Latent print quality in blind proficiency testing: Using quality metrics to examine laboratory performance

Brett O. Gardner[a,*,1], Maddisen Neuman[b,c], Sharon Kelley[a]

[a] Institute of Law, Psychiatry, and Public Policy, University of Virginia, Box 800660, Charlottesville, VA 22908, USA.
[b] Houston Forensic Science Center, Houston, TX, USA
[c] Center for Statistics and Applications in Forensic Evidence (CSAFE), Ames, IA, USA

ARTICLE INFO

Article history:
Received 30 November 2020
Received in revised form 15 March 2021
Accepted 29 April 2021
Available online 7 May 2021

Keywords:
Latent prints
Blind proficiency testing
Quality metrics
Field study

ABSTRACT

Calls for blind proficiency testing in forensic science disciplines intensified following the 2009 National Academy of Sciences report and were echoed in the 2016 report by the President's Council of Advisors on Science and Technology. Both practitioners and scholars have noted that "open" proficiency tests, in which analysts know they are being tested, allow for test-taking behavior that is not representative of behavior in routine casework. This study reports the outcomes of one laboratory's blind quality control (BQC) program. Specifically, we describe results from approximately 2.5 years of blind cases in the latent print section (N = 376 latent prints submitted as part of 144 cases). We also used a widely available quality metrics software (LQMetrics) to explore relationships between objective print quality and case outcomes. Results revealed that nearly all BQC prints (92.0%) were of sufficient quality to enter into AFIS. When prints had a source present in AFIS, 41.7% of print searches resulted in a candidate list containing the true source. Examiners committed *no* false positive errors but other types of errors were more common. Average print quality was in the midpoint of the range (53.4 on a 0-to-100 scale), though prints were evenly distributed across the *Good*, *Bad*, and *Ugly* categories. Quality metrics were significantly associated with sufficiency determinations, examiner conclusions, and examiner accuracy. Implications for blind testing and the use of quality metrics in routine casework as well as proficiency testing are discussed.

1. Introduction

The National Academy of Sciences (NAS) released a congressionally mandated, and highly influential report, in 2009, which described multiple problems with current forensic science practice and resulted in widespread attention and calls for reform (e.g., [20]). One concern highlighted in the NAS report was the current state of proficiency testing in many disciplines. To receive accreditation, crime laboratories typically must administer proficiency testing to analysts as part of a quality assurance program. Proficiency testing serves an important role due to its ability to identify weaknesses in laboratory procedures, monitor performance, and confirm adequate competencies among individual analysts and across laboratories. However, as emphasized in the 2009 NAS report, "A particular need exists for routine, mandatory proficiency testing that emulates a realistic, representative cross-section of casework" (p. 25). There are several methods of administering proficiency tests, with a primary distinction between *open* proficiency tests—in which analysts are aware they are being tested—and *blind* proficiency tests—in which analysts are unaware they are being tested. The authors of the NAS report noted that proficiency testing in some disciplines "is not sufficiently rigorous" (p. 206), and the American Society of Crime Laboratory Directors/Laboratory Accreditation Board (ASCLD/LAB) has recommended that laboratories complete blind proficiency testing as a more precise means of assessing analyst proficiency (as cited in [18]).

In the years that followed the NAS report, other scholars have reiterated similar concerns about open proficiency testing and called for widespread implementation of blind testing (e.g., [1,13,14,20,23]). Critiques generally assert that open proficiency tests do not generalize to real-world casework because analysts' test-taking behavior is not representative of routine casework and because the tests are simply too easy. Indeed, an early study found that analysts behave differently during proficiency testing than during routine

* Corresponding author.
E-mail address: brett.o.gardner@gmail.com (B.O. Gardner).
[1] ORCID ID: https://orcid.org/0000-0001-5175-8960

https://doi.org/10.1016/j.forsciint.2021.110823

analyses involving casework ([3]; see also [7]), and numerous scholars have noted the lack of difficulty in proficiency tests (e.g., [2,13,14,15,16,17]).

1.1. Blind proficiency testing in practice

In 2015, the Houston Forensic Science Center adopted the recommendations made by numerous organizations (e.g., ASCLD/LAB, American Statistical Association) and detailed in the 2009 NAS report for blind proficiency testing by implementing a blind quality control (BQC) program [9]. The intent of the BQC program is to supplement open proficiency tests that are required for accreditation and to provide a way to monitor the entire quality management system from evidence submission to reporting of results. The program was implemented in the Latent Print section in November 2017, and Quality Division personnel routinely track case outcomes and print quality associated with all blind cases.

The BQC program is facilitated and maintained by HFSC's Quality Division, which is organizationally separate from the laboratory sections; thus, BQC cases are prepared and introduced into the workflow by personnel without any connection to the actual testing. The BQC cases are created to mimic real casework with the intent that the analysts will be completely unaware that the cases are not authentic and the cases will therefore receive no special treatment by the analysts. The target submission rate is 5% of the average number of cases completed per month during the previous year. Thus, the 5% submission goal for the Latent Print section was 10 BQC cases per month during 2018, and 9 cases per month during 2019 administered across the entire latent print unit.

