2019

Latent Print Proficiency Testing: An Examination of Test Respondents, Test-Taking Procedures, and Test Characteristics

Brett O. Gardner
University of Virginia

Sharon Kelley
University of Virginia

Karen D.H. Pan
University of Virginia

Latent Print Proficiency Testing: An Examination of Test Respondents, Test-Taking Procedures, and Test Characteristics

Abstract

Proficiency testing is a key component of quality assurance programs within crime laboratories and can help improve laboratory practices. However, current proficiency testing procedures contain significant limitations and can be misinterpreted by examiners and court personnel (Garrett & Mitchell, 2018). To evaluate some of these limitations, we surveyed latent print examiners (n = 198) after they completed a Collaborative Testing Services, Inc. proficiency test. Additionally, we evaluated test performance and used a quality metric algorithm to evaluate the quality of test prints. Results do not suggest that respondents are dissimilar to the broader examiner population, although they may engage in different behaviors when completing tests versus casework. Findings show that proficiency testing contains prints of high quality and is perceived as both relatively easy and representative of casework. The test discriminated between inexperienced and experienced respondents, and verification procedures were largely ineffective in reducing errors. Objective quality metrics may provide a path forward to improving proficiency testing in a measurable manner.

Keywords

forensic science, latent prints, proficiency testing, quality metrics, human factors

Disciplines

Forensic Science and Technology

Comments

This is a manuscript of an article published as Gardner, Brett O., Sharon Kelley, and Karen DH Pan. "Latent Print Proficiency Testing: An Examination of Test Respondents, Test-Taking Procedures, and Test Characteristics." *Journal of forensic sciences* (2019). Posted with the permission of CSAFE.

JOURNAL OF FORENSIC SCIENCES

Latent Print Proficiency Testing: An Examination of Test Respondents, Test-Taking Procedures, and Test Characteristics

Journal:	*Journal of Forensic Sciences*
Manuscript ID	JOFS-19-404.R1
Manuscript Type:	Paper
Keywords:	Forensic Science, latent prints, proficiency testing, quality metrics, human factors, latent print examiners

Latent Print Proficiency Testing: An Examination of Test Respondents, Test-Taking Procedures, and Test Characteristics*

Brett O. Gardner,[1] Ph.D.; Sharon Kelley,[1] J.D., Ph.D.; and Karen D.H. Pan,[2] M.S.

[1]Institute of Law, Psychiatry, and Public Policy, University of Virginia, Charlottesville, VA.

[2]Department of Statistics, University of Virginia, Charlottesville, VA.

Corresponding author: Brett O. Gardner, Ph.D. E-mail: bgardner@virginia.edu

*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, and University of Virginia.

Running head: Latent Print Proficiency Testing

ABSTRACT: Proficiency testing is a key component of quality assurance programs within crime laboratories and can help improve laboratory practices. However, current proficiency testing procedures contain significant limitations and can be misinterpreted by examiners and court personnel (Garrett & Mitchell, 2018). To evaluate some of these limitations, we surveyed latent print examiners (n=198) after they completed a Collaborative Testing Services, Inc. proficiency test. Additionally, we evaluated test performance and used a quality metric algorithm to evaluate the quality of test prints. Results do not suggest that respondents are dissimilar to the broader examiner population, although they may engage in different behaviors when completing tests versus casework. Findings show that proficiency testing contains prints of high quality and is perceived as both relatively easy and representative of casework. The test discriminated between inexperienced and experienced respondents and verification procedures were largely ineffective in reducing errors. Objective quality metrics may provide a path forward to improving proficiency testing in a measurable manner.

KEYWORDS: forensic science, latent prints, proficiency testing, quality metrics, human factors

Accrediting agencies generally require that forensic science laboratories administer proficiency tests at regular intervals to assess analysts' competence as part of a larger quality assurance program. Such testing is important because proficiency tests can help identify problematic laboratory procedures, train forensic analysts, and ensure minimum levels of competency both within and across laboratories. Thus, routine proficiency testing is a useful tool that may be used to improve laboratory—and analyst—practices. However, in the absence of much research detailing the error rates of many forensic science disciplines (*see* [1]), analysts sometimes cite proficiency test results as evidence that their work in a real case is accurate (2). Further, proficiency test results are typically the only "objective" evidence offered to justify admitting the testimony of a latent print examiner in court. These claims may hold intuitive appeal, but are problematic for multiple reasons. Indeed, there are several important limitations to current proficiency testing that prevent one from generalizing test results to real-world work performance.

One concern in extrapolating proficiency test results to routine casework is that test respondents may not be representative of the broader population of examiners. Collaborative Testing Services, Inc. (CTS) is a primary provider of forensic proficiency tests, and there are no restrictions regarding who may order and complete a proficiency test. We therefore do not know whether examiners who complete proficiency tests are more or less experienced than practicing examiners. As demonstrated by a recent study in which practicing attorneys completed a CTS proficiency testing in latent print examination, we cannot even be certain that all respondents are forensic analysts (3). Scholars have noted that test respondents "should be representative of examiners who testify in court," but little is known about the training and work experience of

those who complete proficiency tests (4,[5], pp. 1091-1092). The observed error rates in current proficiency tests may therefore over- or underestimate real-world performance because we cannot be sure that the same population completes proficiency testing.

