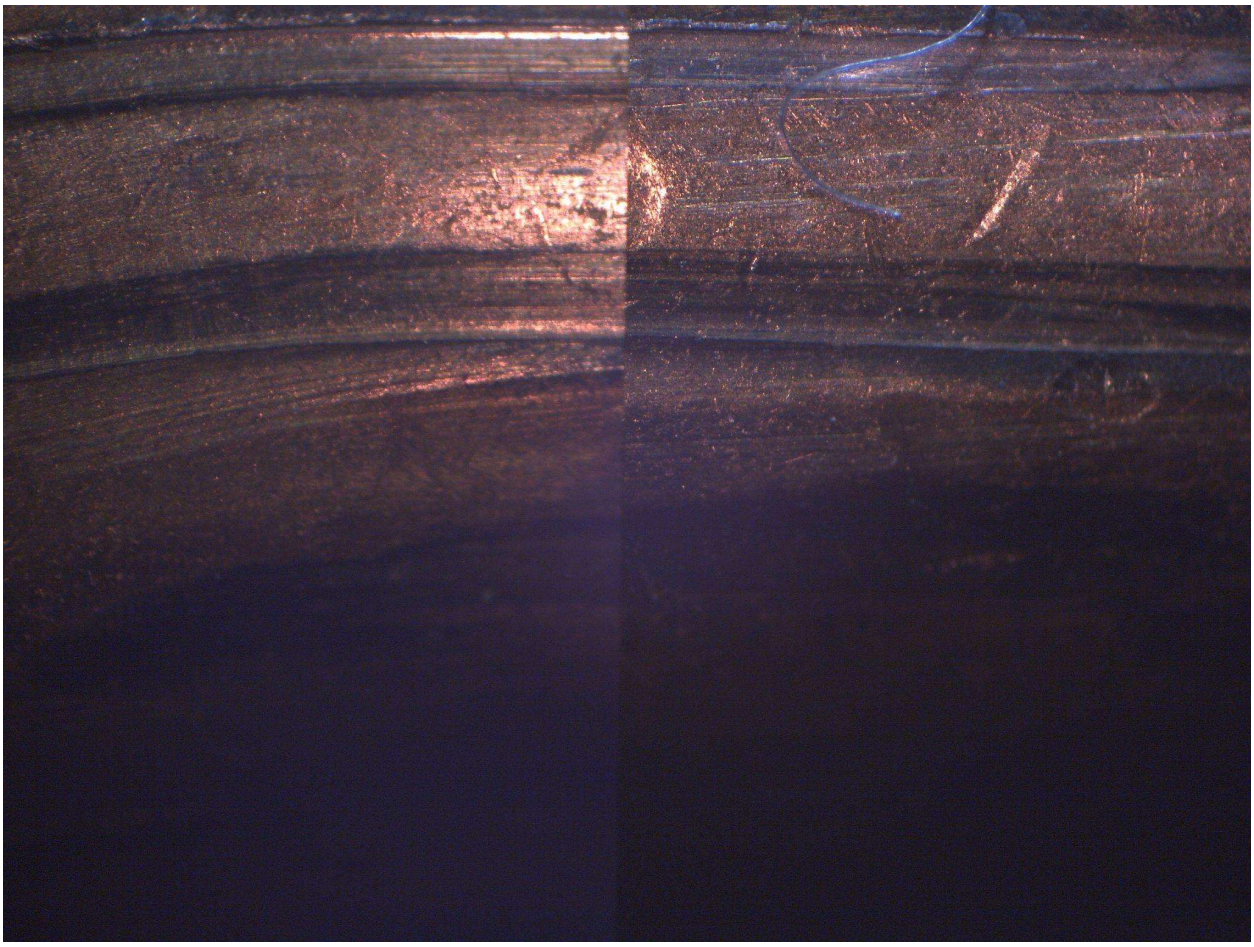


Figure A19: Image from Examiner #4. Item 3 (left) vs. Item 5 (right).

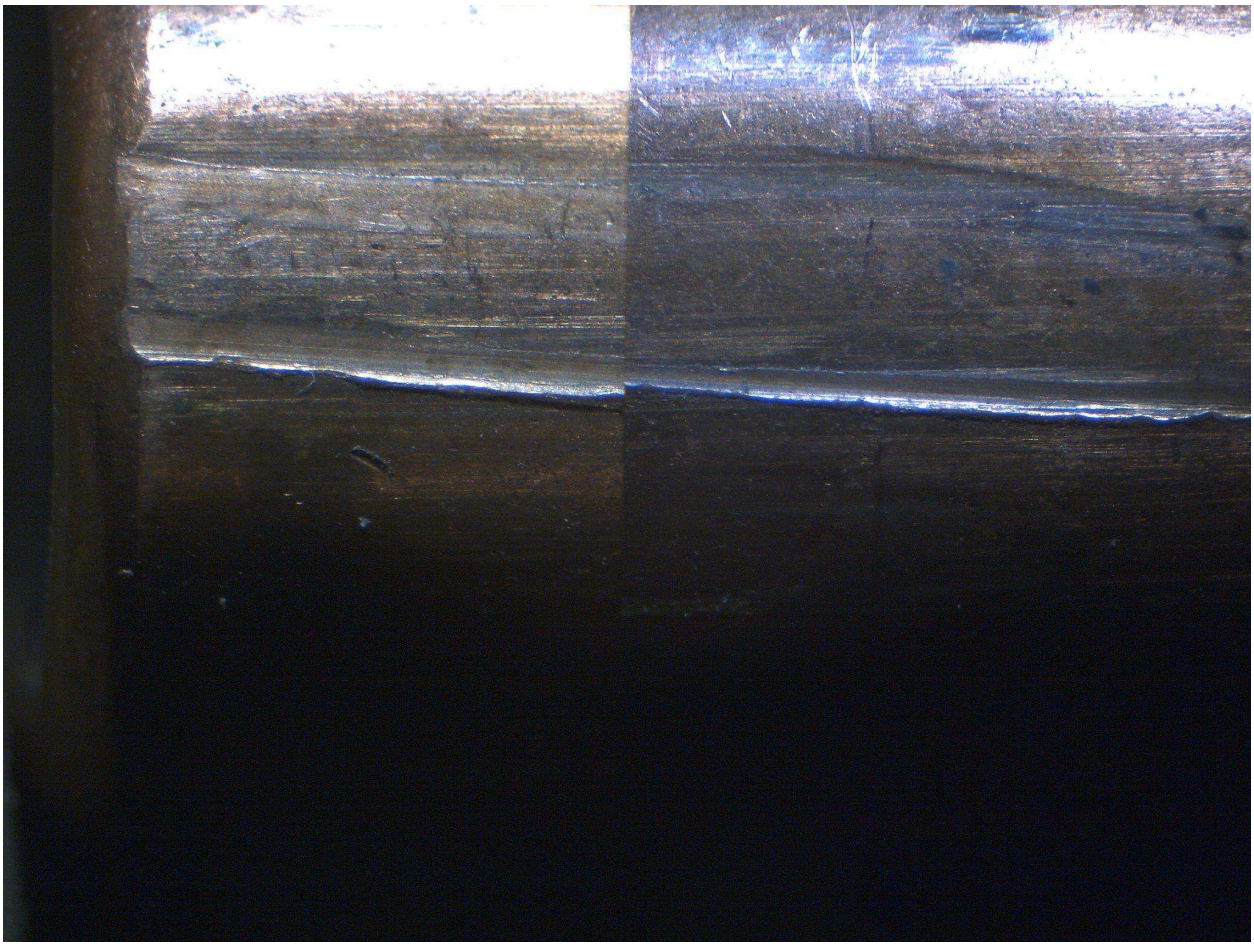


Figure A20: Image from Examiner #4. Item 1 (left) vs. Item 1 (right).

Figure A21: Image from Examiner #4. Item 1 (left) vs. Item 2 (right).