1.2. Print quality metrics

In recent years, researchers have developed multiple quality metrics with the goal of objectively evaluating the quality and/or clarity of latent prints (for a brief review, see [22]). All quality metrics utilize algorithms that incorporate different aspects of prints to calculate a score (e.g., the number of features, ridge contrast, blur versus print clarity) and generally can be categorized as either a global or feature-specific metric. Global metrics such as the Latent Quality Metrics (LQMetrics) software within the FBI's Universal Latent Workstation (ULW) [8,24] and the Defense Fingerprint Image Quality Index (DFIQI) software [22] provide overall scores for the quality and clarity of an entire latent print. Feature- or minutiae-specific metrics provide individual scores for certain aspects of each print. For example, a quality metric developed by Peskin et al. [19] assesses the gradient of contrast intensity around a particular feature, and the smudge noise quality estimator metric (SNoQE) [21] assesses the relative noise (i.e., smudge vs. dryness) associated with print features. Kellman et al. [12] also developed several quantitative measures of prints (e.g., total area, block contrast) related to print quality.

As a whole, quality metrics are particularly needed in latent print comparison because the discipline relies heavily on human visual processing, perception, and judgment. Without print quality metrics, there are no objective methods to assess the representativeness of proficiency tests because assessment of print difficulty is dependent on subjective judgment. Quality metrics provide not only a quantitative metric of fingerprint quality, but a deterministic and objective score that is not dependent upon a single examiner. Despite the large number of anecdotal claims that proficiency tests are much easier than actual casework (e.g., [2,13,14,16,17]), there was no empirical support for these assertions until recently.

Scholars have used such metrics to evaluate open proficiency tests [7,11] and routine casework [15] in recent years. Two recent studies used global quality metric scores to evaluate Collaborative Testing Services (CTS) proficiency tests and discovered that, not only do examiners believe open proficiency testing is relatively easy, but the prints contained in such tests are generally of high quality [7,11]. Very few prints received quality scores suggesting poor quality. Although participants reported a perception that prints on the test generally reflected the demands of casework [7], using quality metrics, Koertner and Swofford [15] found that prints contained in open proficiency tests are not particularly representative of actual casework, with open proficiency tests containing prints of higher quality and lower complexity.

1.3. Current study

Research has begun examining open proficiency testing and, to a lesser extent, actual casework in latent print units [6,25]. But research has yet to examine results from blind proficiency programs due to the recency of their implementation. Despite strong intuitive and theoretical rationale for the argument that blind proficiency testing may better represent actual casework (and perhaps result in more accurate error rate estimates), there is no empirical data examining such programs. In this study, we sought to:

(1) Describe preliminary results from a blind proficiency testing program within a latent print unit of a crime laboratory.
(2) Examine the quality of prints submitted as part of the program via commonly used and widely available quality metrics software.
(3) Examine the potential association between latent print quality and resulting sufficiency determinations and conclusions.

2. Method

In the winter of 2017, HFSC implemented a blind quality control program in latent print comparison. Since its implementation, the Quality Division within the laboratory has developed and inserted 290 blind cases/requests for analysis into the latent print comparison unit as of August 4, 2020. Twenty cases were submitted first to latent print processing (i.e., cases in which the print was placed on a surface and first needed to be developed and photographed) and then proceeded through the laboratory's standard workflow to latent print comparison, and 270 cases were submitted as lift cards directly to the latent print comparison unit.

Of the 290 blind cases inserted into casework, we were able to obtain print images for 144 cases, with report dates spanning approximately two years (i.e., January 9, 2018 to January 8, 2020).[2] All print images were scanned by practicing latent print examiners as part of routine casework. Some images contained non-minutiae photo-editing performed by the examiner. However, we were unable to identify which images had been edited for every print.

In total, examiners reviewed 376 latent prints submitted as part of the 144 blind cases/requests for analysis. Most blind cases involved only one latent print; however, some cases involved as many as 13 latent prints. The median number of prints in blind cases was 2 (SD = 2.41). The majority of latent prints were fingerprints (94.3%; n = 350) or palm prints (4.9%; n = 18). Very few were joint impressions or unspecified impressions (0.8%; n = 3).[3]

[2] Per the sectional standard operating procedure, examiners are only required to maintain images of prints that were annotated electronically; thus, prints that were not examined (e.g., prints that were deemed to be of no value for comparison purposes) are not routinely retained.

[3] The remaining 5 of 376 prints were not attributed to an anatomical source because examiners determined them to be of no comparative value and did not consider them to be latent prints.

2.1. HFSC laboratory procedures

Standard operating procedures within HFSC are likely unique from many laboratories—specifically the use of *Preliminary AFIS Associations*, described below—though the initial workflow is likely more common. Upon request for comparison, examiners first make a determination about the print's suitability for comparison and sufficiency to be searched in an automated fingerprint identification system (AFIS). At this stage, examiners may conclude that a print: (1) has no comparative value (and further analysis is therefore precluded), (2) has comparative value, but is of insufficient quality to be searched in AFIS, or (3) has comparative value, and is of sufficient quality to be entered into AFIS.

