3-2020

How do latent print examiners perceive proficiency testing? An analysis of examiner perceptions, performance, and print quality

Sharon Kelley
University of Virginia

Brett O. Gardner
University of Virginia

Daniel C. Murrie
University of Virginia

Karen D.H. Pan
University of Virginia

Karen Kafadar
University of Virginia

How do latent print examiners perceive proficiency testing? An analysis of examiner perceptions, performance, and print quality

Abstract

Proficiency testing has the potential to serve several important purposes for crime laboratories and forensic science disciplines. Scholars and other stakeholders, however, have criticized standard proficiency testing procedures since their implementation in laboratories across the United States. Specifically, many experts label current proficiency tests as non-representative of actual casework, at least in part because they are not sufficiently challenging (e.g., [1], [2], [3], [4]. In the current study, we surveyed latent print examiners (n = 322) after they completed a Collaborative Testing Services proficiency test about their perceptions of test items. We also evaluated respondents' test performance and used a quality metric algorithm (LQMetrics) to obtain objective indicators of print quality on the test. Results were generally consistent with experts' concerns about proficiency testing. The low observed error rate, examiner perceptions of relative ease, and high objective print quality metrics together suggest that latent print proficiency testing is not especially challenging. Further, examiners indicated that the test items that most closely resembled real-world casework were also the most difficult and contained prints of the lowest quality. Study findings suggest that including prints of lower quality may increase both the difficulty and representativeness of proficiency testing in latent print examination.

Keywords

Forensic science, Latent prints, Proficiency testing, Quality metrics

Disciplines

Forensic Science and Technology

Comments

Kelley, S., Gardner, B.O., Murrie, D.C., Pan, K.D.H., Kafadar, K., How do latent print examiners perceive proficiency testing? An analysis of examiner perceptions, performance, and print quality.Science & Justice , March 2020,60(2);120-127. Doi: 10.1016/j.scijus.2019.11.002 Posted with permission of CSAFE.

How do latent print examiners perceive proficiency testing? An analysis of examiner perceptions, performance, and print quality

Sharon Kelley, Brett O. Gardner, Daniel C. Murrie, Karen D.H. Pan, Karen Kafadar

Institute of Law, Psychiatry, and Public Policy, University of Virginia, USA
Department of Statistics, University of Virginia, USA

ARTICLE INFO

Keywords:
Forensic science
Latent prints
Proficiency testing
Quality metrics

ABSTRACT

Proficiency testing has the potential to serve several important purposes for crime laboratories and forensic science disciplines. Scholars and other stakeholders, however, have criticized standard proficiency testing procedures since their implementation in laboratories across the United States. Specifically, many experts label current proficiency tests as non-representative of actual casework, at least in part because they are not sufficiently challenging (e.g., [1–4]. In the current study, we surveyed latent print examiners (n = 322) after they completed a Collaborative Testing Services proficiency test about their perceptions of test items. We also evaluated respondents' test performance and used a quality metric algorithm (LQMetrics) to obtain objective indicators of print quality on the test. Results were generally consistent with experts' concerns about proficiency testing. The low observed error rate, examiner perceptions of relative ease, and high objective print quality metrics together suggest that latent print proficiency testing is not especially challenging. Further, examiners indicated that the test items that most closely resembled real-world casework were also the most difficult and contained prints of the lowest quality. Study findings suggest that including prints of lower quality may increase both the difficulty and representativeness of proficiency testing in latent print examination.

1. Introduction

The National Commission on Forensic Science [5] defines proficiency testing as the "evaluation of participant performance against pre-established criteria by means of interlaboratory comparisons" (p. 3). The field of forensic science widely acknowledges the importance of such testing as proficiency tests serve many purposes within crime laboratories. Proficiency testing, often part of laboratories' quality assurance processes, can help train incoming forensic analysts and establish base levels of competency within and across laboratories. Routine proficiency testing can also identify problematic laboratory procedures and ultimately improve analyst practice. Finally, proficiency tests can provide performance indicators for external constituencies, including accrediting bodies for ISO standards (e.g., ANSI National Accreditation Board) and members of the legal system (e.g., judges, lawyers).

1.1. Current proficiency testing in latent print examination

Proficiency testing materials can be created and distributed internally by individual laboratories, or laboratories can purchase commercially available tests or reference materials from different groups (e.g., RTI International, Ron Smith & Associates, Inc.). Collaborative Testing Services (CTS) is one of the primary suppliers of proficiency tests for numerous forensic science disciplines, including latent print examination. CTS has offered forensic proficiency testing since the late 1970s to provide an external indication of a crime laboratory's proficiency. Today, approximately 900 laboratories in more than 80 countries participate in CTS proficiency testing [6]. Currently, CTS distributes a latent print examination proficiency test to hundreds of laboratories twice yearly. After receiving test results from all participating laboratories (laboratories choose whether to share test results), CTS publishes anonymous summary reports online.

1.2. Criticisms of proficiency testing in forensic science

Despite the value of proficiency testing, experts have raised concerns about current proficiency testing practices. These concerns generally fall into one of two categories: 1) the representativeness of examiners' behavior during proficiency testing and 2) the difficulty of items on proficiency tests. Regarding examiner behavior: An early study by Cembrowski and Vanderline [7] found that analysts behave

https://doi.org/10.1016/j.scijus.2019.11.002

differently during proficiency testing than during routine analyses. Specifically, surveyed analysts indicated that they engaged in special practices (e.g., conducting an analysis with multiple instruments, spending additional time on an analysis) to ensure accurate conclusions during proficiency testing that they do not engage in during routine casework. Relatedly, calls for proficiency tests to be conducted blindly (i.e., respondents complete testing without awareness that they are being tested) began in earnest in the 1990s after the development of DNA evidence (e.g., [8,9]). Many scholars shared the opinion that, "non-blind proficiency tests may not provide a good indicator of the error rate in actual case work because the technicians may be unusually diligent and cautious when they know they are being observed and tested" [10]. Indeed, the National Academy of Sciences issued a report in 1992 asserting that, "laboratory error rates must be continually estimated in blind proficiency testing" [9]. However, in the 25 years following such calls to action, blind proficiency testing has still not become standard [11].