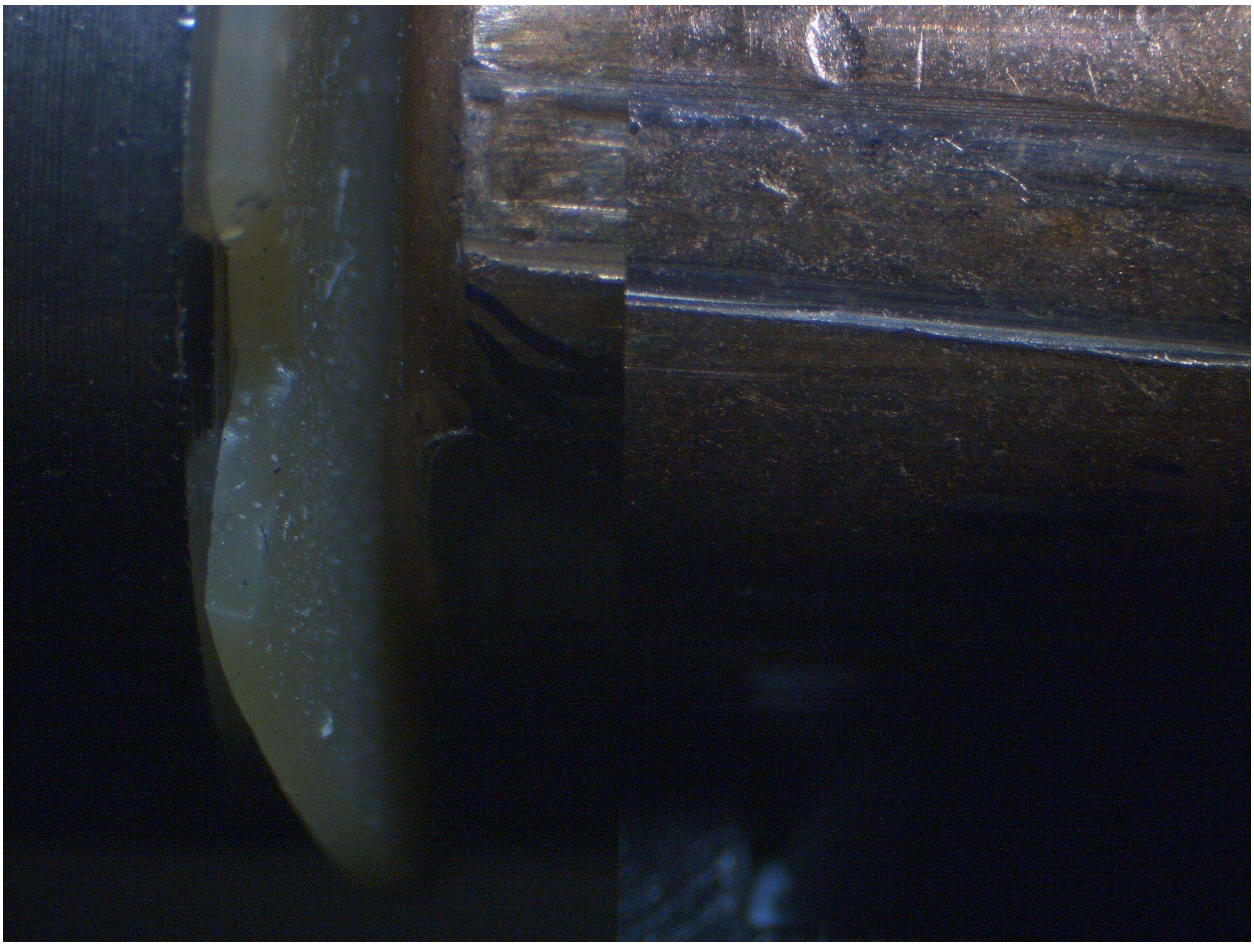


Figure A22: Image from Examiner #4. Item 1 (left) vs. Item 3 (right).

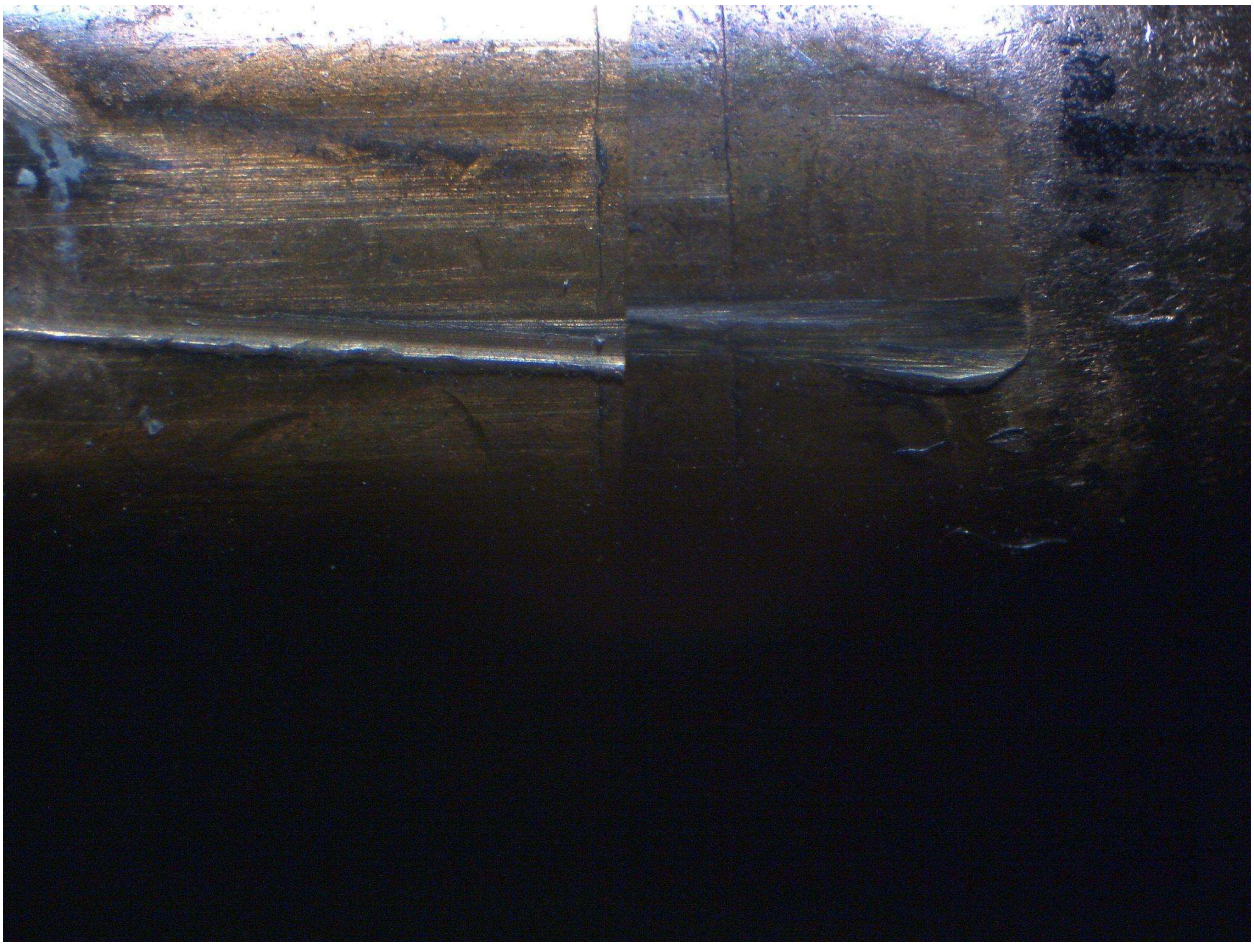


Figure A23: Image from Examiner #4. Item 1 (left) vs. Item 1 (right).

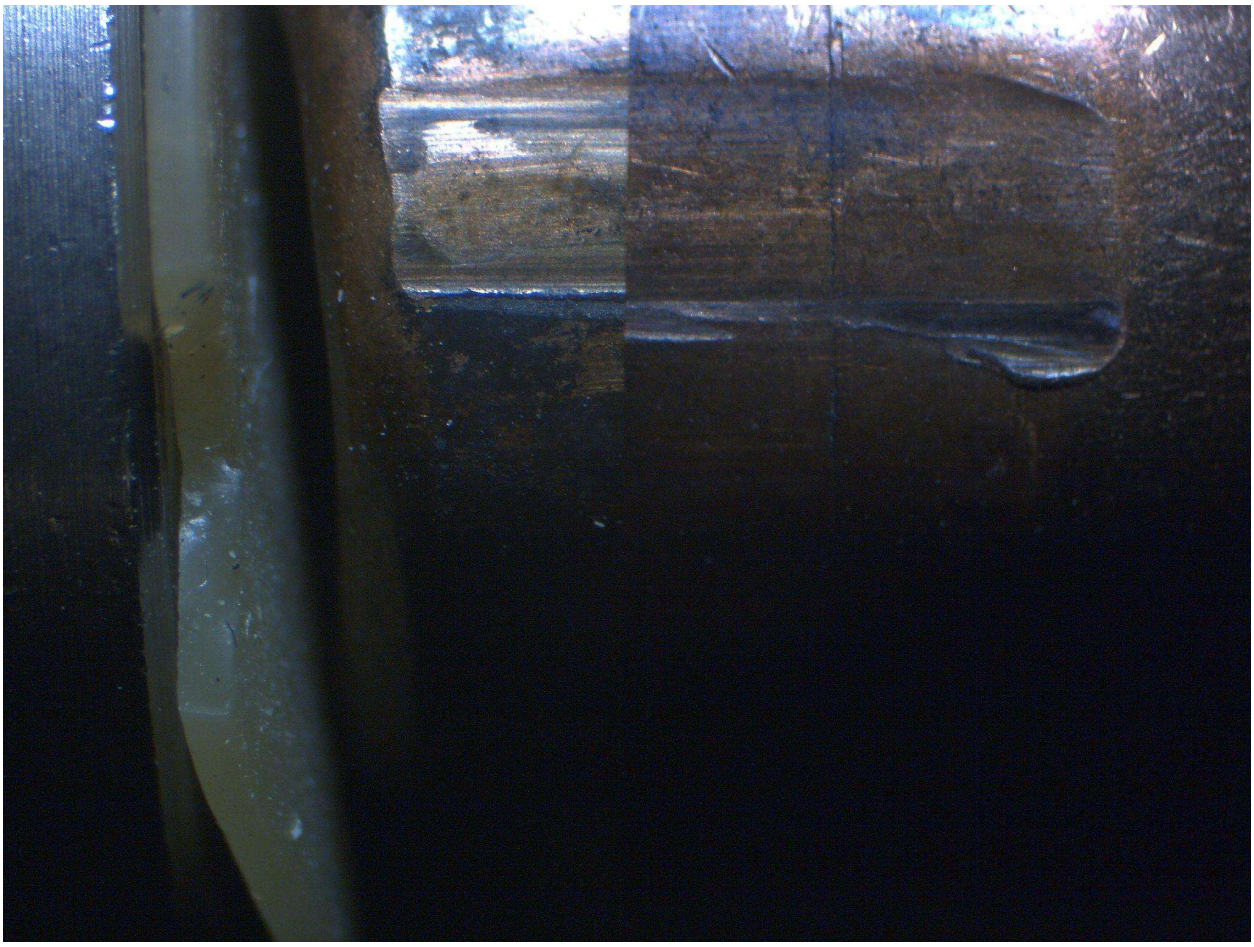


Figure A24: Image from Examiner #4. Item 1 (left) vs. Item 1 (right).