If a latent print is searched in AFIS, examiners subsequently make one of three conclusions depending on the AFIS candidate list outcome[4] and further examination. Examiners may conclude that there is *No Association*, meaning the latent print does not appear to correspond to any print on the AFIS candidate list. Conversely, examiners may, based upon corresponding characteristics between the latent print and the candidate image, conclude that the latent print may have originated from the same source as a candidate image. This conclusion is referred to as a *Preliminary AFIS Association* (PAA) and is provided to the requesting agency. If desired, the requesting agency may request a confirmatory comparison of a PAA conclusion, and the latent print is fully examined for agreement or disagreement of features before an "official identification" is declared. To the authors' knowledge, the use of PAAs as investigative leads is uncommon and was implemented to increase examiner efficiency (i.e., examiners only confirm preliminary identifications when requested to do so, thereby reducing workload in some cases). HFSC emphasizes that PAAs are not identifications and simply represent investigative leads; official identifications never result from PAAs alone. At the same time, confirmatory comparisons of PAAs have consistently resulted in official identifications with only one exception. During the data collection period for the current study, 336 PAA conclusions were confirmed (as part of routine casework *and* the BQC program), with all resulting in official identifications. Finally, examiners may conclude that an AFIS search resulted in a *Reverse Hit*. When an AFIS search does not result in a PAA, the print may be registered to the unsolved latent file (ULF). Both the local AFIS, IDEMIA (formerly MorphoTrak), and federal AFIS, Next Generation Identification (NGI), can register latent impressions to their respective ULFs. The latent prints registered to the ULF are continuously searched against new record prints. In sum, examiners may indicate potential identifications by concluding *Preliminary AFIS Association* or *Reverse Hit*, or may indicate that no potential identification exists by concluding *No Association.*

2.2. Blind quality control (BQC) program procedures

A detailed description of HFSC's BQC program for latent print comparison is provided by Hundl and colleagues [9]. In brief, the Houston Police Department (HPD) is the primary agency that submits requests to HFSC, so BQC cases are created to mimic HPD submissions in the submission process, packaging, and request wording. Before submitting BQC samples, a worksheet with relevant case information (e.g., subject name, offense type, date) is generated. No aspects of the cases include authentic case information, but Quality Division personnel inform their cases through data contained in real-world cases. For example, the offense date and time are created to be within a reasonable proximity to the submission date, and the offense type is informed by common offenses in actual casework (e.g., burglary, theft).

HFSC was granted permission by the Harris County Sheriff's Office to enter five sets of record prints containing fictitious individual information into the local AFIS. HFSC staff volunteers contribute their prints and the donors are provided an alias in AFIS. Though some BQC cases have an associated alias, others do not (i.e., they are submitted by donors without an alias in the local AFIS). The majority of the evidence received by the Latent Print Comparison section is in the form of latent lift cards submitted by HPD officers, and therefore the BQC program uses identical latent lift cards. Moreover, the latent lift cards are packaged in a manner that mimics HPD packaging and, with the aid of HFSC's evidence technicians, is submitted into the workflow in the same way as real evidence.

Examiners' performances on BQC cases are deemed satisfactory or not. For cases with an alias, examiners' performance is dependent on multiple factors, including limitations of the AFIS algorithm and the quality of the latent print. If determined to be of sufficient quality, the latent print is searched in the local AFIS. If the search results in a PAA and no confirmation is requested, the analysis is deemed satisfactory if the PAA identifies the correct alias. Cases that result in a PAA can be requested for full confirmation. Full confirmations are deemed satisfactory if an identification is reported to the correct alias. If a print searched in AFIS does not result in a hit, then the candidate list generated by AFIS is reviewed to determine whether the appropriate alias was on the list. If the alias was not on the list, then the analysis is deemed satisfactory. Submissions with no alias in AFIS are deemed satisfactory if the analyst concludes that there was no AFIS association.

2.3. Latent Quality Metrics (LQMetrics) software

Latent Quality Metrics (LQMetrics) is a global metric that is among the more widely used quality metrics available. Because of its increasing use in practice and empirical research (e.g., [7,11]), we employed LQMetrics in the current study.

To objectively evaluate the quality of each latent print, we entered all obtained images into FBI's Universal Latent Workstation (ULW) LQMetrics software ([8,24]; see [10] for additional information about its development). LQMetrics is a software tool for latent print examiners that outputs four broad metrics relating to print quality. The four quality metric scores range from 0 to 100, with higher scores indicating higher quality. The overall latent Quality score represents the predicted probability that an "image-only search" would return a candidate list containing the correct mate, assuming the mate is of sufficient quality and the images sufficiently overlap. As an example, a Quality score of 65 is interpreted as a 65% chance that a search returns the correct mate. Compared against qualitative assessments of quality, prints assessed to be of "good" quality corresponded to latent quality scores of 65–90, "bad" to scores of 45–65, and "ugly" to scores of 20–45 [24]. The Value for Individualization (VID) score represents the probability that an examiner would determine a print to be of sufficient quality for individualization, and the Value for Comparison (VCMP) score represents the probability that an examiner would determine a print to be of sufficient quality for either individualization or exclusion. Finally, the overall Clarity score does not represent a probability, but indicates the level and quantity of friction ridge detail, with larger scores indicating greater level and/or quantity of detail.

3. Results

3.1. Sufficiency determinations

Of the 376 latent prints submitted as part of 144 blind cases, 92.0% (n = 346) were determined to be of sufficient quality to enter

[4] The local, county-wide AFIS used by HFSC generates a list of 10 candidates. The state-level and federal AFIS algorithms output a list of 5 candidates, with examiners reporting that they rarely examine candidates not identified in the AFIS lists.

Table 1
Latent Quality Metrics and examiner sufficiency determinations of submitted prints.

