NATIONAL COMMISSION ON FORENSIC SCIENCE

NIST National Institute of Standards and Technology U.S. Department of Commerce

Views of the Commission
Optimizing Human Performance in Crime Laboratories through Testing and Feedback

Subcommittee
Human Factors
Status
Initial Draft

Date of Current Version	24/05/16
Approved by Subcommittee	27/05/16
Approved by Commission	[*dd/mm/yy*]

Note: This document reflects the views of the National Commission on Forensic Science, and does not necessarily represent the views of the Department of Justice or the National Institute of Standards and Technology. This document does not formally recommend any action by a government entity, and thus no further action will be taken upon its approval by the Commission.

Overview

How might testing be done to assess and improve the performance of forensic science service providers (FSSPs) on routine analytic tasks? What kind of testing programs will be most helpful for achieving goals such as: (1) establishing the conditions under which analytic methods can (and cannot) be expected to achieve accurate results; (2) providing feedback to help examiners improve their skills; (3) estimating the rate of error for routine analytic tasks and better understanding variables that affect the rate of error? What steps should be taken to facilitate the development of effective performance testing in forensic laboratories in order better to achieve such goals?

1. Why Performance Testing is Needed and What it Might Accomplish

Forensic science plays a vital role in the criminal justice system. It is therefore essential that forensic science service providers take steps to assure the accuracy and reliability of their analyses and the overall quality of their work. Crime laboratories should strive to be *high reliability organizations*.[1]

[1] See, Roberts, Some Characteristics of High-Reliability Organizations. *Organization Science*, 1, 160-177 (1990); Weick, K. E., & Sutcliffe, K. M. (2007). *Managing the Unexpected: Resilient Performance in and Age of Uncertainty, Second Edition*. San Francisco, CA: Jossey-Bass. The term "high reliability organization" (HRO) is used by management and organizational theorists to describe organizations that have succeeded in avoiding catastrophic failures in environments in which such failures are possible. Among the most important characteristics of HROs are their intensive focus on identifying potential sources of error and their efforts to evaluate their own performance in an open, forthright manner that avoids blaming and stigmatization.

Human judgment plays a central role in forensic science. In order for forensic science to realize its full potential for contributing to justice, practitioners must make good judgments about a variety of matters, such as what evidence to collect, what examinations and tests are needed, how to interpret results, and how to report and explain the results.

These tasks are challenging. The accuracy, reliability and validity of forensic science evidence depend on human performance on these tasks. The quality of forensic evidence suffers when human performance on these tasks is less than optimal.

In order to achieve and maintain optimal levels of performance on challenging tasks, people need feedback. They need information on how well they are doing; they need to be told when their judgments are sound and when they are mistaken, incomplete, or otherwise sub-optimal.[2] Research on human performance, and practical experience in a variety of domains, has consistently shown the importance of feedback in optimizing human performance on such tasks.[3]

For some challenging intellectual tasks feedback is automatic. A pilot who makes a navigation error, for example, is likely to learn of the error quickly when the plane fails to reach the expected destination. For many of the most vital judgments made by forensic science service providers, however, feedback is not automatically available, or is incomplete and unreliable. This is particularly true for analytic judgments. An analyst evaluating whether a latent print was made by a suspect, for example, will rarely know with certainty the ground truth of the matter.[4] There is only one consistently reliable way to provide valid feedback to such analysts on the accuracy of their performance—they must be tested using samples for which the ground truth is known.
Tests of this type have the potential to achieve a variety of important goals. The design and implementation of the tests would vary according to the particular goal sought, but all tests would involve laboratory analysis of samples of known origin in order to assess the accuracy of the analytic results.