Regarding the difficulty of proficiency tests: Multiple sources, including accreditation standards (e.g., ISO/IEC 17043), professional organizations (e.g., Scientific Working Group on Friction Ridge Analysis, Study and Technology), and national reports (e.g., the 2016 report by the President's Council of Advisors on Science and Technology) call for proficiency testing that resembles the demands of casework. At the same time, scholars have long criticized current proficiency test standards for not adequately representing "real-world" casework and for the lost opportunity to provide a source for error rate estimates. Of course, research on examiner performance has increased over the last several years, including work by Ulery and colleagues (e.g., [12,13]), Langenbug (e.g., [14]), and several others (e.g., [15])—moving the field closer to reasonable estimates of error rates. However, there is still a need to ensure that proficiency tests represent the challenges associated with actual casework.

Koehler has repeatedly commented on the "notoriously easy" level of difficulty in typical proficiency tests ([1,2,11,16], p. 3). In 2008, Koehler reported that CTS tests "tend to be conducted under unreasonable test conditions (e.g., non-blind conditions that use relatively easy materials)" ([1], p. 1091). Other experts have expressed similar sentiments, stating that "the prints used in the FBI proficiency test are so easy they are a joke" [17] or noting that proficiency tests "have for the most part been extremely easy, far easier than the challenges that can be faced by examiners in actual casework" [3]. On the occasions tests are more challenging, the forensic science community has not consistently welcomed the change. For instance, Koehler [1] reported that, after 22% of laboratories made at least one misidentification on a latent print proficiency test in 1995, the forensic science community was disrupted and subsequent tests have appeared less difficult. Indeed, during hearings conducted by the National Commission on Forensic Science, the President of CTS reported "that he has been under commercial pressure to make proficiency test easier" [18].

Clearly, there are indicators that current proficiency testing is perhaps too easy and, taken together, the aforementioned criticisms have led some scholars to assert that "proficiency testing in forensic science is frequently worthless as a true indicator of examiner proficiency" [4]. However, as noted by Haber and Haber [19], "the profession lacks a quantitative measure of print quality" and thus, assessments of difficulty often rely on subjective judgments (p. 95). Thus, objectively evaluating or systematically adjusting the difficulty of tests has been a complicated undertaking. The development of quality metrics, reviewed below, offers one possible mechanism for assessing the difficulty of existing proficiency tests and, moving forward, creating proficiency tests with samples representative of casework.

1.3. Quality metrics

In recent years, multiple quality metric algorithms have emerged. These algorithms provide not only a quantitative metric of fingerprint quality, but a deterministic and objective score independent of any single examiner. Algorithms may utilize different aspects of latent fingerprints in calculating a score (e.g., the contrast between ridges and troughs, blur versus clarity of a print, number of features). Quality metric scores fall into two general classes. The first is a global metric which provides a single score for the latent print (regardless of whether a latent is of uniform quality or certain portions are of higher/lower fidelity), and the second is a feature- or minutiae-specific metric which provides individual scores for each marked minutiae.

One of the better known quality metrics is Latent Quality Metrics (LQMetrics), which is included in the FBI's Universal Latent Workstation (ULW), an interactive software tool for latent print examiners. Output includes four metrics directly related to quality and nine additional metrics automatically calculated from a latent fingerprint [20]. Research suggests that one of the LQMetrics, the overall clarity score, is highly correlated with latent print examiners' subjective judgments of print difficulty [21,22]. Using the overall clarity score, Koertner and Swofford [22] compared prints from 13 years of CTS latent print proficiency tests to 215 prints from normal casework. Results indicated that proficiency tests prints received significantly higher clarity scores than prints from casework.

1.4. Current study

Scholars have raised multiple concerns with latent print proficiency testing (e.g., [1–4,11]), but little is known about how examiners perceive such tests and how included prints score on an objective measure of quality. Although scholars have opined that "the tests that examiners take are generally so easy, unrealistic, and otherwise unlike casework," only one study [22] offers an empirical analysis of such claims [2]. Thus, the present study: 1) explores latent print examiners' opinions of proficiency testing, 2) explores how examiner opinions relate to performance on proficiency tests, and 3) examines both subjective and objective indicators of print quality in current proficiency tests. By doing so, we hope to summarize prevailing opinions held by those who routinely take latent print proficiency tests. We also intend to provide another objective examination of print quality in the prints included in current proficiency testing.

2. Method

2.1. Participants and procedure

We collaborated with CTS to add survey questions addressing respondents' perceptions of test items to a latent print examination proficiency test that was shipped to respondents in August 2017. At the end of the testing period, 438 respondents submitted completed tests. Respondents were latent print examiners who presumably practice in multiple countries given the international adoption of CTS's forensic proficiency tests. We do not report demographic information regarding the examiners because such information is not collected by CTS during the standard proficiency testing process. Of the respondents who submitted completed tests, approximately two thirds (66.2%; 290 of 438) also submitted answers to our survey items. We later received survey responses from an additional 32 examiners who completed the survey but did not submit test results to CTS. Thus, our final sample of latent print examiners who submitted survey responses was 322.

To augment our understanding of examiners' perceptions regarding the latent prints used in the proficiency test, we also examined all fingerprints depicted in the test using a global quality metric: LQMetrics. LQMetrics, which is included in the FBI's Universal Latent Workstation (ULW), outputs four scores directly related to print quality calculated from the information contained in a latent print [20].

2.2. Measures

CTS Latent Print Examination Proficiency Test. CTS is the largest provider of latent print examination proficiency testing and the company ships two unique tests every year in January and August. The test is offered in one of three formats: physical copies of digitally produced photographs, digital images retrieved from a DVD, or digital images retrieved from a website. In the current sample, most respondents used physical copies of photographs (69.9%) to complete the test, although some examiners used images from a DVD (23.9%) or website (6.2%).

Regardless of test format, respondents were provided 11 latent prints and four sets of known finger and palm prints belonging to four different individuals. In each set of known prints, examiners received a full-hand print (including palm prints), and a completed 10-print card (including rolled and simultaneous prints). Thus, examiners have access to one image of each latent print and three to four separate images of each known fingerprint (i.e., prints included in full-hand image, individual prints on 10-print card, and simultaneous prints made in lower section of 10-print card).