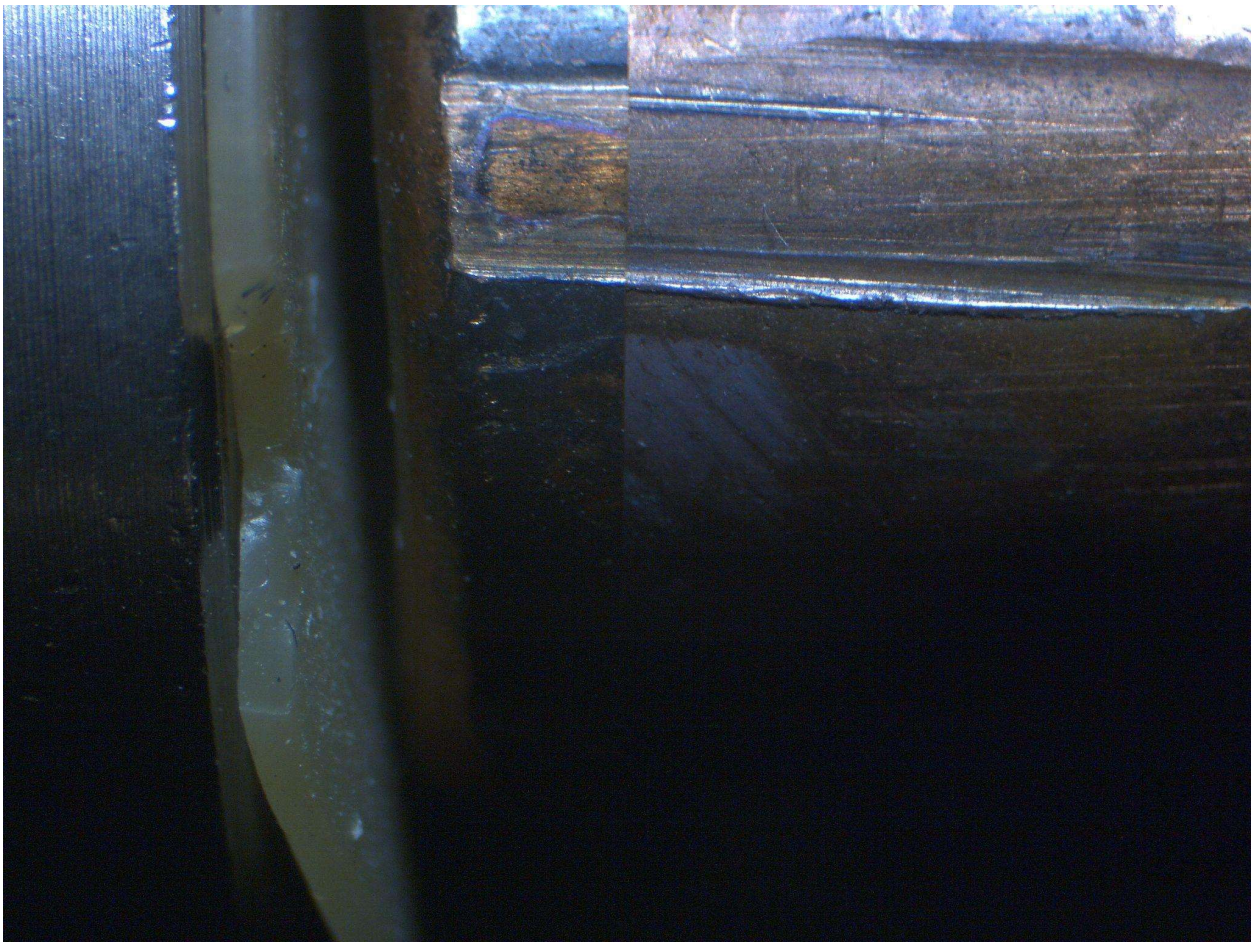


Figure A25: Image from Examiner #4. Item 1 (left) vs. Item 1 (right).

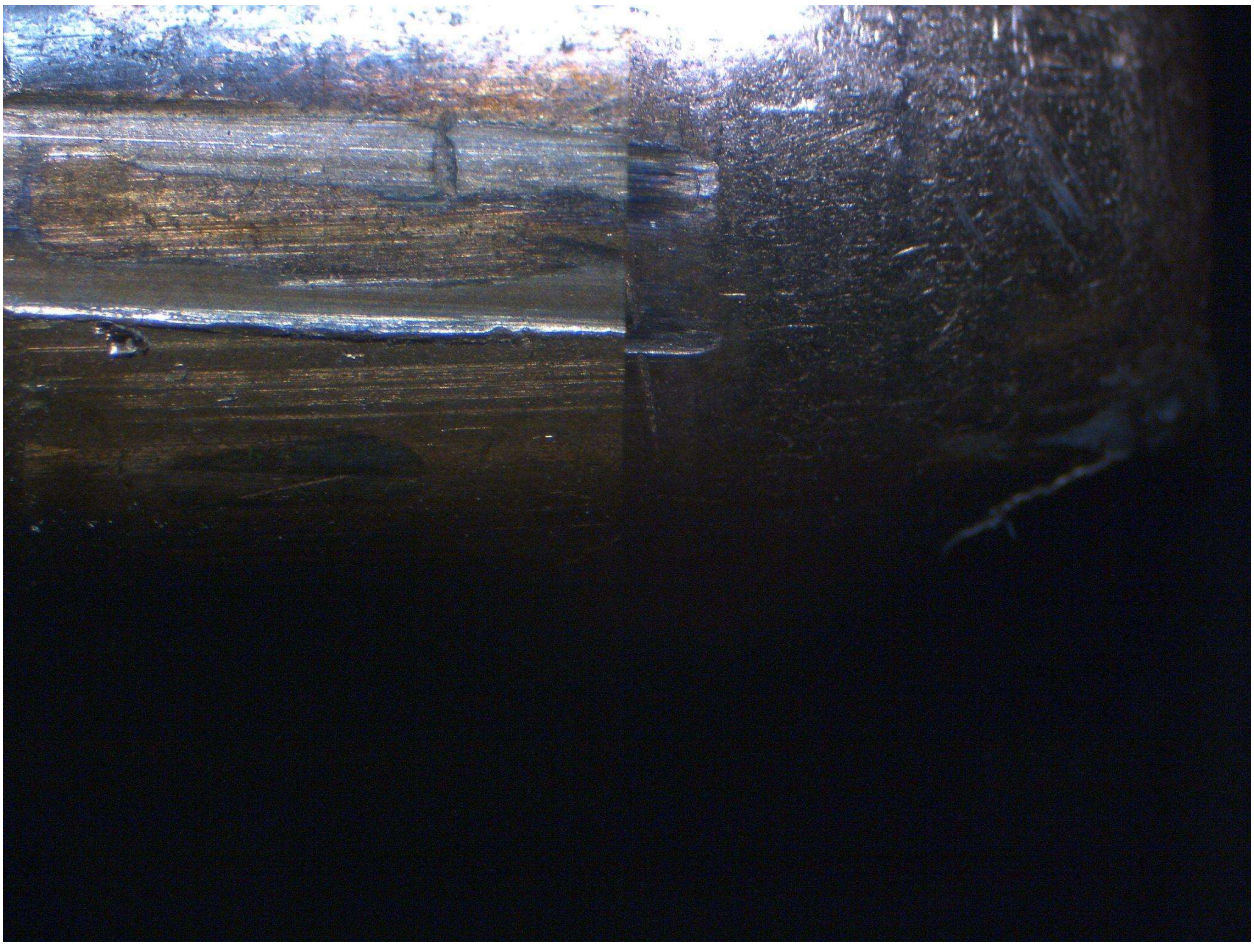


Figure A26: Image from Examiner #4. Item 1 (left) vs. Item 1 (right).