Another important limitation to current proficiency testing regards examiners' behavior during testing and routine casework. Put simply, examiners may behave differently during proficiency testing than during routine analyses (6). Indeed, an early study found that surveyed chemists, hematologists, and pathologists acknowledged engaging in special practices (e.g., using multiple instruments to conduct an analysis, dedicating additional time to an analysis) that they did not otherwise engage in during routine casework to ensure that they reached accurate conclusions during proficiency testing (7). Another study found that some laboratories compared proficiency test results before submitting final conclusions to national programs (8). The Clinical Laboratory Improvement Act of 1988 subsequently required laboratories to indicate that proficiency test samples were treated in a similar manner to materials processed during routine procedures. But this requirement does not apply to many forensic science disciplines whose analyses do not involve human material (e.g., latent print examination). Moreover, research suggests that forensic analysts may unknowingly alter their behavior under observation and due to the context of completing a proficiency test (see the literature documenting contextual effects in forensic sciences; e.g., 1,9,10). As a result, numerous scholars and professional associations have called for the implementation of blind proficiency tests in forensic crime laboratories, although such practice remains uncommon (1,4,11,12). To date, there is no information describing *how* analysts complete open proficiency testing (e.g., individual vs. group work) and whether such procedures resemble daily casework (12).

Finally, another limitation to current proficiency testing is that commercial test materials may differ in important ways from routine, real-world evidence. To use an analyst's performance on a proficiency test as a proxy for their performance in real-world casework, the items on proficiency tests should approximate samples encountered by examiners in daily casework (4,12). In the context of latent print examination, if examiners regularly process partial prints of low quality, then proficiency tests should also feature partial prints of low quality. However, there has been widespread speculation that the prints used in proficiency tests are much easier to examine than real-world prints, thus rendering proficiency tests unrepresentative of authentic cases (e.g., 3,13,14). To empirically assess this issue, a recent study (S. Kelley, personal communication, April 16, 2019) surveyed respondents who completed the Fall 2017 CTS latent print examination proficiency test and used a quality metric algorithm (i.e., LQMetrics) to objectively evaluate the quality of included prints. Examiners described the test as fairly easy and were exceedingly confident in their answers. Such perceptions appeared justified given the low error rate on the test (examiners correctly responded to 99.6% of items [3,177 of 3,190]) and high quality scores of included prints. Further, items perceived as similar to casework were also perceived as more difficult and contained prints of lower quality. Thus, recent research supports experts' concerns that current proficiency testing is unchallenging and suggests that decreasing the quality of included prints may both increase test difficulty and representativeness.

Current Study

Although the literature supports proficiency testing as a key component of quality assurance programs within laboratories (for a brief review, *see* 15), there are important limitations to current testing procedures. Without addressing such limitations, proficiency tests

may be a misleading metric of competence or proficiency in the field, and both internal oversight (e.g., lab managers) and external oversight (e.g., accrediting bodies) become less meaningful (16). In the present study, we expanded on a recent examination of latent print examination proficiency tests (S. Kelley, personal communication, April 16, 2019). We surveyed examiners on a recent CTS proficiency test to gather empirical data regarding test respondents, test-taking procedures, and test characteristics. Specifically, we sought to provide insight into: 1) whether those who complete latent print proficiency tests are representative of the broader latent print examiner population, 2) whether typical test-taking procedures resemble routine case processing procedures, and 3) whether included prints resemble prints found in daily casework. We also evaluated respondents' test performance and used a quality metric algorithm (LQMetrics) to objectively assess the quality of prints used in proficiency testing.

Method

Participants and Procedure

We collaborated with CTS to add survey questions to a latent print examination proficiency test that was shipped to respondents in January 2019. CTS received completed tests from 331 respondents a few months later. Given the international adoption of CTS proficiency tests, respondents were presumably latent print examiners who practice in multiple countries. Detailed demographic information regarding the examiners' age, sex, and ethnicity is unavailable because such information is not collected by CTS during the standard proficiency testing process. Of the respondents who submitted completed tests, we received survey responses from 59.8% (198 of 331). Thus, our final sample of latent print examiners who submitted survey responses was 198, although we evaluated the test responses of 331 examiners.

To provide an objective assessment of the CTS test, we also examined all depicted fingerprints using a global quality metric: LQMetrics. LQMetrics is included in the Federal Bureau of Investigation's (FBI) Universal Latent Workstation (ULW) and outputs three scores directly related to print quality (as described below) that are calculated from the information contained in a latent print (17).

Measures

CTS Latent Print Examination Proficiency Test

As the largest provider of latent print examination proficiency testing, CTS 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 most recent test, most respondents used physical copies of photographs (56.5%) to complete the test, although some examiners used images from a DVD (25.1%) or website (18.4%). Of those that completed the survey, 45.5% used physical photographs, 31.3% used images from a DVD, and 23.2% used images from a website.