Examiners were asked to compare the prints and report their findings concerning the 11 latent prints, with each latent print representing one test item. Responses must be attributed to a specific finger or palm (e.g., left palm, right middle) or marked as "Not Identified." Thus, response options on the test are different from the standard latent print comparison outcomes of identification/individualization (i.e., a definitive identification of an individual as the source of a latent print), inconclusive (i.e., there is insufficient information to determine whether an individual is the source of a latent print), or exclusion (i.e., a determination that an individual is not the source of a latent print). On the current test, nine latent prints (seven fingerprints and two palm prints) were made by one of the four individuals who had provided known prints; print sources were equally distributed across the four individuals (i.e., three individuals were the source of two prints each, one individual was the source of three prints). Two latent prints (Q3 and Q10) were made by an unidentified individual. CTS notes that all latent prints included on its proficiency tests are "analyzed and confirmed by an external advisor to present appropriate detail, clarity, and difficulty" ([23], p. 2).[1]

Over the prior seven proficiency test administrations dating back to 2014, 88% of respondents have completed the exam without providing any erroneous[2] responses. Respondents performed better on the current CTS proficiency test (Test No. 17-5171/2/5) as only 3% (14 of 438) examiners gave an erroneous response (for additional detail, see [23]).

Supplemental Survey. A brief survey was included at the end of the CTS proficiency test asking examiners about their perceptions of the test items. Specifically, the survey asked participants to separately rate the level of challenge and similarity to casework of each latent print on the test (i.e., 11 latent prints) using an 11-point scale. Thus, challenge level could range from 0 = *Extremely easy* to 10 = *Extremely challenging*; similarity to casework could range from 0 = *Nothing like casework* to 10 = *Exactly like casework*. Additionally, we asked examiners to identify both the *least* and *most* challenging latent print, and to rate their confidence in the accuracy of their decisions using an 11-point scale ranging from 0 = *No confidence*, to 10 = *Extremely confident*. Regarding the latent print that examiners identified as most challenging, examiners identified the characteristic(s) that caused the print to be challenging from among the following: *Limited points to compare*, *Distortion*, *Overdeveloped/Underdeveloped ridge detail*, *Image quality*, *Other (please explain)*.

Latent Quality Metrics (LQMetrics). LQMetrics (Version 6.6) is included in the FBI's Universal Latent Workstation (ULW), an interactive software tool for latent print examiners, and outputs four metrics directly related to quality and nine additional metrics automatically calculated from a latent fingerprint[3] [20]. Directly relating to quality are four scores ranging from 0 to 100: latent quality, value for individualization (VID), value for comparison (VCMP), and overall clarity. The latent quality score gives the predicted probability of an "image-only search" returning a candidate list that contains the correct mate, assuming the mate is of sufficient quality and the images overlap sufficiently (i.e., a score of 90 is interpreted as a 90% chance a search returns the mate). VID and VCMP scores are interpreted analogous to the overall LQMetrics score. Thus, a VID score provides the probability that an examiner would believe the print to have sufficient quality for individualization, and a VCMP score provides the probability that an examiner would believe the quality to be sufficient for either individualization or exclusion. Finally, LQMetrics also provides an overall clarity score. This scores describes the level and quantity of friction ridge detail within the print. Unlike the three aforementioned scores, the overall clarity score does not represent a probability, although it also ranges from 0 to 100. Higher scores indicate increased friction ridge detail. For simplicity, we refer to the overall latent quality score as the LQMetrics score in this study.

Compared against qualitative assessments of quality, prints assessed to be of "good" quality by examiners corresponded to latent quality scores of 65–90, "bad" to scores of 45–65, and "ugly" to scores of 24–45 [20]. LQMetrics also outputs nine other variables automatically calculated from the latent print itself (e.g., area size of clear level 3 detail, largest contiguous area of ridge flow). However, we limited our analyses to the three overarching quality scores summarizing a print's quality (i.e., latent quality score, VID, VCMP) and the overall clarity score. We report descriptive statistics for all scores in Table 1 but only report analytic results for latent quality scores.[4]

3. Results

3.1. Examiner perceptions of proficiency prints

Perceived Difficulty and Similarity to Casework. Fig. 1 depicts respondents' average ratings regarding the perceived difficulty of latent prints and their similarity to casework. The mean level of perceived difficulty across all items was 4.27 (SD = 1.90), indicating that, in general, participants found the questions to be relatively easy (given that the average difficulty rating fell below the midpoint of the 0–10 scale). At the item level, average difficulty ratings ranged from 2.29 (a majority of respondents rated Q4 as the least challenging print) to 5.80 (a plurality of respondents rated Q9 as the most challenging print).

Examiners typically perceived the latent prints in the test to be similar to their casework (M = 6.97; SD = 2.28). Average similarity ratings varied minimally across latent prints, ranging from 6.39 (Q4 was rated as the least similar to casework) to 7.44 (Q9 was rated as the most similar to casework). Interestingly, the latent prints rated as the most, and least, challenging were also respectively rated as the most, and least, similar to casework. Indeed, items perceived as more similar to casework were also perceived as more challenging (r[306] = 0.30, $p < .001$).

Respondents' published comments on the proficiency test also provide anecdotal evidence regarding the perceived difficulty of the test. Specifically, one examiner noted that,

Latent prints were not challenging enough. The pattern was visible

[1] Further information about how proficiency tests are created and assessed can be found on the CTS website, cts-forensics.com.

[2] CTS refers to erroneous responses as "inconsistent results" in its summary reports and communications with latent print examiners given that such answers are inconsistent with the answers provided by external verifiers.

[3] LQMetrics is currently limited to fingerprint analysis, although the user guide notes that a future version will support analysis of palm prints.

[4] Results did not significantly change when we re-ran all analyses using overall clarity scores instead of latent quality scores.

Table 1
Latent quality metrics for prints included on proficiency test.