[2] These principles are widely recognized in personnel management. *See e.g.*, the statement on the importance of providing feedback to employees by the United States Office of Personnel Management: https://www.opm.gov/policy-data-oversight/performance-management/performance-management-cycle/monitoring/feedback-is-critical-to-improving-performance/

[3] Thorndike, E. L. (1927). The law of effect. *American Journal of Psychology,39,* 212-222; Ammons, R. B. (1956). Effects of knowledge of performance: A survey and tentative theoretical formulation. *Journal of General Psychology,54,* 279-299; Annett, J. (1969). *Feedback and human behaviour*. Harmondsworth, Middlesex, England: Penguin Books. Arkin, R. M., & Walts, E. A. (1983). Performance implications of corrective testing. *Journal of Educational Psychology, 75,* 561-571; Causer, Barach & Williams (2014). Expertise in medicine: using the expert performance approach to improve simulation training. *Medical Education*, 48(2): 115-23. .

[4] While it has been argued that the results of latent print analysis are tested through the adversarial process of the trial itself, *see e.*g., *US v. Havvard*, 260 F.3d 597 (7th Cir. 2001) , this kind of "testing" is woefully inadequate from a scientific perspective. Risinger, D.M. & Saks, M.J. (1996) Science and Non-Science in the Courts: Daubert Meets Handwriting Identification Expertise 82 *Iowa L. Rev.* 21, at 33-34 and 41 fn. 100. One cannot draw reliable inferences about the accuracy of the forensic science evidence presented in a criminal case from the conviction or acquittal of a defendant; indeed, such arguments are tautological to the extent that the forensic science evidence contributed to the legal outcome that provides the "test" of its validity.

(1) Validation

One possible goal of such tests is assessment of the circumstances under which analytic methods can (and cannot) produce accurate results.[5] A testing program could be designed to explore the boundary conditions within which forensic assessments of a given type are accurate, and beyond which accuracy suffers. This would require that the tests involve samples (or test comparisons) that are expected to be highly challenging—samples that will test the limits of the method. Easy tests that labs always get right would have little value for learning the conditions under which the analytic method will fail. The tests must be difficult enough to induce errors.[6]

(2) Training and Improvement

Highly challenging tests of the type just described would also be valuable for helping experienced examiners improve their skills. For example, latent print examiners sometimes need to make critical decisions about whether a low quality latent print (e.g., a print containing limited detail or distortions) can accurately be identified, or whether the comparison should be deemed inconclusive. Feedback on their accuracy when making such judgments would be invaluable for helping examiners improve their decision making in such cases.

(3) Error Rate Estimation

In order to evaluate the probative value of forensic science evidence for proving a proposition (e.g., that a latent print was made by a particular individual), it is important to know the rate at which laboratory analysis produces erroneous conclusions. The 2009 NAS report declared: "The assessment of the accuracy of conclusions from forensic analyses and the estimation of relevant error rates are key components of the mission of forensic science."[7] Performance testing involving samples of known origin could provide valuable information about the rate at which errors occur in various types of analysis.[8] Testing for the purpose of error rate estimation should involve samples designed to replicate those routinely encountered in casework.[9] It would be misleading to attempt to estimate error rates from performance on samples designed to be unusually challenging and difficult. However, it might well be useful to estimate the rate of error separately

---

[5] In 2009, the National Academy of Sciences commented on the need for such tests. It called for research to "address issues of accuracy, reliability, and validity in the forensic science disciplines," saying that such research is needed to "establish the limits of reliability and accuracy that analytic methods can be expected to achieve as the conditions of forensic evidence vary." National Academy of Sciences (2009) *Strengthening Forensic Science in the United States: A Path Forward*. Washington, D.C.: The National Academies Press, p. 23, Recommendation 3 (hereinafter 2009 NAS report)

[6] Engineers often test products and systems by subjecting them to conditions (e.g., stress, strain, pressure) in excess of normal service parameters. Known as "accelerated life testing," this process is useful for uncovering faults and potential modes of failure in a relatively short time. *Nelson, W. (1980). "Accelerated Life Testing - Step-Stress Models and Data Analyses". IEEE Transactions on Reliability (2): 103.doi:10.1109/TR.1980.5220742;* Donahoe, D.; Zhao, K.; Murray, S.; Ray, R. M. (2008). "Accelerated Life Testing". *Encyclopedia of Quantitative Risk Analysis and Assessment*.doi:10.1002/9780470061596.risk0452.ISBN 9780470035498. Including challenging samples in a performance testing program for forensic laboratories would have similar benefits.