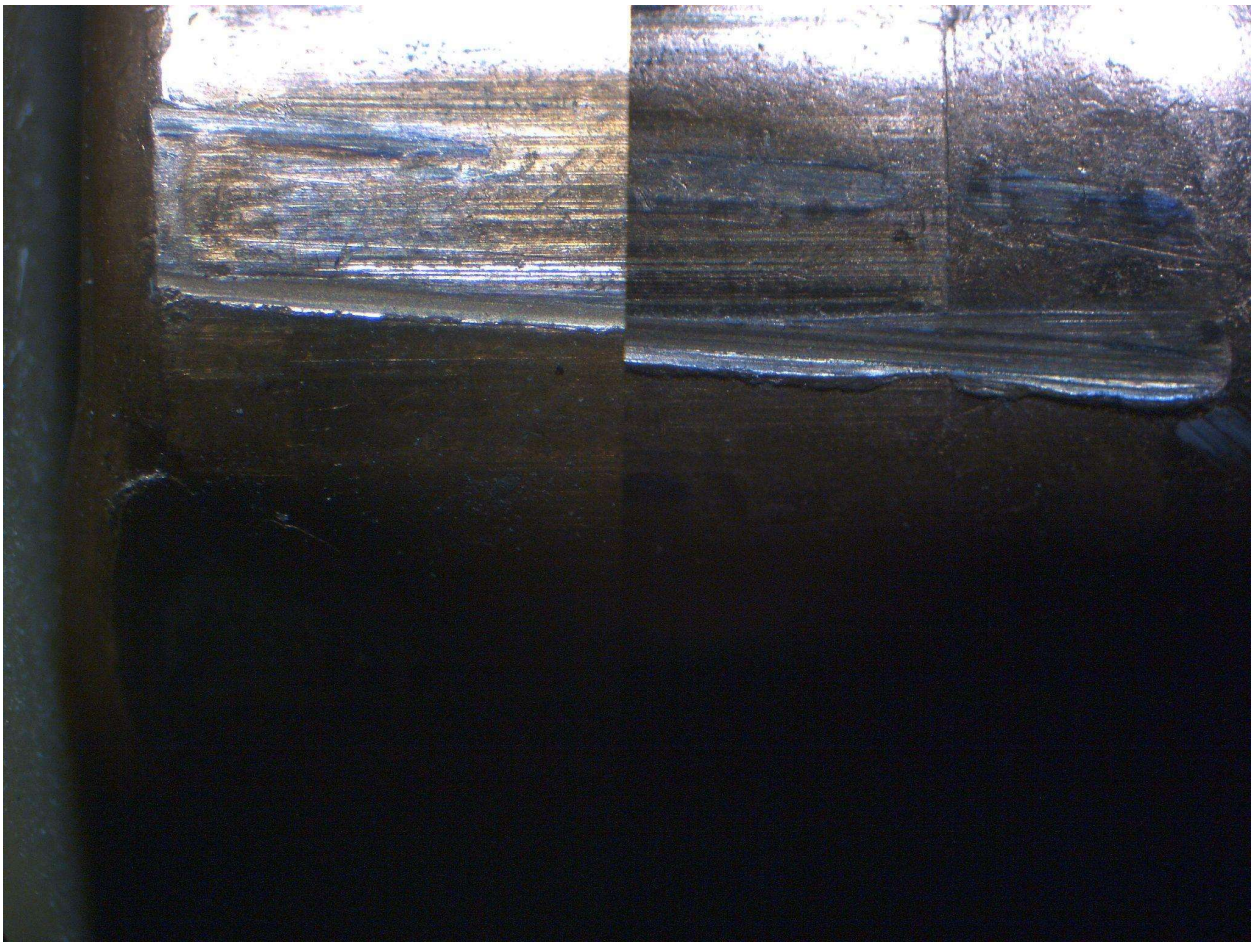


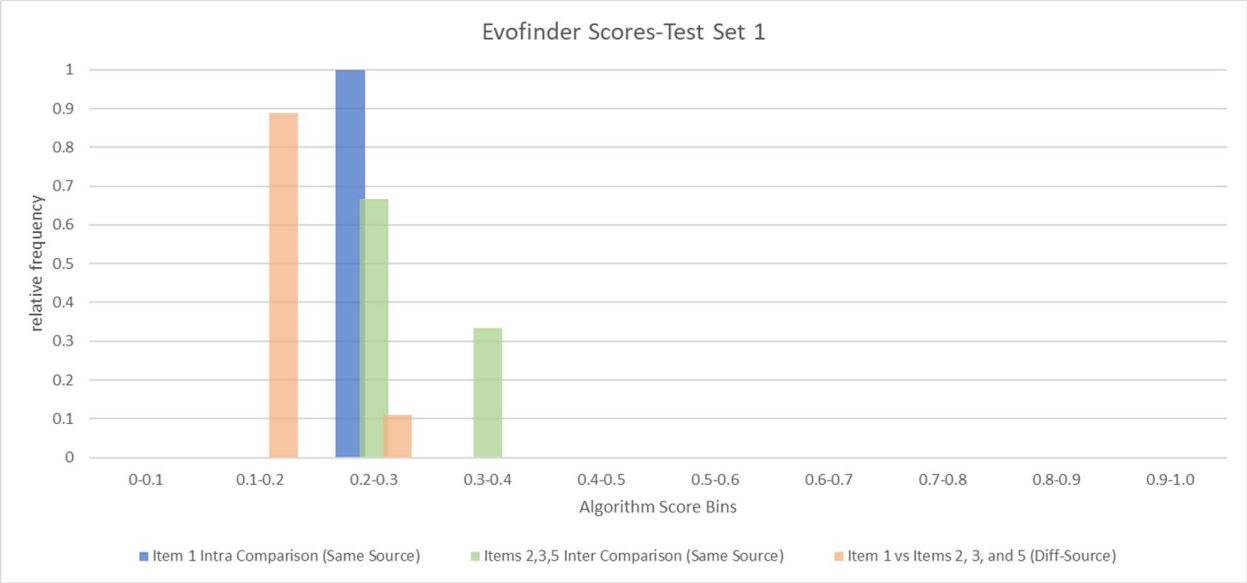
Figure A27: Image from Examiner #4. Item 1 (left) vs. Item 1 (right).

Appendix B

Relative Frequency of Evofinder Algorithm Scores from 10 CTS Test Sets

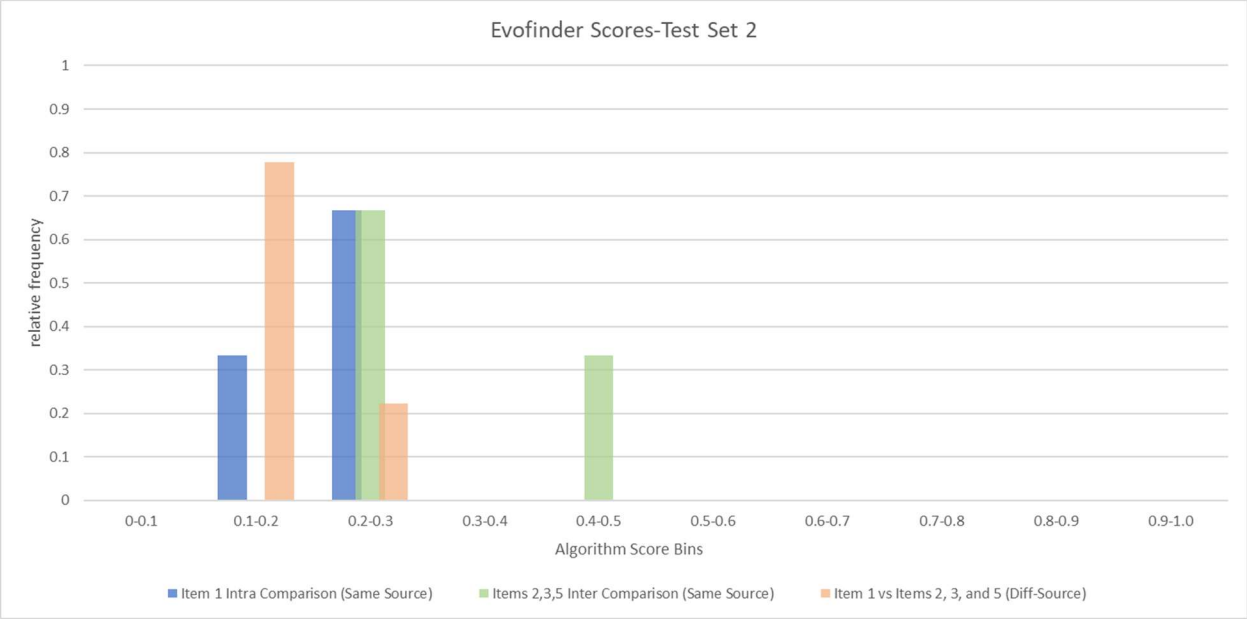
Test Set 1

	Item 1 T1	Item 1 T2	Item 1 T3	Item 2	Item 3	Item 5
Item 1 T1		0.229	0.215	0.171	0.185	0.166
Item 1 T2			0.236	0.208	0.153	0.198
Item 1 T3				0.158	0.168	0.193
Item 2					0.284	0.307
Item 3						0.276