Regardless of test format, respondents were provided 12 latent prints and four sets of known prints belonging to four different individuals. Examiners received a full-hand print (including palm prints) and a completed 10-print card (including rolled and simultaneous prints) in each of the four known-print sets. Thus, examiners had 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 [sometimes two print images were provided in full-hand image], individual prints on 10-print card, and simultaneous prints made in lower section of 10-print card).

To complete the proficiency test, examiners compare the prints and report their findings concerning each latent print, with each latent print representing one test item (12 total). On the

current test, nine latent prints were made by one of the four individuals who had provided known prints and three latent prints (items Q1, Q7, and Q9) were made by an individual whose prints were not provided. New to the spring 2019 proficiency test, CTS included non-powder developed and improperly oriented latent prints "in response to customer feedback for more challenging prints that represented the diversity of their casework" ([18], p. 3). CTS noted that the quality level for the non-powder developed prints was equitable to the other prints. Further, examiners were provided two identical images of non-powder developed latent prints (a standard purple print and a gray-scale print) of the two test items that employed non-powder developed prints (i.e., items Q5 and Q12). CTS indicated that they intend to continue providing similar prints in future tests.

Supplemental Survey

For purposes of this study, CTS included a brief survey at the end of the current proficiency test asking examiners to describe: 1) their experience in work, education, and testimony (e.g., "Typically, in how many cases do you testify on latent print evidence every year?"), 2) their proficiency test-taking procedures (e.g., "Approximately how many hours did you spend completing this test?"), and 3) their perceptions of test items. Specifically, participants separately rated the clarity of each latent print and comparison known print, as well as the level of challenge for each test item, using an 11-point scale ranging from 0 = *Extremely easy/poor clarity,* to 10 = *Extremely difficult/high clarity.*

Latent Quality Metrics (LQMetrics)

LQMetrics is included in the FBI's Universal Latent Workstation (ULW), an interactive software tool for latent print examiners. LQMetrics outputs three broad metrics directly relating to print quality and nine additional metrics calculated from a latent fingerprint (17; LQMetrics is

currently limited to fingerprint analysis, although the user guide notes that a future version will support analysis of palm prints). The three broad quality metric scores range from 0 to 100, and consist of: latent quality, value for individualization (VID), and value for comparison (VCMP). 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). For clarity, we refer to the overall latent quality score as the LQMetrics score in this study. A VID score represents the probability that an examiner would believe a print to have sufficient quality for individualization; and a VCMP score represents the probability an examiner would believe the print quality is sufficient for either individualization or exclusion. VID and VCMP scores are interpreted analogous to the overall LQMetrics score.

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 in one examination (17). Although 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), we limited our analyses to the three overarching quality scores summarizing a print's quality (i.e., latent quality score, VID, VCMP).

Results

Test Respondents

Most examiners that completed the proficiency test indicated that they held either a Bachelor's degree (46.0%) or a Master's degree (30.8%). Some examiners held an Associate's degree (10.1%) whereas fewer examiners did not endorse education beyond a high school

diploma (8.1%). Only two examiners (1.0%) indicated that they held a doctoral degree, and 4.0% either did not respond or endorsed another education level (e.g., "institute diploma").

On average, respondents reported that they had 11.9 years of experience (*Mdn* = 11.0; *SD* = 8.2) working as a latent print examiner. Respondents reported work experience ranged from 1 to 37 years. Regardless of education level or years of work experience, most examiners indicated that they had completed hundreds of hours of training regarding latent print examination. Indeed, only 7.6% (15 of 172) of examiners endorsed less than 100 hours of occupation-specific training.

Examiners' accounts of testimony experience varied widely, although most (80.7%) reported that they typically testify regarding latent print evidence on five or fewer occasions every year. More specifically, 12.3% of examiners indicated that they typically did not testify on a yearly basis, 19.8% testify once every year, 15.5% testify twice every year, 11.8% testify three times yearly, 7.0% testify four times yearly, and 8.6% testify an average of five years every year. Few examiners (9.1%; 17 of 187) reported that they typically testify on more than 10 occasions every year. Regarding the content of their testimony, examiners' responses indicated that they were more likely to be questioned about proficiency testing than error rates in their discipline. Approximately half of examiners (53.2%; 101 of 190) reported that they had been questioned about proficiency testing while testifying whereas only one-third of examiners (38.9%; 74 of 190) reported that they had been questioned about error rates in latent print examination. A chi square test of independence revealed a significant association between the two experiences, χ^2 (1, N = 189) = 46.53, $p < .001$, Cramer's $V = .50$. Examiners that had been questioned about proficiency testing were 4.59 times more likely to have also been questioned about error rates than examiners who had not been questioned about proficiency testing.

Test-Taking Procedures

Approximately one in four examiners (22.8%; 44 of 193) indicated that their laboratory required proficiency testing beyond that provided by CTS. Regarding the current proficiency test, examiners typically completed the work individually (96.9%; 190 of 196), although a few indicated that they completed the test as part of a group (3.1%; 6 of 196). We did not exclude responses from individuals who completed the test as part of a group because current proficiency testing procedures do not bar examiners from doing so. However, we acknowledge that the reported demographics for these six respondents may not fully represent the demographics of all involved examiners. Moreover, approximately two-thirds of respondents (65.8%; 127 of 198) reported that their test responses had been verified by another examiner before submission. Examiners typically reported that they completed testing in 9.5 hours on average (*Mdn* = 6.0; *SD* = 10.4). However, responses were positively skewed, with most respondents indicating that they completed the test in 8 hours or less (68.8%) but others endorsing time estimates ranging from 1 to 50 hours.