[7] 2009 NAS report, at 122.

[8] See Jonathan J. Koehler, Forensic Science Accuracy: Why We Know So Little and How to Learn More. Unpublished Manuscript, Northwestern University School of Law, available at: http://ssrn.com/abstract=2773255

[9] *Id*.

for different types of comparisons or classes of samples if the level of difficulty (and hence the expected rate of error) differed for different types of cases or different types of samples.

(4) Proficiency Testing

At the present time, most laboratories require analysts to take periodic proficiency tests. These tests are vital for assuring that analysts have the minimal level of competency needed to perform at a satisfactory level, but current proficiency testing programs do not achieve many of the important benefits that could be achieved with more comprehensive performance testing programs. Proficiency tests have several limitations: analysts know they are being tested (which may cause them to perform differently during proficiency tests than when performing casework); the tests involve relatively few samples; and the tests are typically designed to be relatively easy for a competent analyst to pass.[10] In their current form, proficiency tests have limited value for establishing the limits of reliability and accuracy that analytic methods can be expected to achieve as the conditions of forensic evidence vary. [11] These tests provide little useful feedback to forensic analysts on the limits of their expertise when dealing with difficult cases or marginal evidence and hence have little value for helping experienced analysts hone and improve their skills. They are not designed to determine error rates.[12]

(5) Quality Control and Quality Assurance

In clinical medicine, routine testing of laboratory performance has been a prominent feature of laboratory analysis since 1988, when Congress passed the Clinical Laboratory Improvement Amendments (CLIA). CLIA has been credited with bringing about remarkable improvement in the quality of clinical testing; it brought an end to scandals that previously plagued the field of medical testing. [13] CLIA requires medical laboratories to participate in routine blind proficiency testing using samples supplied by designated test providers. The proficiency test samples must be tested in the same manner as patient specimens, at the same time as patient specimens, by the same personnel that routinely test the patient specimens, and using the same test system that is routinely

[10] The President of Collaborative Testing Services, an organization that provides test samples that are widely used for proficiency testing in forensic laboratories, told the Commission during its seventh meeting (August 10, 2015) that he has been under commercial pressure to make proficiency tests easier.

[11] Proficiency tests are all-too-often analogous to the driving tests one must pass in order to obtain a driver's license. They test whether drivers possess a minimal level of competency, not how well they can drive under challenging conditions. The typical test required to obtain a driver's license would be useless for assessing how well drivers perform on a high speed race track, or when it would be safe for a driver to negotiate icy mountain roads. These tests serve a valuable function by weeding out truly incompetent drivers, but they do not provide typical drivers the kind of feedback they need to improve their skills, nor do they provide insights into when, due to challenging conditions, it becomes unsafe even for a competent operator to drive.

[12] Indeed, Collaborative Testing Service specifically warns that its proficiency tests were not designed and should not be used for error rate estimation. Collaborative Testing Services, Inc.CTS STATEMENT ON THE USE OF PROFICIENCY TESTING DATA FOR ERROR RATE DETERMINATIONS, March 30, 2010 at 3, http://www.ctsforensics.com/assets/news/CTSErrorRateStatement.pdf ("The design of an error rate study would differ considerably from the design of a proficiency test. Therefore, the results found in CTS' Summary Reports should not be sued to determine forensic science discipline error rates").