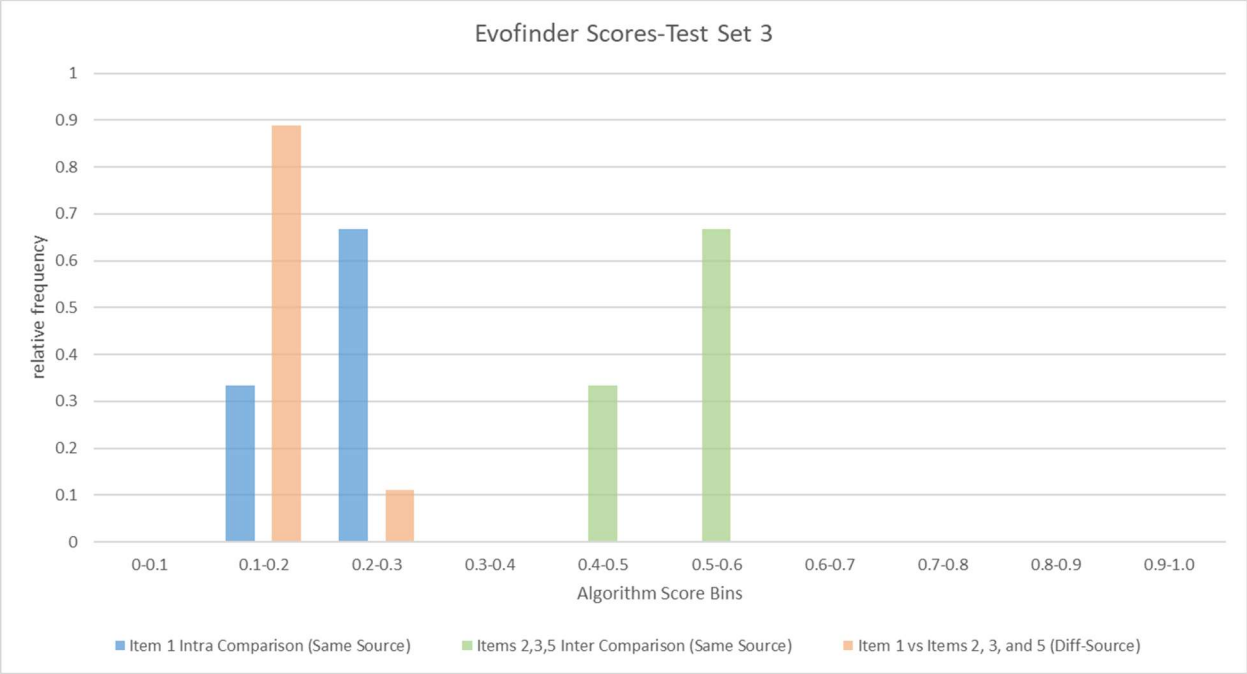
Test Set 2

	Item 1 T1	Item 1 T2	Item 1 T3	Item 2	Item 3	Item 5
Item 1 T1		**0.171**	**0.28**	**0.207**	**0.195**	**0.171**
Item 1 T2			**0.212**	**0.182**	**0.168**	**0.174**
Item 1 T3				**0.18**	**0.212**	**0.171**
Item 2					**0.221**	**0.479**
Item 3						**0.266**



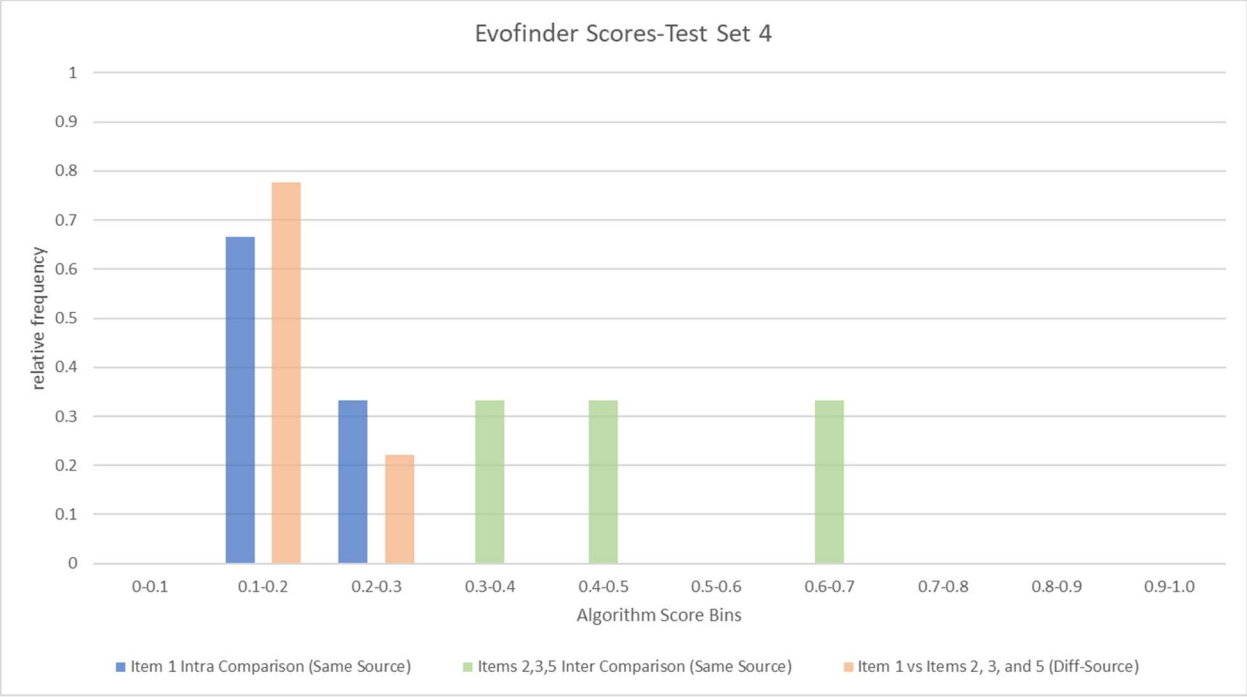
Test Set 3

	Item 1 T1	Item 1 T2	Item 1 T3	Item 2	Item 3	Item 5
Item 1 T1		**0.198**	**0.204**	**0.156**	**0.2**	**0.191**
Item 1 T2			**0.226**	**0.181**	**0.192**	**0.185**
Item 1 T3				**0.25**	**0.167**	**0.155**
Item 2					**0.554**	**0.535**
Item 3						**0.266**



Test Set 4

	Item 1 T1	Item 1 T2	Item 1 T3	Item 2	Item 3	Item 5
Item 1 T1		0.203	0.192	0.209	0.158	0.169
Item 1 T2			0.168	0.176	0.235	0.173
Item 1 T3				0.18	0.169	0.176
Item 2					0.488	0.612
Item 3						0.393

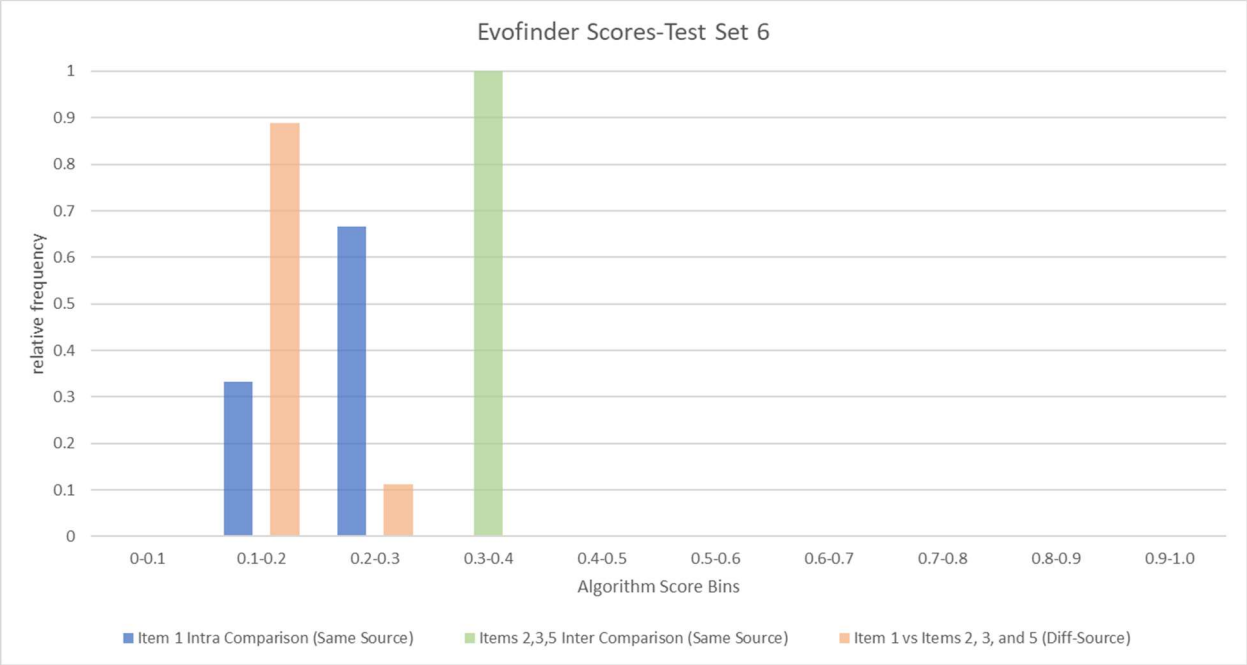


Test Set 5

	Item 1 T1	Item 1 T2	Item 1 T3	Item 2	Item 3	Item 5
Item 1 T1		**0.188**	**0.197**	**0.139**	**0.147**	**0.159**
Item 1 T2			**0.212**	**0.177**	**0.15**	**0.162**
Item 1 T3				**0.179**	**0.177**	**0.169**
Item 2					**0.225**	**0.351**
Item 3						**0.292**