Prints	Objective LQMetrics											
	Latent quality score			Overall clarity score			VID			VCMP		
	M	*Mdn*	Range	*M*	*Mdn*	Range	*M*	*Mdn*	Range	*M*	*Mdn*	Range
Latent prints (n = 9)	74.44	72.00	(60–88)	44.33	43.00	(32–56)	98.22	98.00	(96–100)	99.44	99.00	(99–100)
Known fingerprints (n = 124)	92.41	96.00	(55–99)	70.60	70.00	(33–94)	99.64	100	(95–100)	99.97	100	(99–100)
● Full-hand prints (n = 44)	88.30	95.00	(55–99)	61.86	66.00	(33–86)	99.30	100	(95–100)	99.91	100	(99–100)
● 10-print card (n = 40)	98.80	99.00	(96–99)	84.35	85.00	(70–94)	100	100	–	100	100	–
● Simultaneous 10-print (n = 40)	90.55	94.50	(73–99)	66.48	65.00	(42–93)	99.68	100	(98–100)	100	100	–
Source fingerprints (n = 22)	93.41	96.50	(71–99)	71.05	72.00	(39–91)	99.77	100	(98–100)	99.95	100	(99–100)

Note. VID = Value for individualization; the probability an examiner would assess a latent to be of sufficient quality for individualization. VCMP = Value for comparison; the probability an examiner would assess a latent to be of sufficient quality for individualization or exclusion. Range of scores described in parentheses. Latent Quality Scores were generally heavily skewed left.

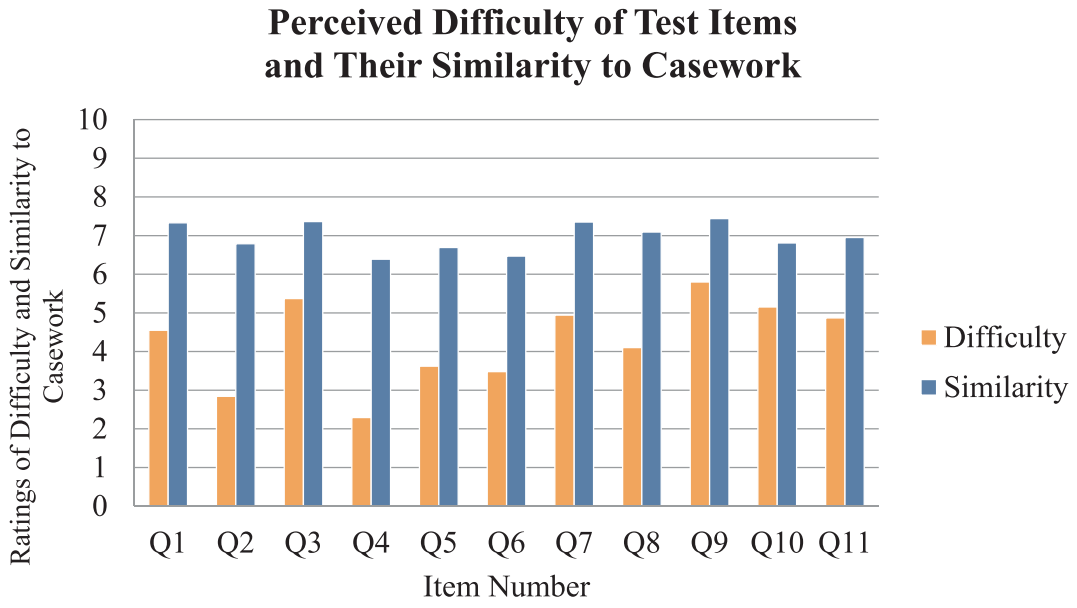

Fig. 1. Perceived Difficulty of Test Items and Their Similarity to Casework. n = 296 to 314.

in every latent, which is not good. The latent prints from palms were very easy to locate too. In the future we would like to have more challenging test latents: latents where the pattern is not visible, only partial latents, etc. ([23], p. 27).

However, this sentiment was not held by all respondents as another examiner commented, “It will be easier if electronic images were given” ([23], p. 27).

Least Challenging Latent Print. In addition to providing item ratings for every latent print, we asked participants to identify the least challenging print on the test and to indicate their confidence in their conclusion regarding that print. Two latent prints emerged as the consensus picks for least challenging items on the test. Almost two thirds of examiners (63.7%; 200 of 314) endorsed item Q4 as the least challenging item and 16.6% (52 of 314) of examiners identified item Q2 as the least challenging. A sizable minority of respondents identified items Q5 (7.3%; 23 of 314) and Q6 (5.4%; 17 of 314) as the least challenging items, with other examiners identifying various other prints. Items Q9 and Q10 were the only items not identified by any examiners as the least challenging on the test.

Regarding examiners’ confidence in the latent print they identified as least challenging, almost all respondents (96.5%) endorsed maximum confidence in their conclusions (i.e., they endorsed a score of 10 on the scale of 0 to 10). On average, examiners reported a 9.95 (SD = 0.27) on the scale from 0 = *No Confidence*, to 10 = *Extremely Confident*, with no examiner endorsing less than an 8.

Most Challenging Latent Print. Examiners’ responses regarding which print was most challenging were more varied than were responses identifying the least challenging prints. Item Q9 was most frequently identified as the most challenging print (39.2% of respondents; 124 of 316), followed by item Q3 (17.7%; 56 of 316) and Q11 (15.8%; 50 of 316). Fewer examiners identified item Q7 (11.7%; 37 of 316) and Q10 (8.9%; 28 of 316) to be the most challenging. Items Q2 and Q4 (the two items most often rated as least challenging) were the only items not identified by any examiners as the most challenging on the test.

Examiners endorsed less confidence in conclusions regarding their most challenging items; however, their expressed confidence remained very high. Indeed, approximately three fourths of respondents (77.5%; 245 of 316) endorsed maximum confidence regarding the item they perceived as most challenging. On average, examiners endorsed a 9.44 (SD = 1.36), with responses ranging from 2 to 10.