	Latent Quality Metrics			
Sufficiency determinations	Quality	Clarity	VID	VCMP
AFIS Quality (n = 343–345)	55.0 (9–99)	35.0 (11–87)	89.7 (53–100)	98.2 (92–100)
Insufficient AFIS Quality (n = 23)	40.9 (11–80)	28.2 (11–50)	80.4 (59–99)	96.4 (92–100)
No Comparative Value (n = 6)	15.0 (9–30)	18.5 (12–29)	62.7 (58–79)	93.0 (92–96)
Total (n = 372–374)	53.4 (9–99) SD = 20.8	34.3 (11–87) SD = 12.5	88.6 (53–100) SD = 12.1	98.0 (92–100) SD = 2.3

Note: Mean scores are provided, latent quality metric score ranges are in parentheses. VID = Value for Individualization score. VCMP = Value for Comparison score.

into AFIS, and 6.1% (n = 23) were determined to be of comparative value, but insufficient quality to enter into AFIS. Very few prints were determined to be of no comparative value (1.6%; n = 6) or of value only for exclusionary purposes (0.3%; n = 1). It is important to note that we only reviewed available images of latent prints uploaded by examiners, and examiners do not upload all reviewed lifts. As a result, our sample generally represents *latent prints entered into AFIS*, but excludes many reviewed lifts that were determined to be of no comparative value and therefore were not uploaded and available for review. In other words, our sample represents all lifts reviewed by examiners and determined to contain latent prints, but is not inclusive of all lifts reviewed by examiners. The few images we obtained of prints deemed to be of no comparative value were uploaded because the examiner initially suspected that they may be of value, before ultimately deciding that they were not.

3.1.1. Sufficiency determinations and Latent Quality Metrics

Table 1 reports the mean quality metrics of all latent prints. Overall, prints received average Quality scores of 53.4 (SD = 20.8) and average Clarity scores of 34.3 (SD = 12.5), although scores varied widely. When categorized using traditional print quality categories [24],[5] submitted prints were evenly distributed across *Good* (35.7%; n = 133), *Bad* (30.6%; n = 114), and *Ugly* (33.8%; n = 126) categories.[6] Nevertheless, average VID scores suggest that all prints were more likely than not to be determined to be of sufficient quality for individualization. Indeed, the print with the lowest quality was still predicted to have a 53% chance of being deemed sufficient quality.

Table 1 also demonstrates the association between quality metrics and examiner determinations of sufficiency. All metrics were significantly associated with sufficiency determinations and exhibited similar patterns of association. Quality scores were significantly associated with sufficiency determinations, $F(2, 369) = 16.73$, $p < .001$, $\eta p^2 = 0.08$. Bonferroni post hoc analyses indicated that prints determined to be AFIS quality (n = 343) received higher quality scores than did prints determined to be of insufficient quality for AFIS entry (n = 23; d = 2.02; p = .003), and these latter prints received higher quality scores than prints determined to be of no comparative value, in turn (n = 6; d = 1.26; p = .02). Clarity scores ($F(2, 371) = 8.37$, $p < .001$, $\eta p^2 = .04$), VID scores ($H(2) = 21.99$, $p < .001$), and VCMP scores ($H(2) = 24.41$, $p < .001$) were also significantly associated with sufficiency determinations. These metrics did not significantly differ between prints determined to be of insufficient quality for AFIS entry and prints determined to be of no comparative value (Clarity: d = 1.07; p = .26; VID: d = 1.28; p = .09; VCMP: d = 1.27; p = .09), but this is likely due to the small sample size of the latter (n = 6).

Put more simply in categorical terms, Fig. 1 indicates that prints belonging to the *Ugly* quality category were less likely to be determined to be of sufficient quality for AFIS entry and were more likely to be of no comparative value than other prints. However, *Good* and *Bad* quality prints did not differ in their sufficiency determinations.

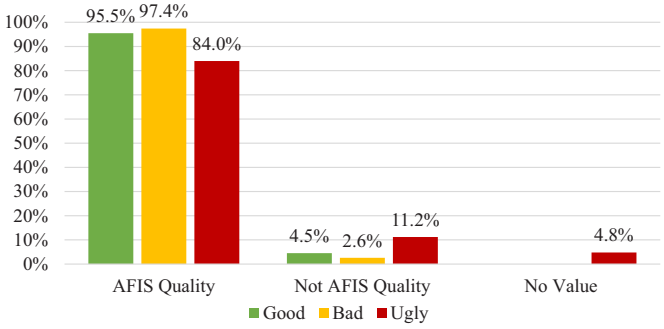

Fig. 1. Examiner sufficiency determinations among *Good*, *Bad*, and *Ugly* latent prints. Note: N = 372 prints. There were 133 *Good* prints, 114 *Bad* prints, and 125 *Ugly* prints.

3.2. Examiner conclusions

Table 2 depicts examiner conclusions for the latent prints submitted as part of HFSC's BQC program. Despite knowledge of ground truth, it may be misleading to describe all instances that did not result in conclusions consistent with ground truth as errors because the quality of submitted prints varied widely and, therefore, AFIS candidate lists did not always include the corresponding alias. At the same time, all instances in which prints are entered into AFIS and the resulting candidate lists do not include the correct source warrant further exploration. This table indicates that, based on ground truth, examiners arrive at the correct conclusion in about half of submitted prints (51.1%; n = 192). Examiners did not commit any false positive errors (i.e., concluding a potential association exists when the source was absent), but 41.0% of prints were deemed to have no association with other prints despite the source being in AFIS (i.e., as an HFSC alias). An additional 8.0% of prints were not entered into AFIS due to determinations of insufficient quality.