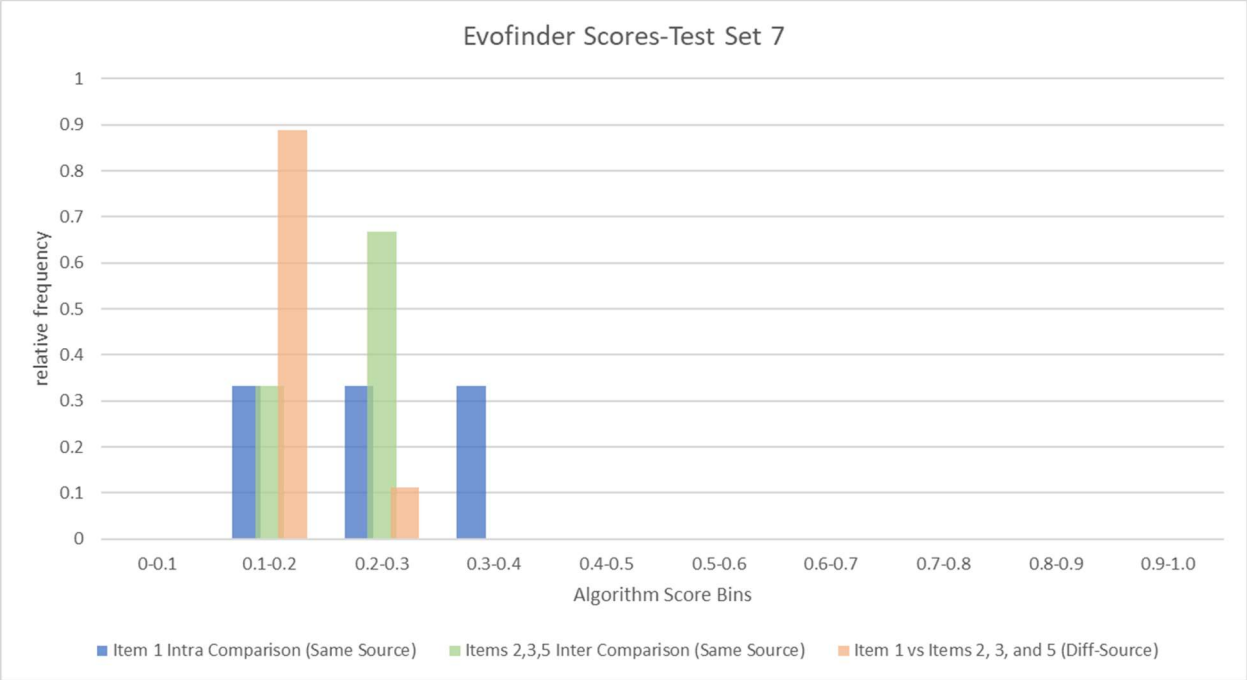
Test Set 6

	Item 1 T1	Item 1 T2	Item 1 T3	Item 2	Item 3	Item 5
Item 1 T1		0.161	0.202	0.161	0.2	0.176
Item 1 T2			0.216	0.171	0.13	0.228
Item 1 T3				0.176	0.171	0.15
Item 2					0.347	0.334
Item 3						0.316



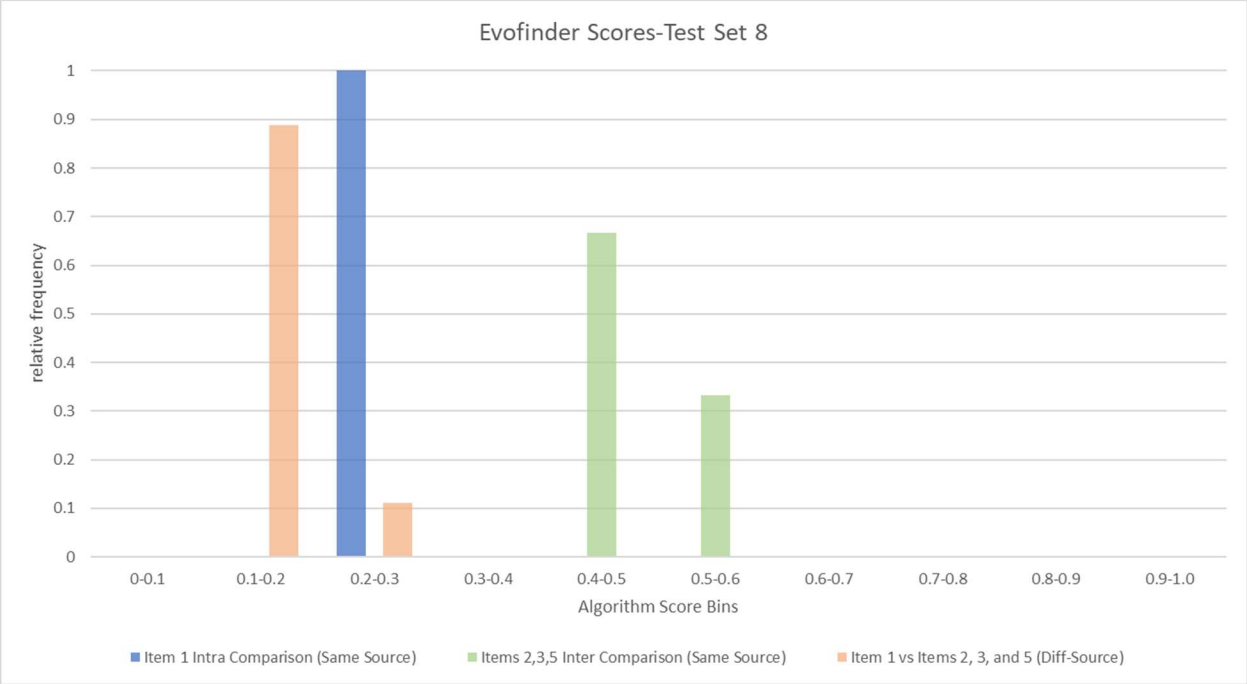
Test Set 7

	Item 1 T1	Item 1 T2	Item 1 T3	Item 2	Item 3	Item 5
Item 1 T1		0.376	0.185	0.164	0.17	0.192
Item 1 T2			0.215	0.157	0.205	0.183
Item 1 T3				0.18	0.181	0.174
Item 2					0.268	0.272
Item 3						0.146



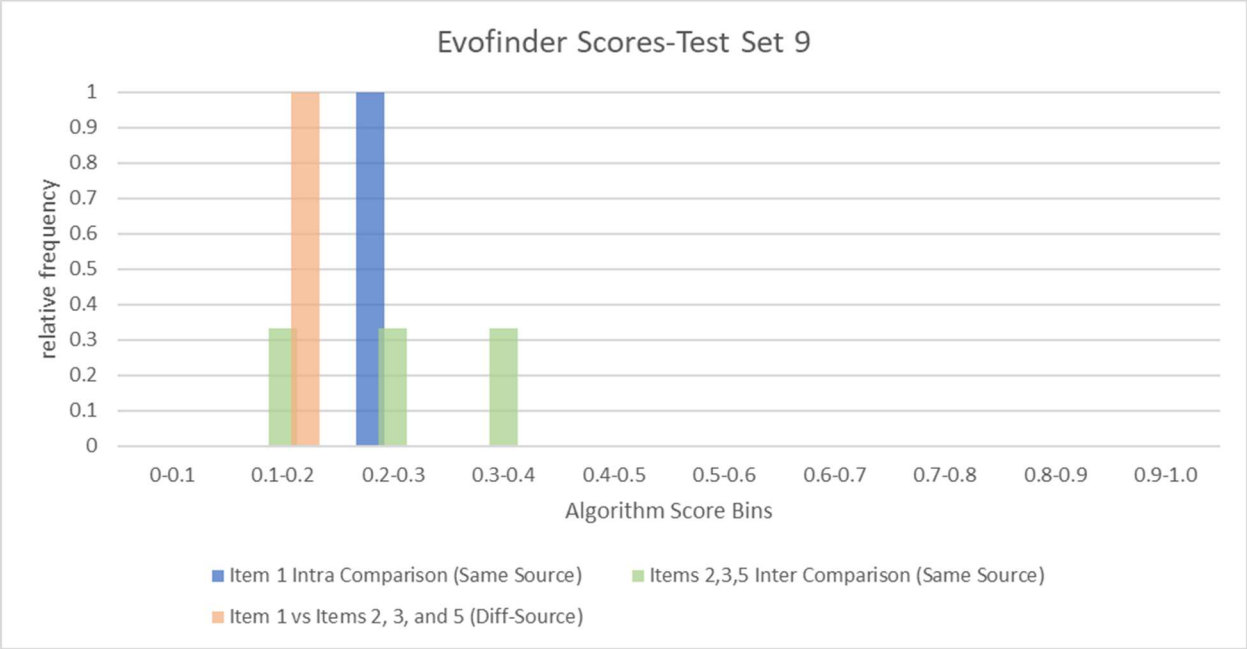
Test Set 8

	Item 1 T1	Item 1 T2	Item 1 T3	Item 2	Item 3	Item 5
Item 1 T1		0.206	0.207	0.181	0.18	0.196
Item 1 T2			0.255	0.176	0.173	0.2
Item 1 T3				0.154	0.199	0.223
Item 2					0.455	0.506
Item 3						0.453