Examiners also specified what characteristics about the latent print they perceived as most challenging contributed to its difficulty. The majority of respondents identified two or more characteristics (55.0%; 177 of 322). Fig. 2 indicates that image quality (55.6%; 175 of 315) and distortion (47.6%; 150 of 315) were most commonly cited as reasons a particular item was challenging. Among “Other” reasons provided by examiners, many respondents (31.7%; 32 of 101) indicated that poor known print quality made items challenging whereas others noted that exclusions were more difficult than identifications (9.9%; 10 of 101) or

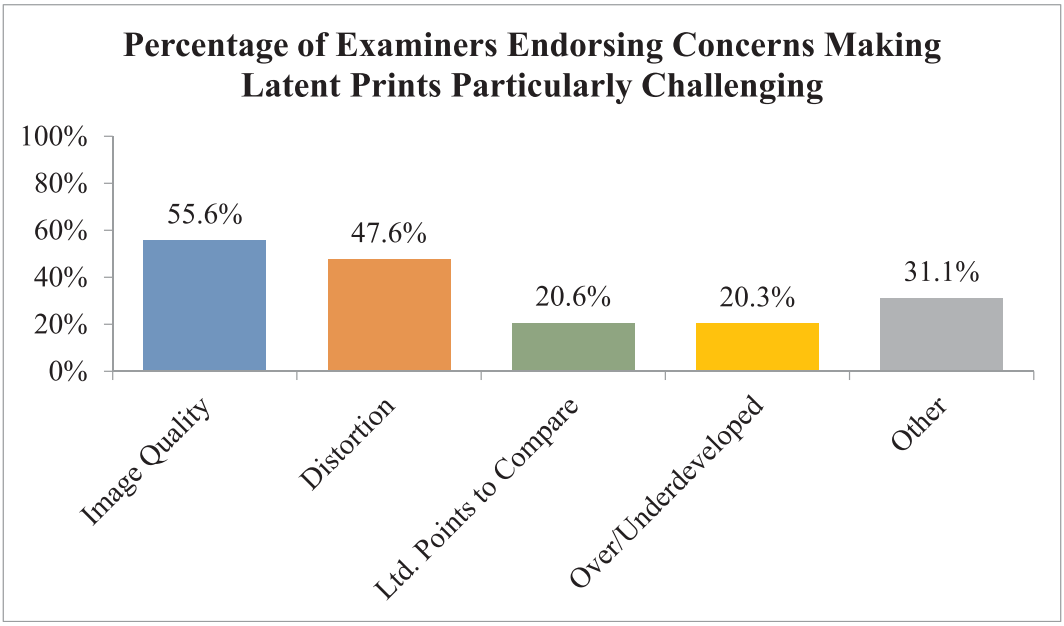

Fig. 2. Percentage of Examiners Endorsing Concerns Making Latent Prints Particularly Challenging. $n = 315$.

cited concerns in examining palm prints (9.9%; 10 of 101) or creases (7.9%; 8 of 101).

Regarding the difficult characteristics associated with specific latent prints, we examined commonly cited characteristics among the five most challenging items. As shown in Fig. 3, image quality and distortion concerns were prominent for four of the five items perceived as most challenging. Examiners who identified items Q9 or Q10 as most challenging were also more likely to endorse concerns regarding limited points to compare than were other examiners. Unlike other respondents, examiners who identified item Q11 as the most challenging item were most likely to cite "Other" concerns about the print, primarily consisting of concerns regarding known print quality, comparing palm prints, and creases. Concern regarding known print quality was also often cited among examiners who identified items Q7 and Q9 whereas concern regarding the difficulty of exclusionary conclusions was often cited among examiners who identified item Q3.

3.2. *Objective measures of proficiency prints*

We calculated LQMetrics scores for 9 of 11 latent prints (two latent

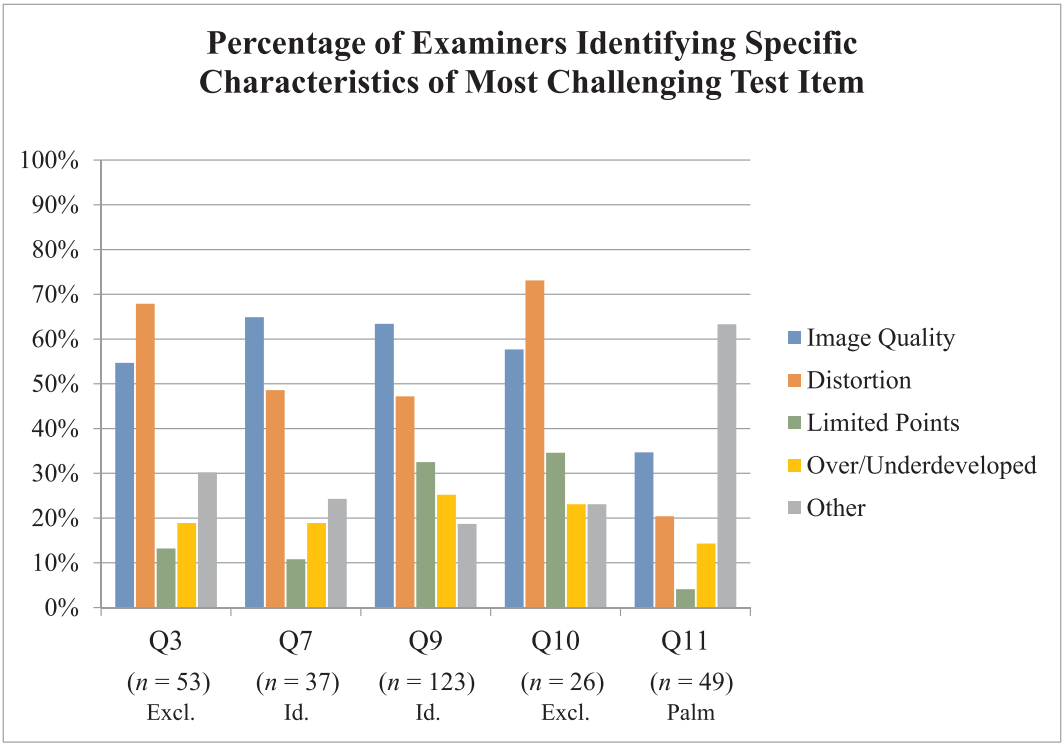

Fig. 3. Percentage of Examiners Identifying Specific Characteristics of Most Challenging Test Item. Excl. = correct exclusionary conclusion. Id. = correct identification conclusion. Palm = latent print derived from palm.

prints were palm prints [i.e., Q5 and Q11] and could not be scored via LQMetrics) and all provided known fingerprints ($n = 124$). As reported in Table 1, the average LQMetrics score for latent prints was fairly high at 74.44. The average LQMetrics score for all known fingerprints was 92.41, indicating that the overall quality of the known prints was extremely high. Further, the source fingerprints (i.e., known prints determined to be the source of a latent print on the test; $n = 22$) were of even higher quality, with an average LQMetrics score of 93.41. Overall clarity scores among latent prints were closer to the midpoint of the scale, with a mean of 44.33. Known prints had a mean overall clarity score of 70.60. Taken together, the objective measures of print quality indicate that these latent prints fall into the category of "good" quality prints while "bad" and "ugly" prints were not well represented. However, the clarity scores suggest that the latent prints on the test did not unambiguously depict the presence or attributes of features (e.g., pores, minutiae)—suggesting a higher level of difficulty than quality scores alone might indicate.