Of the 302 prints with a true source present in AFIS, 41.7% (n = 126) of prints resulted in the correct source being displayed among the top 10 AFIS candidates. When listed, the correct candidate print was almost always listed first (n = 85) or second (n = 9), although the source candidate was listed as low as 10th on one occasion. There were two occasions when the source candidate was listed in the AFIS results (in first and second positions) and the examiner concluded that no association existed. Therefore, we can conclude that, when the AFIS candidate list returned the correct source print, the false negative error rate in the current dataset is 1.6%.

3.2.1. Examiner conclusions and Latent Quality Metrics

Table 3 details the average Quality and Clarity scores of prints according to examiners' ultimate conclusions regarding each print. Prints that resulted in correct *Preliminary Associations* were of the highest quality and most clear, followed closely by prints that resulted in correct *No Hit* conclusions. Prints that did not result in AFIS searches were of the lowest quality and least clear.

[5] The FBI (2015) suggest cut scores of 45 and 65 to delineate *Ugly*, *Bad*, and *Good* print quality categories.

[6] LQMetrics could not provide a score for three images in our study and we therefore omitted these from analyses involving quality metric scores.

Table 2
Examiners' conclusions regarding latent prints in the blind quality control program.

	Ground truth	
Examiner conclusion	Association (302 source-present prints)	Exclusion (74 source-absent prints)
Preliminary Association	33.0%; *n* = 124 (*Correct Association*)	0.0%; *n* = 0 (*False Positive*)
No Association	41.0%; *n* = 154 (*False Negative*)	18.1%; *n* = 68 (*Correct Exclusion*)
No AFIS Search	6.4%; *n* = 24 (*Potential False Inconclusive*)	1.6%; *n* = 6 (*Potential False Inconclusive*)

Note: *N* = 376 latent prints submitted as part of blind quality control program.

Table 3
Latent Quality Metrics and examiners' conclusions within the blind quality control program.

	Ground truth		
Examiner conclusion	Association (302 source-present prints)	Exclusion (74 source-absent prints)	Total
Preliminary Association	Quality: 63.0 Clarity: 39.2 (*Correct Association*)	N/A (*False Positive*)	Quality: 63.0 Clarity: 39.2
No Association	Quality: 48.4 Clarity: 31.7 (*False Negative*)	Quality: 55.0 Clarity: 35.0 (*Correct Exclusion*)	Quality: 50.4 Clarity: 32.7
No AFIS Search	Quality: 35.8 Clarity: 26.6 (*Potential False Inconclusive*)	Quality: 35.7 Clarity: 24.5 (*Potential False Inconclusive*)	Quality: 35.8 Clarity: 26.2
Total	Quality: 53.4 Clarity: 34.4	Quality: 53.4 Clarity: 34.1	

Note: *N* = 376 latent prints submitted as part of blind quality control program. Mean scores are provided.

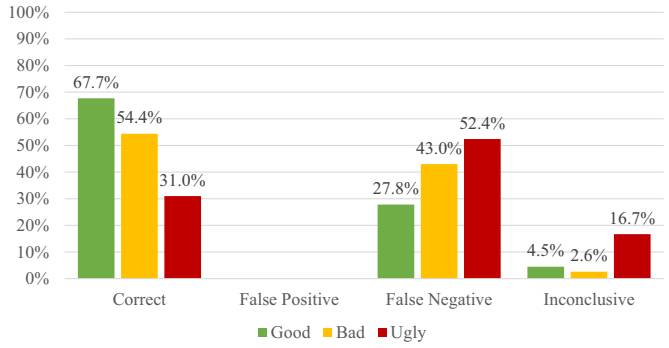

Fig. 2. Examiner conclusions among *Good*, *Bad*, and *Ugly* latent prints. Note: *N* = 373 prints. There were 133 *Good* prints, 114 *Bad* prints, and 126 *Ugly* prints.

Put more simply, print quality (as categorized by *Good, Bad,* or *Ugly*) was significantly associated with examiner conclusions and ultimate accuracy, $\chi^2(4, N = 373) = 34.01$, $p < .001$, Cramer's $V = .25$, 95% CI [.19, .32]. As depicted in Fig. 2, *Good* latent prints were 2.18 times more likely to result in correct conclusions than were *Ugly* prints, whereas *Ugly* prints were 3.71 times more likely to result in inconclusive conclusions (i.e., no AFIS searches) than were *Good* prints.

3.3. Case-level sufficiency determinations and examiner conclusions

Of 144 submitted BQC cases, the large majority (95.8%; *n* = 138) contained at least one print determined to be of sufficient quality to enter into AFIS. This means that only 4.2% (*n* = 6) of submitted blind cases did not result in an AFIS search. Examiners sometimes correctly associate only one latent print of nine in a case. However, one association is sometimes all that is required for a case to proceed with the correct conclusion (i.e., identification of correct subject) after a full comparison is performed. Therefore, assessing error at the case level, rather than the print level, is perhaps more reflective of 'real world' outcomes in that correctly identifying only a portion of provided prints can still lead to an accurate outcome in a case.