Test Characteristics

Subjective Assessment of Proficiency Prints

Across all test items, respondents indicated that they did not find the test particularly challenging (*M* = 3.5; *SD* = 1.5). They also described the clarity of the latent prints as relatively high (*M* = 6.8; *SD* = 1.7) and the clarity of source prints as even higher (*M* = 8.0; *Mdn* = 8.7; *SD* = 2.2). Despite respondents' report of the test's relative ease and fairly high level of print clarity, examiners indicated that the test was fairly representative of typical casework (*M* = 7.0; *SD* = 2.0; Modal response = 8). This may indicate that examiners similarly perceive routine casework

as fairly easy. However, examiners who perceived the test to be more difficult also perceived the test to be more representative of typical casework, $r[169] = .27$, $p < .001$.

As Figure 1 demonstrates, item Q7 ($M = 1.8$; $SD = 2.2$) was described as the least challenging item and contained the latent print with the highest perceived clarity ($M = 8.5$; $SD = 2.2$). Item Q9 was described as the most challenging item ($M = 6.3$; $SD = 2.7$) whereas item Q12 contained the latent print perceived to be of the lowest clarity ($M = 5.0$; $SD = 2.4$). There was less variability in examiner perceptions of source print clarity across items – item Q3 contained the source print with the highest perceived clarity ($M = 8.4$; $SD = 2.4$) and item Q2 contained the source print with the lowest perceived clarity ($M = 7.5$; $SD = 2.4$). Overall, perceptions of latent print and source print clarity were positively correlated, $r[159] = .67$, $p < .001$. Perceptions of item difficulty was negatively correlated with latent print clarity ($r[163] = -.26$, $p = .001$) but not source print clarity ($p = .16$)

Objective Assessment of Proficiency Prints

We calculated LQMetrics scores for 10 of 12 latent prints (two latent prints were palm prints and could not be scored via LQMetrics) and all provided known fingerprints ($n = 128$). As reported in Table 1, latent prints within the test were of fairly high quality ($M = 72.60$; $Mdn = 73.50$) and the quality of all known prints was even higher ($M = 89.83$; $Mdn = 94.00$). The source fingerprints (i.e., known prints determined to be the source of a latent print on the test; $n = 23$) were of similarly high quality ($M = 88.35$; $Mdn = 93.00$). Ultimately, LQMetrics scores for all included prints indicate that, as a whole, prints on the current CTS proficiency test belonged to the category of "good" quality prints. Indeed, three of the 10 latent prints, and only one of 23 source prints fell into the "bad" category of print quality; none fell in the "ugly" category.

Association Between Subjective and Objective Assessments of Prints

Examiner perceptions of item difficulty were marginally associated with LQMetrics scores for latent (r[8] = -.62, p = .05) and source fingerprints (r[5] = -.72, p = .07). Again, two test items contained two images (i.e., a purple and gray-scale image) of an otherwise identical non-powder developed latent print. LQMetrics scores for the two latent prints varied slightly, and we thus averaged the two quality scores to obtain one latent print quality score for these two test items. Examiner perceptions of latent and source print clarity were similarly associated with latent print LQMetrics scores (r[8] = .57, p = .09), and source print LQMetrics scores (r[5] = .75, p = .05), respectively. The lack of formal significance is likely due to the small number of analyzed prints (ns = 7 to 10). Interestingly, examiner perceptions of source print clarity appeared to be significantly associated with the quality scores of latent prints (r[5] = .89, p = .01).

We next averaged the quality metric scores for each latent print with the quality metric scores of its source fingerprints (when provided) to compute an overall quality score for each test item. We calculated overall print quality scores for 7 of 12 items. The overall print quality of test items was M = 83.78, with scores ranging from 72.20 (Q6) to 94.75 (Q3). A formal test of the association between this quality score and examiners' perceptions of item difficulty (r[5] = -.77, p = .04) suggested a non-zero association, even given the small number of prints. Although perceptions of latent print clarity were not associated with overall quality scores, perceptions of source print clarity appeared to be highly correlated with overall quality scores, r[5] = .92, p = .003.

Test Performance

Of the 331 examiners that submitted completed tests, the vast majority (95.5%; 316 of 331) did not provide an erroneous response to any test item. Almost 200 examiners also provided survey responses, and of those, 93.4% (185 of 198) did not provide an erroneous response to any test item. Only 13 respondents in the current sample made an error on the proficiency test, and only three respondents committed more than one error. This means that, of the 2,376 test items submitted in the current sample, there were only 21 erroneous conclusions (i.e., 99.1% overall accuracy rate across items).