3.3. Association between examiner perceptions and objective measures

LQMetrics scores for latent prints were not associated with examiner perceptions of item difficulty ($r[7] = -0.34$, $p = .37$) or similarity to casework ($r[7] = -0.25$, $p = .52$). We also averaged LQMetrics scores (i.e., the overall latent quality score) for source prints of each latent print ($n = 7$)[5] to assess the overall source print quality. A formal test between the average quality of source prints and examiner perceptions of item difficulty or similarity to casework showed little association ($r[5] = -0.67$, $p = .10$; $r[5] = -0.54$, $p = .21$, respectively). The lack of formal significance likely is due to the small number ($n = 7$) of analyzed prints. It is worth noting, however, that the item that was perceived as most difficult and similar to casework (Q9) contained the lowest quality ($M = 81.00$) and least clear ($M = 53.50$) source prints, and all trends followed the same pattern (i.e., lower quality prints were associated with increased perceptions of difficulty and similarity to casework).

To compute an overall quality score for each test item, we averaged the quality metric scores for each latent print with the quality metric scores of its source prints (when provided). Although the resulting value reflects a mathematical combination of multiple images (a process with no equivalency in real casework), there is still merit in being able to assign numerical values to each test item that reflect the clarity of both the latent print and source prints. We were able to calculate overall print quality scores for 7 of 11 items[6], with an average LQMetrics score of 84.36 and scores ranging from 75.00 (Q9) to 92.67 (Q6). Overall, this quality score appeared to be associated with examiners' perceptions of item difficulty and similarity to casework; formal tests of the Pearson correlation coefficient ($r[5] = -0.75$, $p = .05$; $r[5] = -0.82$, $p = .02$, respectively) suggests non-zero associations, even with this small number of prints. Thus, test items perceived as more challenging and more similar to casework contained prints of lower quality. In other words, examiners' perception of item difficulty is correlated with perceptions of similarity to casework, and either outcome appears highly correlated with overall print quality scores: the Pearson correlation coefficients (and approximate 95% confidence intervals, obtained via 500 bootstrap replications) are $r = -0.75$ (95% CI: -0.99, -0.21) and $r = -0.82$ (95% CI: -0.99, -0.59), respectively.

[5] Recall that only 9 of the 11 latent prints had source prints; two latent prints were made by an unidentified individual. Additionally, LQMetrics scores for source prints could not be calculated for items with source palm prints.

[6] Again, of the 11 latent prints on the test, two latent prints were made by an unidentified individual and two were derived from palm prints. Thus, we could only calculate overall quality scores (i.e., average quality metric score of a latent print and its source print[s]) for 7 of 11 items.

3.4. Examiner perceptions, quality Metrics, and test performance

Of the 438 respondents who submitted a completed latent print examination test, only 14 examiners gave an inaccurate response to any test item. Of the 290 respondents who also submitted responses to our survey, only 11 examiners provided an inaccurate response to any test item. Further, only one respondent provided more than one incorrect response (i.e., three incorrect responses). Put differently, of the 3,190 test items submitted in the current sample, only 13 submitted items contained inaccurate conclusions (i.e., 99.6% overall accuracy rate across items). Item Q9 was responsible for the most erroneous responses (53.8% of errors); every one of the 438 examiners was correct on items Q1–Q6.

The extremely high rate of accuracy precluded thorough evaluation of the association between survey responses and responses to test items as virtually all respondents provided the exact same answers to test items. Limited analyses revealed that the 11 examiners who provided erroneous responses did not generally perceive items to be more or less difficult than those who did not provide any inconsistent responses. Again, the average level of perceived difficulty across all items was 4.27 on a scale from 0 = *Extremely Easy*, to 10 = *Extremely Difficult*. Examiners who gave inaccurate responses did not indicate that such items were substantially more or less difficult than typical ($M = 3.83$). Only 2 of the 11 respondents with erroneous responses identified the item for which they provided an inaccurate response as being the most challenging and, of the two, one examiner endorsed maximum confidence in their conclusion regarding that print.

Although we did not perform formal analyses due to the small sample size, average latent print ($M = 73.40$), source print ($M = 91.11$), and combined ($M = 80.22$) LQMetrics scores were lower for test items with errors than for items without errors (latent: $M = 83.00$; source: $M = 96.17$; combined: $M = 87.46$). As previously mentioned, respondents were most likely to offer an erroneous conclusion on item Q9. While this item did not have the lowest quality latent print, the quality scores for its source print images (LQMetrics = 71, 76, 79, & 98) were much lower than the quality scores for the other provided source prints (LQMetrics = 89 to 99). The overall quality score for Item Q9 (LQMetrics = 75.00) also appeared lower than the overall quality scores for all other items (LQMetrics = 82.66 to 92.66).