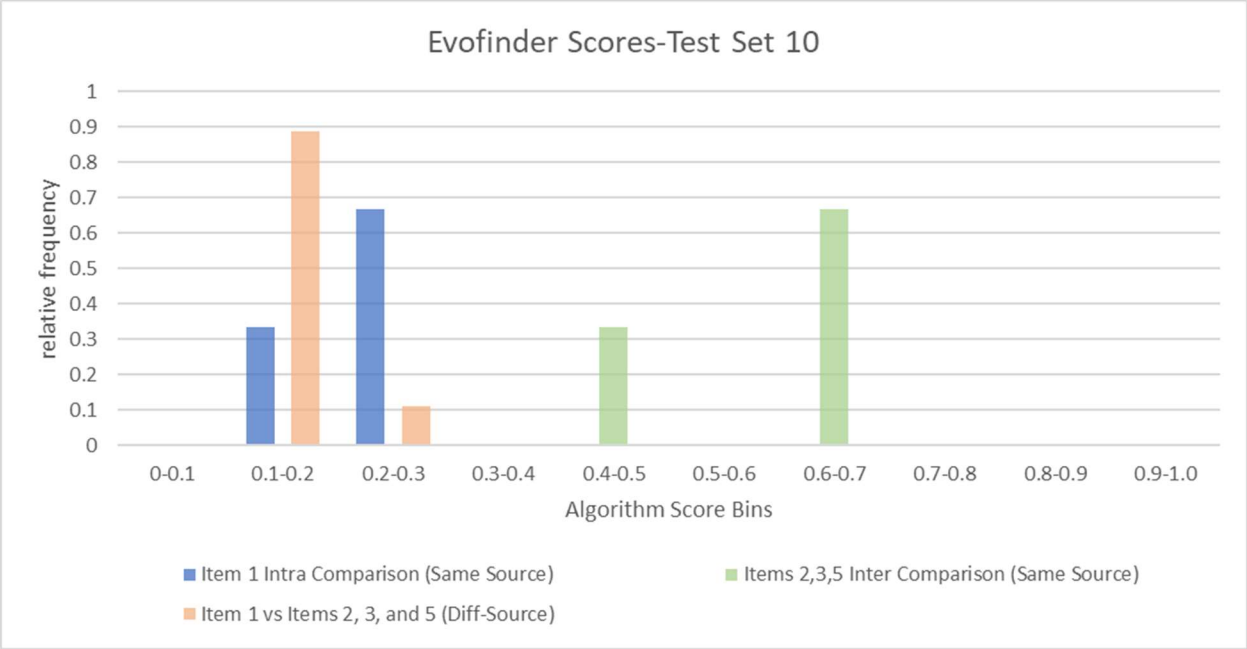
Test Set 9

	Item 1 T1	Item 1 T2	Item 1 T3	Item 2	Item 3	Item 5
Item 1 T1		**0.299**	**0.202**	**0.146**	**0.178**	**0.177**
Item 1 T2			**0.203**	**0.185**	**0.184**	**0.173**
Item 1 T3				**0.177**	**0.16**	**0.187**
Item 2					**0.176**	**0.23**
Item 3						**0.39**



Test Set 10

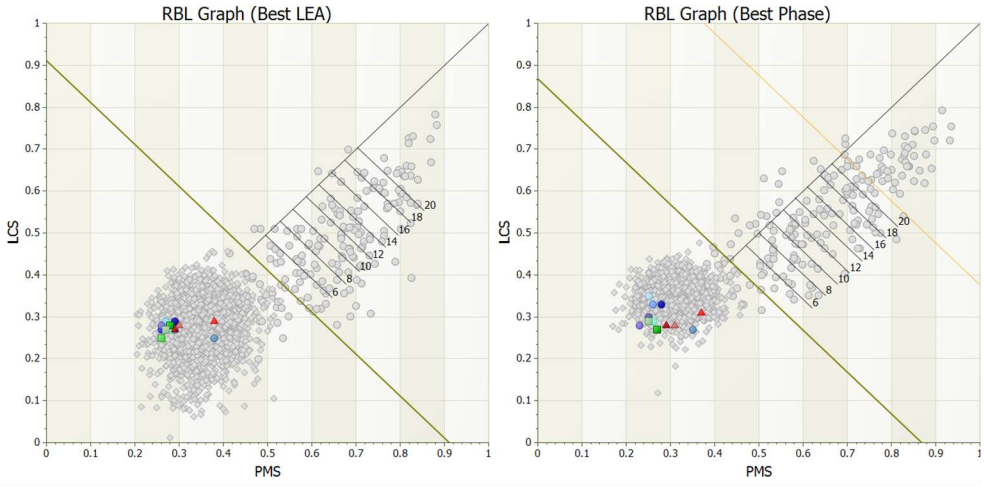
	Item 1 T1	Item 1 T2	Item 1 T3	Item 2	Item 3	Item 5
Item 1 T1		**0.211**	**0.18**	**0.186**	**0.169**	**0.176**
Item 1 T2			**0.244**	**0.161**	**0.177**	**0.182**
Item 1 T3				**0.229**	**0.182**	**0.148**
Item 2					**0.494**	**0.622**
Item 3						**0.613**



Appendix C

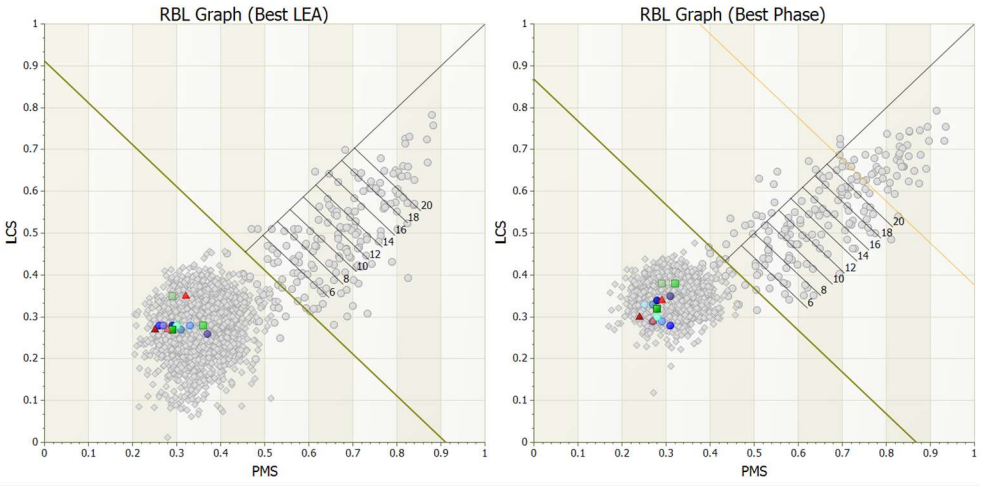
Graphs From Houston Forensic Science Center