HFSC's BQC data at the case level leads to a higher overall accuracy rate compared to data at the print level. Table 4 describes examiner conclusions at the case level for prints submitted as part of the BQC program. As demonstrated, at the case level, at least 60.4% of blind cases result in correct conclusions based on ground truth (as compared to 51.1% of blind *prints* resulting in correct conclusions based on ground truth).

4. Discussion

The results presented here reflect the outcomes of one latent print section's blind quality assurance program over approximately two and a half years. Outcomes from this program—which has creatively inserted cases into the section's workflow to mimic real casework—offer a glimpse into the complete process of latent print comparison, from evidence submission to reporting of results. Importantly, the results also reflect outcomes when examiners were truly blind (i.e., unaware that the cases they were working were quality control cases rather than genuine casework). Consistent with previous research, these BQC cases suggest low rates of false positive errors (zero in the present study) and a higher rate of false negative errors, the overwhelming majority of which involved true sources *not* being included among the top 10 candidates after an AFIS search (rather than examiners failing to identify the correct source on a candidate list).

Examiners determined that nearly all latent prints examined in this data set (92.0%) were of sufficient quality to enter into AFIS. When prints had a source present in AFIS (through an HFSC alias), slightly less than half of prints (41.7%) had the source displayed in the top 10 AFIS candidates, generally in one of the two top positions. Examiners themselves committed no false positive errors (i.e., indicating an association between a latent and an incorrect source) and only two false negative errors (i.e., failing to indicate an association when the correct source was on the candidate list). Finally, when examined at the case level, rather than the print level—which is perhaps more analogous to 'real world' outcomes—overall accuracy rates increased from about 50% to approximately 60%.

These results highlight two limitations inherent in evaluating the outcomes of many blind programs. The first involves how to determine whether examiners correctly labeled prints as insufficient quality for AFIS or as being of no comparative value. Scholars have noted that these types of inconclusive determinations should not be

Table 4
Examiners' case-level conclusions in the blind quality control program.

	Ground Truth	
Examiner conclusion	Association (123 source-present cases)	Exclusion (21 source-absent cases)
Preliminary Association	45.8%; $n = 66$ (*Correct Association*)	0.0%; $n = 0$ (*False Positive*)
No Association	35.4%; $n = 51$ (*False Negative*)	14.6%; $n = 21$ (*Correct Exclusion*)
No AFIS Search	4.2%; $n = 6$ (*Potential False Inconclusive*)	0.0%; $n = 0$ (*Potential False Inconclusive*)

Note: One case involved both a source-present and source-absent source. This case was coded as target-present in this table.

automatically deemed correct [5], but with few objective criteria to guide these determinations it is difficult to evaluate them. In the current sample, we see that the objective quality of prints deemed to be sufficient and insufficient quality for AFIS entry varied widely; *Ugly* prints with Quality scores as low as 9 were nonetheless determined to be of sufficient quality and, conversely, *Good* prints with Quality scores as high as 80 were deemed to be of insufficient quality for AFIS entry. Moving forward, blind proficiency programs such as the one implemented at HFSC may address this limitation by conducting a priori examinations of all submitted prints in order to determine whether prints "should" be deemed of sufficient quality for AFIS entry and "should" appear on AFIS candidate lists. Alternatively, perhaps prints with extremely high Quality scores "should" consistently be determined to be of sufficient quality for AFIS entry. Such procedures would allow the BQC program to more effectively test the accuracy of the entire system. The current findings reveal that, as currently implemented, HFSC's BQC program does not effectively test the accuracy of examiner sufficiency determinations.

The second highlighted limitation concerns AFIS algorithms and the interaction of examiners' markup and AFIS output. As referenced above, based on ground truth, examiners arrived at "correct" conclusions for about half (51.1%) of BQC prints. For the other half of prints, the examiner/AFIS interaction and AFIS limitations appear to have influenced a large proportion of outcomes. Notably, when the correct source was present within the AFIS system as an HFSC alias, the source was *not* among the top 10 AFIS candidates in over half of searches (58.3% of prints). Certainly, some portion of these results might be explained by poor quality of the source print or insufficient overlap between the latent and source print (or some combination thereof). At the same time, results raise questions about how examiners' markup affects AFIS output, and limitations in the AFIS algorithm that can lead to false negative errors. As currently implemented, HFSC's blind proficiency program does not effectively test the accuracy of some examiner conclusions (i.e., no association conclusions and situations in which prints are not searched in AFIS), nor the reasons why some AFIS searches do not produce candidate lists that include the correct source when the correct source is within AFIS. Again, conducting a priori examinations of all submitted images, while time intensive, may address some aspects of this system limitation.

4.1. Quality metrics and BQC prints

Regarding LQMetrics, the mean Quality score of 53.4 indicates that the average BQC print falls in the *Bad* category, though scores varied widely and prints were evenly distributed across the *Good*, *Bad*, and *Ugly* categories. In order for blind proficiency programs to effectively evaluate laboratory performance in routine casework, submitted prints must be representative of routine casework. To this end, there is little available research describing the objective quality of latent prints evaluated in routine casework, but recent literature provides some context. The average LQMetrics Quality score of prints submitted as part of HFSC's BQC program (M = 53.4) is certainly lower than the average Quality scores of latent prints contained in recent open proficiency tests (Ms = 72.6 and 74.4) [7,11]. But more directly, the average LQMetrics Clarity score of submitted prints (M = 34.3) appears quite similar to average Clarity scores of a sample of 215 latent prints taken from normal casework at a United States-based federal laboratory (M = 36.6) [15][7]. Thus, current findings support anecdotal arguments suggesting that the clarity of prints submitted as part of HFSC's BQC program closely resemble the clarity of actual casework and are likely more representative than the prints contained in widely available open proficiency tests.