[13] Westgard, J.O. & Westgard, S.A. (2006). The quality of laboratory testing today: An assessment of metrics for analytic quality using performance data from proficiency testing surveys and the CLIA criteria for acceptable performance. *American Journal of Clinical Pathology*, 125(3) 343-354. http://dx.doi.org/10.1309/V50H4FRVVWX12C79

used for the patient specimens.[14] Feedback from testing has helped laboratory managers identify weaknesses in laboratory systems, identify weaknesses in training or preparation of staff, and detect problems with equipment and procedures. In light of the vital role testing has played in both quality management and quality improvement in medical laboratories[15], it seems likely that a similar program of testing could prove helpful in forensic science.

2. Challenges in Implementing Performance Testing in Forensic Laboratories

A number of challenges must be faced in order to develop an effective performance testing program. What follows is a discussion of some of the key challenges and how they might be overcome.

(1) Blinding of bench-level examiners

In an ideal testing program, the bench-level analysts, whose judgments are critical to the test results, would not know they are being testing. The test samples would be incorporated into routine casework so that analysts would not be able to distinguish tests from actual casework.[16]

Blind testing of forensic scientists is difficult in laboratories where examiners communicate directly with detectives and have access to police reports and other information. To conduct such a test, laboratory managers need to enlist the support of police in preparing simulated case materials. The materials must be sufficiently realistic to pass as a real case. Furthermore, the police might need to provide other false information to the examiner to prevent the examiner from discovering that the case was simulated. Elaborate simulations of this type are feasible and have been conducted successfully to test the accuracy of DNA analysis[17], but they are burdensome and expensive. They may also be problematic in other ways. Some observers may have qualms, for example, about involving police officers in the creation of false or simulated evidence.

Fortunately, blind testing is much easier to implement in laboratories that employ a context management system to shield examiners from task-irrelevant contextual information. If bench-level examiners are typically exposed only to the specimens presented for analysis, along with any task-relevant information passed on by the case manager, the case manager can easily insert a test case occasionally without the examiner being able to distinguish the test from routine casework. The case manager would need to know it was a test; the examiner would not. Successful blind testing programs of this type have been implemented in forensic laboratories.[18] The ability to

[14] A convenient summary of CLIA testing requirements can be found in a brochure prepared by the Center for Medicare and Medicaid Services: https://www.cms.gov/Regulations-and-Guidance/Legislation/CLIA/Downloads/CLIAbrochure8.pdf

[15] Westgard, J.O. & Westgard, S.A. (2006), op cit.

[16] Informing someone that they are being tested can create what psychologists call demand characteristics that change the person's responses. Orne, Martin T. (1962). "On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications.". American Psychologist 17 (11): 776–783. doi:10.1037/h0043424. Individuals who know they are being tested may shift their threshold of decision in ways designed to make them look good. Paulhus, D.L. (1991). Measurement and control of response biases. In J.P. Robinson et al. (Eds.), *Measures of personality and social psychological attitudes.* San Diego: Academic Press. Hence, performance testing will provide a more realistic picture of analytic performance if the analysts do not know they are being tested.

[17] Peterson, J.L., Lin, G., Ho, M. Ying, C., & Gaensslen, R.E. (2003) The feasibility of external blind DNA proficiency testing I. Background and findings. Journal of Forensic Sciences, 48(1) 21-31.

[18] For example, the Houston Forensic Science Center has been conducting blind testing in three of its disciplines: controlled substance, blood alcohol and firearms analysis. It is planning to expand the blind testing program to

implement blind testing in this way is a secondary benefit that arises when laboratories adopt context management systems, as recommended by this Commission (see, *Ensuring that Forensic Analysis is Based on Task-Relevant Information*).

(2) Developing Suitable Test Sets

In order to implement rigorous blind performance testing programs, forensic laboratories will need access to suitable test sets of evidentiary specimens for which ground truth regarding origin is known. To test latent print examiners, for example, laboratory managers will need access to latent prints of known origin that can be inserted into routine casework and compared with same-source and different-source exemplar prints. Similarly, to test DNA analysts laboratory managers will need known-source biological specimens and to test firearms examiners, they will need known-source bullets and shell casings.