HFSC Test Set 1



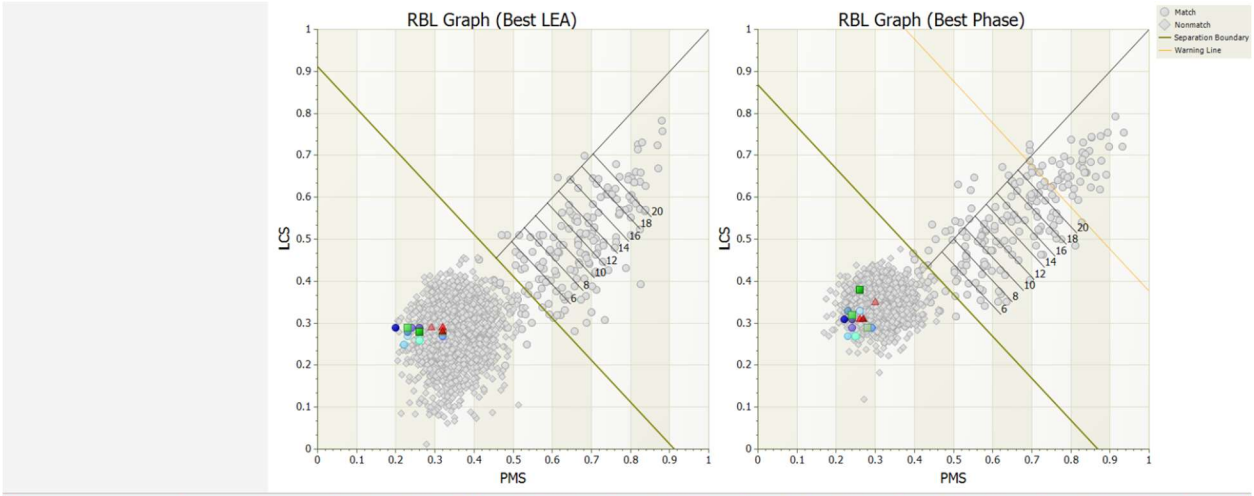
Acquisition Name 1	Acquisition Name 2	Best LEA PMS	Best LEA LCS	Best LEA FMR	Best Phase PMS	Best Phase LCS	Best Phase FMR
2023PT2	2023PT1C	0.38	0.25	0.00	0.35	0.27	0.00
2023PT2	2023PT3	0.38	0.29	0.30	0.37	0.31	0.32
2023PT3	2023PT5	0.30	0.28	0.00	0.31	0.28	0.00
2023PT5	2023PT1B	0.29	0.27	0.00	0.26	0.33	0.00
2023PT3	2023PT1A	0.29	0.28	0.00	0.25	0.30	0.00
2023PT3	2023PT1C	0.29	0.29	0.00	0.28	0.33	0.00
2023PT2	2023PT5	0.29	0.27	0.00	0.29	0.28	0.00
2023PT2	2023PT1B	0.28	0.27	0.00	0.25	0.35	0.00
2023PT1A	2023PT1B	0.28	0.28	0.00	0.27	0.27	0.00
2023PT5	2023PT1C	0.28	0.27	0.00	0.26	0.29	0.00
2023PT3	2023PT1B	0.27	0.29	0.00	0.26	0.29	0.00
2023PT1B	2023PT1C	0.27	0.27	0.00	0.25	0.29	0.00
2023PT5	2023PT1A	0.26	0.28	0.00	0.23	0.28	0.00
2023PT2	2023PT1A	0.26	0.27	0.00	0.25	0.29	0.00
2023PT1A	2023PT1C	0.26	0.25	0.00	0.25	0.29	0.00

HFSC Test Set 2



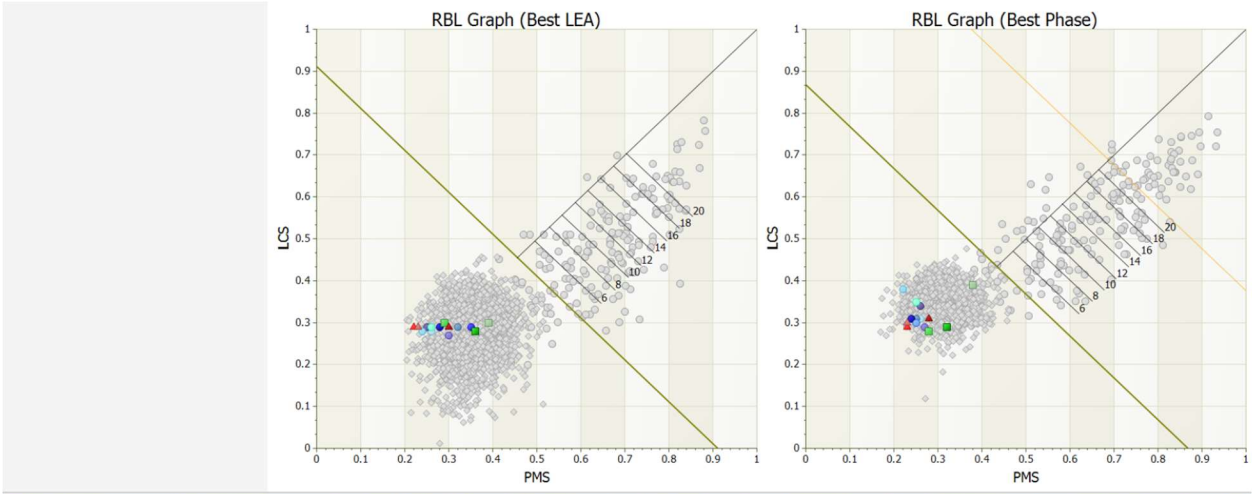
Acquisition Name 1	Acquisition Name 2	Best LEA PMS	Best LEA LCS	Best LEA FMR	Best Phase PMS	Best Phase LCS	Best Phase FMR
2023PT3	2023PT1A	0.37	0.26	0.00	0.31	0.35	0.00
2023PT1A	2023PT1C	0.36	0.28	0.00	0.32	0.38	0.68
2023PT5	2023PT1B	0.33	0.28	0.00	0.29	0.29	0.00
2023PT2	2023PT3	0.32	0.35	0.29	0.29	0.34	0.00
2023PT2	2023PT1C	0.31	0.27	0.00	0.27	0.33	0.00
2023PT2	2023PT1B	0.31	0.27	0.00	0.25	0.33	0.00
2023PT5	2023PT1C	0.30	0.28	0.00	0.28	0.30	0.00
2023PT1B	2023PT1C	0.29	0.35	0.04	0.29	0.38	0.22
2023PT3	2023PT1C	0.29	0.28	0.00	0.28	0.34	0.00
2023PT3	2023PT1B	0.29	0.27	0.00	0.27	0.29	0.00
2023PT1A	2023PT1B	0.29	0.27	0.00	0.28	0.32	0.00
2023PT3	2023PT5	0.28	0.27	0.00	0.27	0.29	0.00
2023PT5	2023PT1A	0.27	0.28	0.00	0.27	0.29	0.00
2023PT2	2023PT1A	0.26	0.28	0.00	0.31	0.28	0.00
2023PT2	2023PT5	0.25	0.27	0.00	0.24	0.30	0.00

HFSC Test Set 3



Acquisition Name 1	Acquisition Name 2	Best LEA PMS	Best LEA LCS	Best LEA FMR	Best Phase PMS	Best Phase LCS	Best Phase FMR
DC5	DC1B	0.32	0.27	0.00	0.29	0.29	0.00
DC2	DC3	0.32	0.29	0.00	0.26	0.31	0.00
DC1B	DC1C	0.32	0.28	0.00	0.28	0.29	0.00
DC2	DC5	0.32	0.28	0.00	0.27	0.31	0.00
DC3	DC5	0.29	0.29	0.00	0.30	0.35	0.00
DC1A	DC1B	0.26	0.28	0.00	0.26	0.38	0.00
DC5	DC1C	0.26	0.26	0.00	0.25	0.27	0.00
DC3	DC1A	0.26	0.29	0.00	0.24	0.31	0.00
DC5	DC1A	0.24	0.29	0.00	0.24	0.29	0.00
DC2	DC1B	0.23	0.27	0.00	0.26	0.33	0.00
DC1A	DC1C	0.23	0.29	0.00	0.24	0.32	0.00
DC2	DC1C	0.23	0.28	0.00	0.23	0.33	0.00
DC2	DC1A	0.23	0.29	0.00	0.26	0.33	0.00
DC3	DC1B	0.22	0.25	0.00	0.23	0.27	0.00
DC3	DC1C	0.20	0.29	0.00	0.22	0.31	0.00

HFSC Test Set 4



Acquisition Name 1	Acquisition Name 2	Best LEA PMS	Best LEA LCS	Best LEA FMR	Best Phase PMS	Best Phase LCS	Best Phase FMR
1B	1C	0.39	0.30	0.50	0.38	0.39	1.89
1A	1B	0.36	0.28	0.10	0.32	0.29	0.00
2	1A	0.35	0.29	0.07	0.32	0.29	0.00
2	1C	0.32	0.29	0.00	0.25	0.31	0.00
5	1A	0.30	0.27	0.00	0.27	0.29	0.00
2	5	0.30	0.29	0.00	0.28	0.31	0.00
1A	1C	0.29	0.30	0.00	0.28	0.28	0.00
3	1C	0.28	0.29	0.00	0.24	0.31	0.00
2	1B	0.26	0.28	0.00	0.26	0.31	0.00
5	1B	0.26	0.29	0.00	0.25	0.30	0.00
5	1C	0.26	0.29	0.00	0.25	0.35	0.00
3	1A	0.25	0.29	0.00	0.26	0.34	0.00
3	1B	0.24	0.28	0.00	0.22	0.38	0.00
3	5	0.23	0.29	0.00	0.23	0.30	0.00
2	3	0.22	0.29	0.00	0.23	0.29	0.00

ERRATUM

In the *Report of the Association of Firearm and Tool Mark Examiners Proficiency Test Review Ad-Hoc Committee on the Results of the Collaborative Testing Service (CTS) Firearms Examination Proficiency Test 23-5262*, dated November 25, 2024, the authors have discovered two errors in the report:

1) On page two of the report, the firearm that fired the items 2, 3, and 5 bullets was incorrectly described as "a second CZ 75 pistol with the same general rifling characteristics as the first CZ 75 pistol". In fact, according to the CTS summary report, the pistol used to fire the items 2, 3, and 5 bullets was a CZ model 40B pistol. This misstatement is not significant, as the CZ model 40B pistol exhibited the same general rifling characteristics as the CZ 75 pistol and, therefore, it does not affect the conclusions or recommendations contained in the report.

2) In the recommendations section on page 23, the recommendation to make the three-dimensional scan data available to the community was inadvertently included as part of recommendation #2 when it should have been its own, stand-alone recommendation.

December 10, 2024