The extremely high rate of accuracy on the current test is not atypical for CTS latent print proficiency tests (S. Kelley, personal communication, April 16, 2019), but the low occurrence of errors on the test also makes it difficult to predict which individuals will provide erroneous conclusions because virtually all respondents provided the exact same answers to test items. Nevertheless, we sought to explore whether examiners who gave erroneous conclusions differed from other examiners in their work experience, test-testing procedures, or perceptions of included prints.

Test Respondents

Examiners who committed an error on the proficiency test were less experienced (M = 7.5 years; SD = 6.9) than examiners who did not commit any errors (M = 12.2 years; SD = 8.2), $t(185) = 2.00$, $p = .047$, $d = .57$. Relatedly, examiners who reported less than 100 hours of occupation-specific training were 5.23 times as likely to commit an error (26.7% committed an error) as examiners who endorsed more than 100 hours of training (5.1% committed an error), χ^2 (1, N = 172) = 9.82, $p = .002$, Cramer's $V = .24$. Examiners who committed an error did not perceive the clarity of test prints differently than others ($ps \geq .76$), and only one examiner rated their erroneous item as the most difficult. However, they did perceive the test as generally more

difficult (M = 4.8; SD = 1.9) than those who did not commit an error (M = 3.4; SD = 1.4), $t(170) = 2.92$, p = .004, d = .95.

Test-taking Procedures

Examiners who committed an error reported spending the same amount of time completing the test as those that did not commit an error (p = .83). Of note, examiners who indicated that their responses had been verified by another examiner before submission were not significantly less likely to commit an error than examiners whose responses were not verified, χ^2 (1, N = 193) = 2.39, p = .12, Cramer's V = .11, although the small sample size may have contributed to the lack of significance. In total, 4.7% of examiners whose responses had been verified committed an error and 10.6% of examiners whose responses had not been verified committed an error.

Test Characteristics

Although we did not run formal analyses due to extremely small sample sizes, examiners seemed to perceive error-free items as less difficult (M = 2.4; SD = 0.8) than other test items (M = 4.4; SD = 1.1). Additionally, test items that all examiners accurately completed appeared to contain latent and source fingerprints of higher quality (overall LQMetrics score: M = 82.00; SD = 10.49) than other items (M = 66.75; SD = 13.30).

Discussion

Test Respondents

Some scholars have questioned whether those who complete latent proficiency tests are representative of the broader latent print examiner population (e.g., 12). At the same time, there is little empirical data regarding the characteristics of the examiner population. A recent study described the educational background of forensic analysts within the United States, and found that most held a Bachelor's (43%) or Master's degree (49%; *citation redacted for review*).

Regarding latent print examiners specifically, a large-scale study of examiner conclusions assessed 169 examiners and found that 50% of participants held a Bachelor's degree and 25% held a graduate or professional degree (19). These estimates appear fairly similar to the proportion of respondents in the current sample with Bachelor's (46%) or Master's degree (31%). Moreover, proficiency test respondents in the current study endorsed similar levels of work experience as a latent print examiner (*Mdn* = 11 years) compared to examiners within Ulery and colleagues' study (19; *Mdn* = 10 years). Finally, proficiency test respondents endorsed broadly equitable levels of testimony experience compared to participants in Ulery and colleagues' 2011 study. Whereas approximately 86% of proficiency test respondents endorsed testifying at least once yearly, 60% of participants in the 2011 study indicated that they had testified within the past year. Approximately 12% of respondents in the current study reported that they did not testify on a yearly basis compared to the 11% of examiners who indicated that they had never testified about a latent print identification. In sum, although literature describing the examiner population is scant, current findings do not suggest that latent print proficiency test respondents are especially dissimilar to the general population of latent print examiners. In any case, examiners who complete proficiency testing are not novices; respondents typically endorsed 11 to 12 years of work experience as a latent print examiner in addition to hundreds of completed training hours.

To our knowledge, this study is the first to describe the frequency with which latent print examiners testify about their work. Approximately two-thirds of examiners (66.5%) testify between one and five times every year. Somewhat surprisingly, more examiners indicated that they had been questioned about proficiency testing than about the actual error rates of their analyses. This suggests that despite the increased emphasis on error rates in forensic science in

recent years, most examiners (61.1%) are not questioned about the error rates of their discipline when communicating results in court. Of course, examiners cannot discuss topics during testimony that are not raised by attorneys, and the current results suggest that additional attorney training may be needed to facilitate comprehensive communication about latent print examination conclusions.

Test-Taking Procedures

A common criticism of current proficiency testing is that examiners are aware that they are being tested and therefore may behave differently during testing than during routine casework (e.g., 12). There is again limited research describing "typical" procedures during routine latent print examination casework beyond the traditional analysis, comparison, evaluation, and verification (ACE-V) process. In fact, there is little empirical data regarding the average time and effort required to perform latent print examinations. A past study (20) surveyed 56 laboratories and found that latent print examiners complete 35 cases per month on average. On the current proficiency test, examiners reported spending anywhere from 1 to 50 hours to examine and form conclusions regarding 12 latent prints. Although most spent 8 hours or less working on the test, some examiners (10.9%; 21 of 192) spent 24 hours or more. Further, a similar number of examiners reported spending one hour (4.7%) versus approximately one work-week (i.e., 40 hours or more; 5.2%) to complete the test. Although we do not know whether such estimates are representative of routine casework, results do suggest that the amount of effort on proficiency testing—as determined by time—among latent print examiners varies considerably.