4. Discussion

Taken together, the results provide empirical support for several conclusions regarding latent print proficiency testing. Primarily, findings suggest that current proficiency testing is not difficult. Examiners gave correct responses to 99.6% of test items during the present test administration, indicating that erroneous conclusions were exceedingly rare. Further, examiners generally described the test as fairly easy ($M = 4.27$ on an 11-point scale ranging from 0 = *Extremely easy*, to 10 = *Extremely challenging*) and expressed substantial confidence in their conclusions. Indeed, approximately three out of four examiners (77.5%) endorsed maximum confidence regarding the item they perceived as most challenging on the test. These findings could be interpreted to reflect not that proficiency testing is easy, but that the vast majority of participating examiners are highly skilled and proficient in their discipline. While this interpretation may be true, the objective quality metric scores of the prints on the test suggest that the majority were of high ("good") quality and, thus, not especially challenging nor representative of latent prints that may be seen in casework—consistent with previous work by Koertner and Swofford [22]. Table 1 indicates that 60 was the lowest quality score for latent prints (near the margin of "good" and "bad" categories) and 71 was the lowest for source prints (distinctly in the "good" category). We emphasize that these scores are typically negatively skewed. Furthermore, the U. S. Federal Bureau of Investigation [20] notes that the range of scores for "good" prints is

from 65 to 90 (if we infer that these qualitative categories originated from a database of real casework latent prints, it would follow that none of the prints had a latent quality score of over 90). Yet, a large number of proficiency test prints had scores of over 90; indeed, a majority had scores of 99. Thus, the exceedingly low observed error rate, examiner perceptions that items were relatively easy, and high objective print quality metrics suggest that latent print proficiency testing is not particularly difficult and may reflect a low bar for proficiency.

Results also suggest that examiner perceptions of test difficulty and examiner confidence are consistent with external metrics. Broadly, examiners described the test as fairly easy and were highly confident in their results, and with good reason; only 11 of 290 survey respondents made any errors on the test. More specifically, examiners rated item Q9 as the most difficult item on the test and Q9 was most frequently identified as the most challenging item. This item was responsible for over half of erroneous conclusions on the test (53.8% of errors were on item Q9) and contained the lowest source-print qualities of any test item. Thus, examiner perceptions aligned with objective indicators of item difficulty and quality suggesting that item Q9 was, in fact, the most difficult on the test. It is encouraging that examiner perceptions are consistent with external indicators of difficulty because examiners often decide to seek consultation or support depending on their perceptions of difficulty. Seeking assistance on complex or challenging cases is one method through which examiners can improve the accuracy of their conclusions.

Despite good overall alignment between examiner perceptions of difficulty and external metrics, respondents were nevertheless divided about which print on the test was most challenging. Nine of the eleven items received at least one “vote” for most challenging print. Thus, results suggest overall difficulty remains, to some degree, a subjective judgment that encompasses factors outside print quality and clarity. Examiner perceptions of item difficulty appeared to be associated with the average quality scores of all prints relating to that item (i.e., our overall quality score describing the latent print and source print images) in this small study. Findings indicate that examiners do integrate perceptions of print quality when assessing the difficulty of a latent print case. In fact, over half of the variance (56%) in examiner perceptions of difficulty can be explained by the quality scores of the latent and source prints. This suggests that print quality plays a significant role in how examiners’ perceive the difficulty of a latent print comparison, but that other aspects are also important (e.g., print distortion, identification vs. exclusionary conclusions).

Although examiner perceptions generally appeared consistent with other metrics, they did deviate from expected outcomes in some regards. Examiners most often identified concerns with latent print quality among test items they perceived to be most challenging (see Fig. 2). Moreover, concern regarding known-print quality was the most prevalent concern identified by examiners beyond the provided options. However, such general concerns are not supported by objective indicators of print quality. Indeed, latent prints received an average LQMetrics score of 74.44 and known prints received an average LQMetrics score of 92.41. Such scores are well into the range of what examiners consider to be “good” quality prints (i.e., scores 65–90). Item Q9 contained the lowest source-print quality score of all items at 71 (and 79; there were two images of the source print). Such a score, while on the low-middle end of “good” quality, indicates that latent prints of “bad” or “ugly” quality are not well represented on this test.

Additionally, examiners who made erroneous conclusions did not appear to perceive the test differently than others and did not demonstrate insight into their errors. Of the 11 examiners who made errors, most did not identify the erroneous item as the most difficult on the test or expressed maximum confidence in their erroneous conclusion. Further, many examiners identified print quality as the basis for their selection of the most challenging item despite metrics suggesting high overall print quality.

Experts have long held that proficiency tests are too easy and are not representative of actual casework (e.g., [1–4,11,22]). Current findings suggest that these two criticisms are closely related. Examiner perceptions of item difficulty explained 9% of the variability in perceptions regarding that item’s similarity to routine casework (a moderate effect size). Put simply, respondents perceived more difficult items to be more similar to routine casework. Additionally, the average quality scores of latent and source prints for an item explained approximately two thirds of the variance (67%) in examiner perceptions of similarity to casework. In light of aforementioned concerns that testing may be relatively easy and unable to generalize to real-world settings, the current findings suggest that latent print comparisons perceived as more challenging and more similar to casework contained prints of lower quality. In other words, examiners indicated that the items that most closely resembled real-world casework were also the hardest and contained prints of the lowest quality.

4.1. Limitations and future directions

These results must be placed in the context of the study’s limitations. We had access to data from only one latent proficiency test provided by one test provider. Although CTS is the leading provider of forensic science proficiency tests, we cannot generalize results from this test administration to all proficiency testing. Moreover, we know little about the examiners who responded to the current test and survey. While we do not have reason to suspect that survey respondents differed meaningfully from examiners who did not respond to the survey, there is scant research describing the “typical” proficiency test respondent. As Koehler [1] noted, it is unknown whether respondents work alone, in groups, or under close supervision while completing proficiency tests. Therefore, we cannot definitively assert that the current results represent independent responses from a representative sample of individual examiners. The National Commission of Forensic Science [5] acknowledged that not all purchased proficiency test results are reported externally because only some laboratories choose to disclose such information. It is important that future research clarify typical procedures behind proficiency testing to better examine the representativeness of proficiency test results and test respondents.

These results point to several areas of future work and research. First, this research should be replicated and include additional queries to better gauge how representative respondents are of the population of latent print examiners and how respondents take the test (e.g., individually, in groups, with or without verification). Second—to the extent that this test was representative of CTS latent print comparison proficiency tests in general—these results suggest that items on proficiency tests cluster on the low end of perceived difficulty and on the high end of print quality. Thus, future tests might include items with a wider range of print quality and clarity. Ideally, this range would reflect the range of print quality seen in real casework, which will be easier to determine as broader research on print quality metrics continues.