Quality metrics were associated with sufficiency determinations following an expected pattern, with AFIS-quality prints receiving higher scores than prints deemed insufficient quality for AFIS, which in turn received higher scores than prints of no comparative value. Categorically, *Ugly* prints were less likely to be useable (in AFIS or otherwise) than prints in the other two categories. Print quality was also strongly associated with examiner conclusions (following an AFIS search) and ultimate accuracy. Notably, *Good* prints were more than twice as likely to result in correct conclusions as were *Ugly* prints.

Finally, LQMetrics results also point to potential AFIS limitations (or limitations in how AFIS is used by examiners), just as BQC findings did. Of the 302 prints with a source present in AFIS, the average Quality score was 53.4, meaning that an image-only search is predicted to return the correct source on the AFIS candidate list 53.4% of the time. However, only 41.7% ($n = 126$) of prints resulted in the correct source being displayed in the top 10 AFIS candidates. Of course, Quality scores are probabilistic—such that there will always be *some* difference between expected results and actual results—but the discrepancy is noteworthy nevertheless.

Overall, these results highlight the utility of blind quality assurance programs as a supplement to open proficiency testing and to note areas of improvement. When properly implemented, such programs have the potential to more effectively test the accuracy of the entire system, including AFIS, compared to standard open proficiency tests. This research also suggests a potential role for incorporating quality metrics into blind programs and routine casework. First, quality metrics offer objective indicators that could help ensure blind cases closely resemble routine casework. Second, based on results documenting the relationship between LQMetrics and sufficiency determinations and examiner conclusions, there could be merit in screening prints for quality as a first step in an analysis. Given the relationship between *Ugly* prints and inconclusive outcomes (e.g., finding the print to be of insufficient quality for AFIS entry), and reduced accuracy in outcomes as compared to *Good* prints – analysts might consider not proceeding with *Ugly* prints, and instead saving their time and decisional efforts for higher-quality prints.

4.2. Limitations and future research

These data reflect a blind quality assurance program implemented within a single laboratory and detail findings of a particular AFIS. From an implementation perspective, HFSC has the

[7] Koertner and Swofford [15] did not report LQMetrics Quality scores in their study.

advantage of being a larger laboratory with a dedicated quality management section, and is able to negotiate circulating aliases through a local AFIS. We recognize that not all laboratories will be able to implement a blind program in exactly this way.

The sample of analyzed BQC prints also reflects all latent prints entered into AFIS, but, for the most part, excludes lifts that were not entered (e.g., prints deemed to have no comparative value). Thus, while results sufficiently represent all *latent prints* submitted as part of the BQC (i.e., prints deemed to be of AFIS quality), we know less about all *lifts* submitted as part of the BQC (i.e., reviewed images that analysts deemed to be of insufficient quality for AFIS entry and therefore were not labeled as latent prints). In this regard, it is possible that some very low-quality prints were dismissed based upon examiner judgment and not included in this sample. It is likely that this sample overrepresents images of high quality and relatively low complexity. At the same time, we again note that the clarity of the current sample appears similar to routine casework in federal laboratories [15].

Future research would, ideally, evaluate the full range of lifts submitted as part of the BQC program. Additionally, subsequent research might explore how to evaluate inconclusive determinations at the sufficiency or conclusion stage. Results suggest that quality metrics could be implemented to help guide this process. For instance, inconclusive determinations on low quality prints are likely to be more "accurate" than inconclusive determinations for higher quality prints. The ambiguity about how best to evaluate these inconclusive decisions also speaks to the importance of shifting away from categorical outcomes generally and towards reporting conclusions probabilistically, though the practice is not currently the norm in latent print comparisons [4].

Finally, results also speak to the need for additional research on interactions between examiner markup and AFIS output and the accuracy of different AFIS algorithms. Although evidence from this study and others does not point to concerns about false positive errors, AFIS algorithms certainly contributed to false negative errors. Thus, future research might explore practices to reduce the false negative error rate.

4.3. Conclusion

In sum, these results demonstrate the utility of a blind quality control program in evaluating the entire system of latent print comparison, from evidence submission to reporting of results, and note procedural areas needing improvement. Although we recognize HFSC's program as it exists cannot be replicated exactly in every laboratory, we hope these results speak to the feasibility of implementing blind testing and the importance of doing so. These data reveal important components of case processing (from sufficiency determinations to final conclusions) and have implications for understanding error at the system level. Notably, while explicit examiner errors were minimal (i.e., no false positive errors and only two false negative errors), results underscore limitations associated with AFIS algorithms and with current blind proficiency procedures that should be explored in future research.

Funding

This work was partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College and University of Nebraska, Lincoln.

CRediT authorship contribution statement

Brett O. Gardner: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing, Visualization, Project administration **Maddisen Neuman:** Investigation, Data curation, Writing - original draft, Writing - review & editing **Sharon Kelley:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing.