The verification stage of case processing has received additional attention in recent years as the field has increasingly focused on error management and quality assurance procedures.

Still, few studies have documented "standard" verification procedures, with the President's Council of Advisors on Science and Technology 2016 report noting that verification procedures vary widely across laboratories (1). A survey of primarily U.S. laboratories found that nearly all (96%) indicated that they verify all identifications, but only one in three laboratories (36%) verified *all* conclusions (20). A more recent examination of one large crime laboratory in the U.S. found that most cases are not verified (56%) and only one in three cases (36%) receive complete verification (i.e., all conclusions are verified by another examiner; 21). Recognizing that verification procedures vary widely, it seems possible, or perhaps even likely, that complete verification rates are higher in proficiency testing (65.8% of tests) than in routine casework. Finally, there was also anecdotal support for discrepant test-taking practices during proficiency testing. For example, one respondent noted, "…Here in my Unit we use the ACE-V. Not for testing but for real case work which eliminates human error" ([18], p. 21). Taken together, the current findings suggest that proficiency test results are not necessarily representative of routine casework because at least some examiners engage in different practices during such tests. Results do not specify whether test results overestimate or underestimate real-world performance, although the exceptionally high rate of accuracy on the test (i.e., 99.1% across items) certainly does not suggest that results underestimate real-world performance.

Test Characteristics

CTS sought to make the current test more difficult by including several non-powder developed and improperly oriented latent prints. However, results indicate that the changes to the test did not significantly affect the difficulty of the exam. An earlier study of CTS proficiency testing (S. Kelley, personal communication, April 16, 2019) found that examiners described the test as fairly easy (M = 4.27 on an 11-point scale ranging from 0 = *Extremely easy,*

to 10 = *Extremely challenging*), endorsed high confidence in their conclusions, and rarely made an error (3.8% of respondents made an error). Moreover, latent prints received an average LQMetrics score of 74.44 (*Mdn* = 72.00) and known prints received an average LQMetrics score of 92.41 (*Mdn* = 96.00). On the current proficiency test, examiners again described the test as easy (*M* = 3.5), with no individual test item receiving an average difficulty rating above 6.3. Respondents described most prints as being of high clarity, and only 6.6% of survey respondents made an error on the test. Objective assessments of print quality also revealed that LQMetrics scores did not change appreciably (latent prints: *M* = 72.60; *Mdn* = 73.50; known prints: *M* = 89.83; *Mdn* = 94.00). Thus, providing examiners with latent prints developed through different techniques and in different orientations did not appear to have a meaningful impact on the difficulty of the test.

Overall, respondents reported that the test approximated real casework (*M* = 7.0). This finding runs counter to one of the growing criticisms of proficiency testing (i.e., the tests are too easy because they do not represent casework). In conjunction with relatively low difficulty ratings, this finding may suggest that much of examiners' routine casework is relatively easy and that, in certain respects, existing proficiency tests do capture the demands of a typical case. Of course, while the subjective perceptions of examiners are important, subsequent research should use objective metrics (e.g., quality metrics) to compare proficiency testing items against randomly selected prints from actual casework. Further, additional results offer a bit more depth to this finding.

Specifically, in keeping with recent research (S. Kelley, personal communication, April 16, 2019), results again revealed a moderate association between examiners' perceptions of test difficulty and representativeness of casework (*r* = .30 in previous study; *r* = .27 in current

study). Specifically, examiners who perceived the proficiency test as more difficult also thought that it was more representative of casework. Thus, there may still be some fodder to the argument that increasing proficiency test difficulty will also increase the test's ability to adequately represent casework. Additionally, current findings also replicate the large association between examiner perceptions of item difficulty and the objective quality score of that item (i.e., the average LQMetrics score of an item's latent and source prints). Indeed, both studies found large, significant effects ([15]: r = -.75 vs. current study: r = -.77). Moreover, the replicated associated between print difficulty and LQMetrics score demonstrates how test providers may impartially and systematically make tests more difficult. While subjective assessments of print difficulty and clarity can vary substantially (clarity ratings of the same print on the current test often ranged from 0 to 10), quality metrics can help test providers objectively evaluate and increase the difficulty of test items.

Test Performance

The vast majority of—but not all—respondents did not make any error on the proficiency test. An integral aspect of proficiency testing is providing insight into the characteristics and procedures of examiners who make erroneous conclusions. Importantly, the current test achieves this objective in that examiners who made an error on the test had less work experience and training than those who made no errors. Thus, although examiners perceive the test to be easy, it appears that the test is at least somewhat effective in differentiating examiners with more or less work experience. Further, examiners who made an error believed the test was more difficult, although they rarely identified the items on which they erred as being particularly difficult or unclear.