The test sets should be designed with the help of experts in statistics and human factors in order to facilitate research on important questions regarding the accuracy of forensic analysis. Different test sets would be needed according to the goals of the test. To achieve some goals (e.g., establishing limits of validity; providing feedback on performance in difficult cases), the test sets would need to be extremely challenging; for other goals, the test sets would need to replicate routine casework.

Tests sets could be designed to examine a variety of important issues. For example, an important issue in latent print analysis is the ability of analysts to deal with distortions of latent prints that might arise from pressure, torsion, double-taps, and other mechanisms. Test sets of known-source latent prints that incorporate various types and degrees of distortion would, when introduced into the flow of casework, help laboratories see how well analysts recognize and deal with distortion. Performance testing with such specimens could help identify conditions under which a particular kind or degree of distortion can lead to inaccuracy. It could also provide invaluable feedback to analysts on their performance when dealing with the challenge of analyzing distorted latent prints, helping them to hone and improve their skills.

Development of suitable test sets for performance testing is a time-consuming and expensive task that will exceed the resources of many forensic laboratories. Laboratories may be able to cooperate in order to share this burden. Known-source latent print images, for example, could be prepared by one laboratory and shared electronically with other labs—creating efficiencies through inter-laboratory cooperation. Nevertheless, it is unrealistic to expect forensic laboratories themselves to bear the entire burden of creating such samples. Assistance from governmental agencies is needed.

latent print analysis and DNA analysis. This program is reportedly working well in Houston and is facilitated by the laboratory's adoption of the kind of context management system recommended by the Commisssion in the views document entitled "*Ensuring that Forensic Analysis is Based on Task-Relevant Information.*" A similar program has been adopted by the Netherlands Forensic Institute.

(3) Overcoming Aversion to Error

In order for the valuable potential of performance testing to be realized, some of the tests run under the program must be sufficiently challenging to induce errors. One cannot assess the strengths and limitations of a system without knowing the circumstances under which the system fails. To identify the boundary conditions within which forensic assessments are accurate, and beyond which accuracy suffers, it is necessary to test performance under marginal conditions—to push the boundaries until accuracy drops off. By analogy, if one wished to assess the strength of various types of chain, one must stress the chains until they break. Little of value would be learned from a study of chain strength in which the chains rarely if ever broke. A low rate of failure indicates an inadequate test. Frequent failure is the hallmark of a rigorous performance testing system.
Because it is desirable that performance testing induce errors, it is imperative to avoid the naïve mindset that associates error of any sort with incompetence. That association might be appropriate for proficiency tests designed to ascertain whether practitioners have the minimal level of proficiency needed to do their jobs. It is inappropriate for a test designed specifically to be highly challenging. On a highly challenging test, the occurrence of an error should be viewed as a valuable opportunity for feedback, learning, and improvement. It is not necessarily an indication of deficiency in the training, diligence or skills of the individual who makes the error.

It should also be recognized that the rate of error in challenging performance testing with marginal samples may have little relevance for predicting the rate of error in more routine forensic tests on samples that are easier to analyze. The rate of error when analysts are pushing the boundaries of their expertise tells us little or nothing about the probability of error when they are well within those boundaries. Performance is almost certain to be better in easy cases than in hard cases.

(4) Testing of Databases

In some forensic science disciplines, the processing of blind test samples may entail searching for matching samples in local, state or national databases. For example, a DNA analyst might search various government DNA databases looking for profiles that match the profile of a test sample. A latent print analyst might search databases of fingerprints looking for a match to a latent print that was submitted as a blind test sample. Allowing analysts to conduct such searches will provide valuable feedback on the overall operation of database systems. If a same-source reference sample is present in the database, these searches will test how often it is found and provide insight into circumstances under which same-source reference samples are missed, or are given a low ranking by search algorithms. These searches will also provide information on the risk of finding non-source reference samples that are similar enough to be misidentified as potential sources. At present, very little information is available about the sensitivity and specificity of identification of samples from databases *as those operations are performed in actual casework.* Feedback on this issue from blind performance testing will provide valuable insights that can be used to make these systems more effective and efficient.