Conflict of interests

As detailed on the title page, the third author (Maddisen Neuman) works at Houston Forensic Science Center, which is the crime laboratory of examination in the current study. Beyond employment at the study location, the authors have no conflicts of interest to disclose.

References

[1] American Statistical Association, ASA board policy statement on forensic science reform, 2010. https://www.amstat.org/asa/files/pdfs/POL-Forensic_Science_Endorsement.pdf.
[2] A. Bayles, Testimony in US v. Plaza, 188, R. Suppl. 2d, Daubert hearing, 2002.
[3] 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.
[4] S.A. Cole, M. Barno, Probabilistic reporting in criminal cases in the United States: a baseline study, Sci. Justice 60 (2020) 406–414, https://doi.org/10.1016/j.scijus.2020.06.001
[5] I.E. Dror, N. Scurich, Commentary on: I. Dror, N Scurich "(Mis)use of scientific measurements in forensic science" Forensic Science International: Synergy 2020 https://doi.org/10.1016/j.fsisyn.2020.08.006, Forensic Sci. Int. Synerg. 2 (2020) 701–702, https://doi.org/10.1019/j.fsisyn.2020.08.006
[6] B.O. Gardner, S. Kelley, M. Neuman, Latent print comparison and examiner conclusions: a field analysis of case processing in one crime laboratory, Forensic Sci. Int. 319 (2021) 1–6, https://doi.org/10.1016/j.forsciint.2020.110642
[7] B.O. Gardner, S. Kelley, K.D.H. Pan, Latent print proficiency testing: an examination of test respondents, test-taking procedures, and test characteristics, J. Forensic Sci. 65 (2020) 450–457, https://doi.org/10.1111/1556-4029.14187
[8] A. Hicklin, J. Buscaglia, M.A. Roberts, Assessing the clarity of friction ridge impressions, Forensic Sci. Int. 226 (2013) 106–117.
[9] C. Hundl, M. Neuman, A. Rairden, P. Rearden, P. Stout, Implementation of a blind quality control program in a forensic laboratory, J. Forensic Sci. 65 (2019) 814–822.
[10] N. Kalka, M. Beachler, A. Hicklin, A latent fingerprint quality metric for predicting AFIS performance and assessing the value of latent fingerprints, J. Forensic Identif. 70 (2020) 443–463.
[11] S. Kelley, B.O. Gardner, D.C. Murrie, K.D.H. Pan, K. Kafadar, How do latent print examiners perceive proficiency testing? An analysis of examiner perceptions, performance, and print quality, Sci. Justice 60 (2020) 120–127, https://doi.org/10.1016/j.scijus.2019.11.002
[12] P.J. Kellman, J.L. Mnookin, G. Erlikhman, P. Farrigan, T. Ghose, E. Mettler, D. Charlton, I.E. Dror, Forensic comparison and matching of fingerprints: using quantitative image measures for estimating error rates through understanding and predicting difficulty, PLOS One 9 (2014) 1–14.
[13] J.J. Koehler, Fingerprint error rates and proficiency tests: What they are and why they matter, Hastings Law J. 59 (5) (2008) 1077–1099.
[14] 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
[15] 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 Identif. 68 (2018) 379–388.
[16] B. Max, J. Cavise, R.E. Gutierrez, Assessing latent print proficiency tests: Lofty aims, straightforward samples, and the implications of nonexpert performance, J. Forensic Identif. 69 (2019) 281–298.
[17] J.L. Mnookin, Of black boxes, instruments, and experts: Testing the validity of forensic science, Epistem.: A J. Soc. Epistemol 5 (2008) 343–358.
[18] National Academy of Sciences, Strengthening Forensic Science in the United States: A Path Forward, National Academies Press, Washington, DC, 2009.
[19] A.P. Peskin, K. Kafadar, A. Dima, A quality pre-processor for biological cell images, in: G. Bebis (Ed.), Advances in Visual Computing. ISVC 2009. Lecture Notes in Computer Science, 5876 Springer, Berlin, Heidelberg, 2009.
[20] President's Council of Advisors on Science and Technology, Report to the President: Forensic science in Criminal Courts: Ensuring Scientific Validity of

Feature-Comparison Methods, Executive Office of the President of the United States, Washington, DC, 2016.
[21] R. Richter, C. Gottschilich, L. Mentch, D.H. Thai, S.F. Huckemann, Smudge noise for quality estimation of fingerprints and its validation, IEEE Trans. Inf. Forensics Secur. 14 (2019) 1963–1974, https://doi.org/10.1109/tifs.2018.2889258
[22] H. Swofford, C. Champod, A. Koertner, H. Eldridge, M. Salyards, A method for measuring the quality of friction skin impression evidence: Method development and validation, Forensic Sci. Int. 320 (2021) 110703, https://doi.org/10.1016/j.forsciint.2021.110703
[23] W.A. Tobin, W.C. Thompson, Evaluating and challenging forensic identification evidence, Champion 12 (2006) 19–20.
[24] US Federal Bureau of Investigation, Universal Latent Workstation (ULW) LQMetrics User Guide. Washington, DC, 2015. https://www.fbibiospecs.cjis.gov/Latent/PrintServices.
[25] A. Rairden, B.L. Garrett, S. Kelley, D. Murrie, A. Castillo, Resolving latent conflict: What happens when latent print examiners enter the cage? Forensic Sci. Int. 289 (2018) 215–222, https://doi.org/10.1016/j.forsciint.2018.04.040