Finally, almost half of examiners who made an error on the test (46.2%) had their responses verified by another examiner. Put differently, verification procedures did not appear to significantly reduce the likelihood of committing an error on the current proficiency test. Of course, the current data relies on examiner self-report and we do not know whether all or only some test responses were verified, whether verifications were conducted blindly, or the time spent verifying test responses. A sizeable body of research supports verification—particularly blind verification—as an effective means to reduce erroneous fingerprint conclusions (e.g., 22). Nevertheless, this is concerning for those who view proficiency testing primarily as a system-level test of a laboratory's proficiency. If existing verification procedures do not reduce error rates, then changes to laboratory procedures are indicated.

Limitations and Conclusion

The current study addresses a dearth in the literature by describing the respondents who complete proficiency testing, test-taking procedures, and test characteristics. However, there are limitations that should be considered when interpreting results. There are still many unknown characteristics of the respondents (e.g., age, sex, laboratory setting), and examiners were not randomly selected in that those who chose to respond to the survey may differ from those who did not (although current findings do not suggest that this is the case). While results shed light on common test-taking procedures (e.g., verification practices, hours spent on test), we relied on self-reported information that may not be accurate in all cases. Further, the current effects (e.g., examiners who reported more training experience made less errors) may not consistently generalize to routine latent print examinations for the same documented reasons that performance on current proficiency tests may not generalize to routine laboratory performance (e.g., 4,12).

There are several concerns in extrapolating proficiency test results to real-world performance relating to characteristics of test respondents, test-taking procedures, and the test itself. Current results do not suggest that proficiency test respondents are dissimilar to the larger examiner population in terms of work experience or educational background, but at the same time, they may engage in different behaviors (e.g., verification) during test-taking compared to routine casework. Most respondents indicated that their responses on the test had been verified by another examiner, but verification procedures had little to no effect on test performance. Analyses of proficiency test characteristics are somewhat conflicted. Examiners described the test as easy and quality metrics suggest that almost all included prints are of high quality. However, examiners also indicated that testing is fairly representative of actual casework. Further, current proficiency testing discriminates between inexperienced and experienced examiners to some extent. Taken together, these findings urge caution in generalizing test performance to real-life performance, but also indicate that the current test achieves some primary objectives of proficiency testing procedures. Importantly, results provide insight into how to make tests more difficult by including prints of lower quality. Objective quality metrics may provide a path forward to increasing test difficulty and representativeness in a measurable and purposeful manner.

Acknowledgement

The authors would like to thank Christopher Czyryca and Samantha Heise for their help organizing and conducting this study.

References

1. President's Council of Advisors on Science and Technology. Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods. Washington, DC: Executive Office of the President of the United States, 2016.
2. Garrett BL, Mitchell G. The proficiency of experts. Univ PA Law Rev 2018;166(4):901–60.
3. Max BP, Cavise J, Gutierrez RE. Three attorneys participate in fingerprint proficiency testing: Results and observations. Proceedings of the Annual Scientific Meeting of the American Academy of Forensic Sciences; 2019 Feb 18-23; Baltimore, MD. Colorado Springs, CO: American Academy of Forensic Sciences, 2019.
4. Koehler JJ. Proficiency tests to estimate error rates in the forensic sciences. Law Probab Risk 2013;12(1):89–98.
5. Koehler JJ. Intuitive error rate estimates for the forensic sciences. Jurimetrics 2017;57:153–68.
6. Langenburg G. A performance study of the ACE-V process: A pilot study to measure the accuracy, precision, reproducibility, repeatability, and biasability of conclusions resulting from the ACE-V process. J Forensic Identif 2009;59(2):219–57.
7. Cembrowski GS, Vanderlinde RE. Survey of special practices associated with College of American Pathologists proficiency testing in the Commonwealth of Pennsylvania. Arch Path Lab Med 1988;112(4):374–6.
8. Boone DJ, Hearn T, Lewis S. Assessment of extent to which laboratories compare results before reporting in national laboratory performance surveys, abstracted. Clin Chem1985;31:115.

9. National Academy of Sciences, National Research Council, Committee on DNA Technology in Forensic Science. DNA technology in forensic science. Washington, DC: National Academies Press, 1992.
10. Risinger DM, Saks MJ, Thompson WT, Rosenthal R. The Daubert/Kumho implications of observer effects in forensic science: hidden problems of expectation and suggestion. Calif Law Rev 2002;90(1):1–56.
11. American Statistical Association. ASA board policy statement on forensic science reform, 2010. https://www.amstat.org/asa/files/pdfs/POL-Forensic_Science_Endorsement.pdf (accessed August 15, 2019).
12. Koehler JJ. Fingerprint error rates and proficiency tests: what they are and why they matter. Hastings Law J 2008;59(5):1077–99.
13. Bayles A. Testimony in *US v. Plaza,* 188, R. Suppl. 2d, Daubert hearing (2002).
14. Mnookin JL. Of black boxes, instruments, and experts: testing the validity of forensic science. Episteme: A Journal of Social Epistemology 2008;5(3):343–58.
15. Peterson JL, Lin G, Ho M, Chen Y, Gaensllen RE. The feasibility of external blind DNA proficiency testing: I. Background and findings. J Forensic Sci 2003;48(1):21–31.
16. Koehler JJ. Forensics or fauxrensics? Ascertaining accuracy in the forensic sciences. Ariz St L J 2017;49:1369–416.
17. U.S. Federal Bureau of Investigation. Universal Latent Workstation (ULW) LQMetrics user guide, 2015. https://www.fbibiospecs.cjis.gov/Latent/PrintServices (accessed August 15, 2019).