Making proficiency tests more similar to casework raises several issues. Among them is the need to make the range of decisions more analogous to casework as well. For instance, if CTS or other companies decide to include low quality prints on their proficiency tests, then allowing examiners to conclude that certain prints have insufficient information to conduct a comparison (i.e., “no value”) or that certain comparisons are “inconclusive” may be necessary. Opening up the range of conclusions on proficiency tests, however, depends on the field reaching some consensus about when these types of conclusions (i.e., no value, inconclusive) are appropriate. Further, if proficiency tests increase in difficulty, the field and its constituencies will need to determine what level of proficiency should be required. Consensus about the level of proficiency required of examiners should also be clearly communicated to external constituencies (the constituencies not involved in setting the level of proficiency), so they can understand how to interpret the results of proficiency testing. Until proficiency tests are designed to better reflect the range of difficulty seen in casework,

external constituencies—particularly legal audiences—should understand that an examiner's record of strong performance on proficiency tests does not necessarily imply expertise with respect to challenging comparisons.

Although the current study cannot offer comprehensive recommendations regarding ways to immediately improve proficiency testing, results do provide insight into controversial aspects of proficiency tests (i.e., difficulty and similarity to casework) and provide guidance for future research.

5. Conclusion

Our study suggests that respondents viewed this particular proficiency test as relatively easy and moderately similar to casework. Respondents' perceptions were largely consistent with objective quality metrics, which revealed that both known and latent prints on the test were of high quality (i.e., in the "good" category of LQMetrics) overall. The general agreement between examiner perceptions of difficulty and LQMetrics scores offers some evidence about the validity of this particular quality metric and adds to the scant literature on the subject (e.g., [22]). Further, the association between examiner perceptions of difficulty, LQMetrics scores, and similarity to casework suggests that the most representative test items contained difficult and unclear prints. We hope that the current study prompts further discussion and research regarding the appropriate methodology of latent print proficiency testing and the objective assessment of print quality.

Declarations of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement #70NANB15H176 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California Irvine, Duke University, and University of Virginia. The authors would like to thank Christopher Czyryca and Samantha Heise for their help organizing and conducting this study.

References

[1] J.J. Koehler, Fingerprint error rates and proficiency tests: what they are and why they matter, Hastings Law Journal 59 (5) (2008) 1077–1099.
[2] J.J. Koehler, Intuitive error rate estimates for the forensic sciences, Jurimetrics 57 (2017) 153–168.
[3] Mnookin, J.L., 2008. Of black boxes, instruments, and experts: Testing the validity of forensic science. Episteme J. Soc. Epistemol. 5, 343–358.
[4] W.A. Tobin, W.C. Thompson, Evaluating and challenging forensic identification evidence, Champion 12 (2006) 19–20.
[5] National Commission on Forensic Science, 2016a. Proficiency testing in forensic science. Retrieved from https://www.justice.gov/archives/ncfs/page/file/831806/download, 1–6.
[6] Collaborative Testing Services, Inc. (2019). Retrieved from https://cts-forensics.com/index-forensics-testing.php.
[7] G.S. Cembrowski, R.E. Vanderlinde, Survey of special practices associated with College of American Pathologists proficiency testing in the Commonwealth of Pennsylvania, Arch. Pathol. Lab. Med. 112 (1988) 374–376.
[8] D.J. Balding, P. Donnelly, Inferring identity from DNA profile evidence, PNAS 92 (1995) 11741–11745.
[9] National Academy of Sciences, National Research Council, Committee on DNA Technology in Forensic Science, 1992. DNA Technology in Forensic Science. Washington, DC: National Academies Press.
[10] W.C. Thompson, Evaluation the admissibility of new genetic identification tests: Lessons from the "DNA War", J. Criminal Law Criminol. 84 (1993) 22–104.
[11] J.J. Koehler, Proficiency tests to estimate error rates in the forensic sciences, Law Probab. Risk 12 (2013) 89–98, https://doi.org/10.1093/lpr/mgs013.
[12] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Accuracy and reliability of forensic latent fingerprint decisions, Proc. Natl. Acad. Sci. 108 (19) (2011) 7733–7738.
[13] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Repeatability and reproducibility of decisions by latent fingerprint examiners, PLoS ONE 7 (3) (2012).
[14] G. Langenburg, A performance study of the ACE-V process: a pilot study to measure the accuracy, precision, repeatability, reproducibility, and biasability of conclusion resulting from the ACE-V process, J. Forensic Identif. 59 (2) (2009) 219–257.
[15] M.B. Thompson, J.M. Tangen, D.J. McCarthy, Expertise in fingerprint identification, J. Forensic Sci. 56 (6) (2013) 1519–1530.
[16] J.J. Koehler, Why DNA likelihood ratios should account for error (even when a National Research Council report says they should not), Jurimetrics 37 (1997) 425–437.
[17] Bayles, A. (2002). Testimony in US v. Plaza, 188, R. Suppl. 2d, Daubert hearing.
[18] National Commission on Forensic Science, 2016b. Views of the commission, Optimizing human performance in crime laboratories through testing and feedback. Retrieved from https://www.justice.gov/archives/ncfs/page/file/864776/download.
[19] L. Haber, R.N. Haber, Scientific validation of fingerprint evidence under Daubert, Law Probab. Risk 7 (2008) 87–109.
[20] Federal Bureau of Investigation. (2015). Universal Latent Workstation (ULW) LQMetrics User Guide. Washington, DC. Retrieved from https://www.fbibiospecs.cjis.gov/Latent/PrintServices.
[21] R.A. Hicklin, J. Buscaglia, M.A. Roberts, Assessing the clarity of friction ridge impressions, Forensic Sci. Int. 226 (2013) 106–117.
[22] A.J. Koertner, H.J. Swofford, Comparison of latent print proficiency tests with latent prints obtained in routine casework using automated and objective quality metrics, J. Forensic Identifi. 68 (2018) 379–388.
[23] Collaborative Testing Services, Inc. (2017). Latent Print Examination Test No. 17-5171/2/5 Summary Report. Retrieved from https://cts-forensics.com/reports/37171_Web.pdf, 1–30.