18. Collaborative Testing Services, Inc. Latent Print Examination Test No. 19-5161/2/5 summary report, 2019. https://cts-forensics.com/reports/19-5161.2.5_Web.pdf (accessed August 15, 2019).

19. Ulery BT, Hicklin A, Buscaglia J, Roberts MA. Accuracy and reliability of forensic latent fingerprint decisions. Proc Natl Acad Sci U S A 2011;108(19):7733–8.

20. Black JP. Is there a need for 100% verification (review) of latent print examination conclusions? J Forensic Identif 2012;62(1):80–100.

21. Rairden A, Garrett BL, Kelley S, Murrie D, Castillo D. Resolving latent conflict: what happens when latent print examiners enter the cage? Forensic Sci Int 2018;289:215–22.

22. Ulery BT, Hicklin A, Buscaglia J, Roberts MA. Repeatability and reproducibility of decisions by latent fingerprint examiners. PLOS One 2012;7(3):e32800.

TABLE 1—*Latent quality metrics for prints included on proficiency test.*

	Objective LQMetrics								
	Latent Quality Score			VID			VCMP		
Prints	***M***	***Mdn***	**Range**	***M***	***Mdn***	**Range**	***M***	***Mdn***	**Range**
Latent prints (n = 10; gray-scale)	72.60	73.50	(52 – 91)	97.20	98.50	(91 – 100)	99.50	99.50	(99 – 100)
Known fingerprints (n = 128)	89.83	94.00	(52 – 99)	99.38	100	(91 – 100)	99.90	100	(99 – 100)
• Full-hand prints (n = 48)	83.27	88.50	(52 – 99)	98.73	100	(91 – 100)	99.77	100	(99 – 100)
• 10-print card (n = 40)	98.88	99.00	(98 – 99)	100	100	-	100	100	-
• Simultaneous 10-print (n = 40)	88.65	90.50	(64 – 99)	99.53	100	(96 – 100)	99.95	100	-
Source fingerprints (n = 23)	88.35	93.00	(64 – 99)	99.35	100	(96 – 100)	99.83	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. On the current proficiency test, CTS provided images of two non-powder developed latent print for two items. The test items included both a purple and gray-scale image of an otherwise identical latent print. The color of the non-powder developed prints did not substantially affect the LQMetrics score (i.e., 64 vs 54 [gray-scale]; 74 vs 79 [gray-scale]).

Figure Legends

FIG. 1—*Perceived Difficulty and Clarity of Test Items. n = 163 to 174. 0* = Extremely easy/poor clarity, *10* = Extremely difficult/high clarity. *Items Q1, Q7, and Q9 do not have Source Clarity scores because a source print was not provided.*

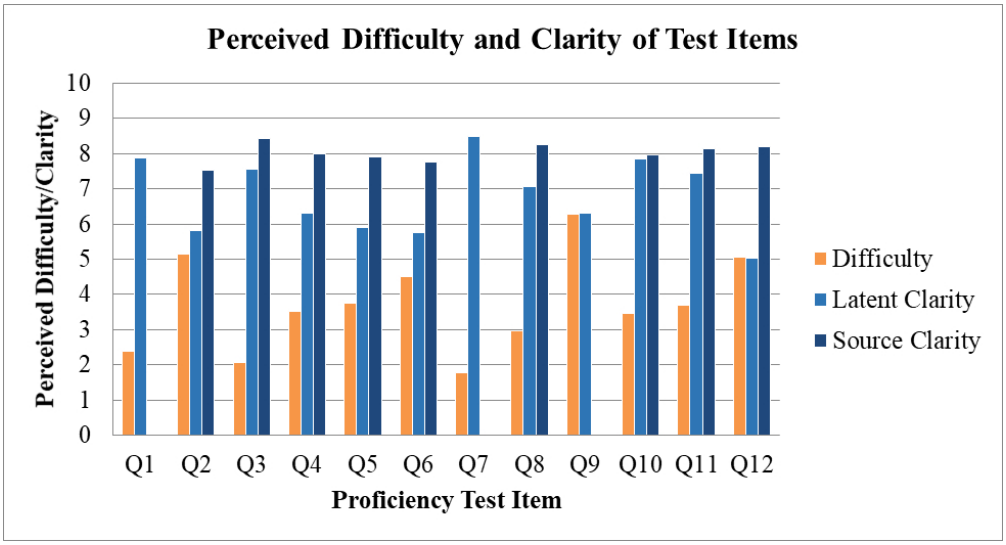

FIG. 1—Perceived Difficulty and Clarity of Test Items. n = 163 to 174. 0 = Extremely easy/poor clarity, 10 = Extremely difficult/high clarity. Items Q1, Q7, and Q9 do not have Source Clarity scores because a source print was not provided.