Before database systems are tested in this manner, it may be necessary to change some of the rules that currently govern the purposes for which crime laboratories may gain access to governmental

databases.[19] These rules (as currently written) may not authorize forensic laboratories to access databases in the manner described here for the purpose of quality assurance and testing, or may create uncertainty about whether databases may be used in this way.

Crime laboratories will also need to take steps to assure that specimens created for purpose of performance testing are not permanently entered into databases in a manner that might create the false impression that these samples are associated with a crime scene. For example, laboratories might require that entry of samples into a database be authorized by section managers or quality assurance personnel who know which samples are from real crimes, and which are tests.[20]

3. Commission Recommendations

The National Commission on Forensic Science recommends that the following steps be taken to facilitate and promote performance testing in forensic laboratories:

(1) Funding of Pilot Programs

In order to facilitate development of performance testing programs, the Commission recommends that the Department of Justice and other funding agencies provide funding to laboratories willing to establish such programs. Pilot projects in which laboratories establish performance testing programs while monitoring how well these programs work, will be particularly valuable in charting a path forward on this issue. The practical experience of laboratories that pioneer the development of such programs should be recorded and disseminated for the benefit of the entire forensic science community. A period of trial and error will undoubtedly be necessary to learn how best to set up and run effective performance testing programs. Funding agencies should both support trial efforts and provide incentives to encourage laboratories to undertake these efforts.

(2) Creation of Test Sets

It is the view of the Commission that a government agency, such as NIST, should play a leading role in creating test sets for laboratory performance testing. This is a function that will be most efficient if handled in a centralized manner by an agency with expertise in testing. NIST has made valuable contributions to forensic DNA testing by providing mixed biological samples to laboratories for proficiency testing. In the view of the Commission, it would be it would be desirable for NIST to expand its efforts in this arena to include the creation of test sets for internal blind testing of forensic laboratories.

---

[19] These rules are found in a variety of sources, including the state and federal legislation that authorized the creation of the databases and in memoranda of understanding (MOU's) between agencies that operate the databases (e.g., the Federal Bureau of Investigation) and the laboratories and law enforcement agencies that are granted access to the databases.

[20] This step would not prevent analyst from searching samples against a database while remaining blind to the source of the samples. It is possible to search both latent fingerprints and DNA profiles against government databases without making the item a permanent part of the database. Latent prints, for example, must be registered before they become a permanent part of the FBI's fingerprint database, which was formerly known as the Integrated Automated Fingerprint Identification System, or IAFIS, and is currently known as the Next Generation Identification System. Laboratories can search latent prints against the database without "registering" them; and prints that are "registered" can later be "unregistered," which results in their removal for the system.

(3) Avoiding Misrepresentations about Error in the Courtroom

The Commission urges state and federal judges to consider carefully circumstances under which information about the rate of error in performance testing should be admissible in the courtroom. While the results of performance tests will be valuable and enlightening on a number of important issues, it would be misleading to equate the rate of error on test samples designed to be highly challenging with the rate of error for cases in general or with the probability of error in a specific case, particularly if the case involved relatively easy or straightforward analysis. Forensic scientists may hesitate to engage in the challenging kind of testing called for here if the results for highly challenging cases are unfairly used to impugn their performance on more routine cases. Consequently, if the results of performance testing are admitted as evidence in the courtroom, it should only be under narrow circumstances, and with careful explanation of the limitations of such data for establishing the probability of error in a given case.[21]

(4) Revision of Rules Regarding Database Access

It is the Commission's view that any rules regarding database access that preclude the kind of quality assurance testing program discussed here should be changed in order to allow such testing to proceed. Because the key rules at issue are contained in memoranda of understanding between the Federal Bureau of Investigation and the various agencies it serves, the Commission recommends that the Department of Justice work with the FBI to identify and revise any language or agreements that prohibit, or appear to prohibit, access to these databases for the kind of performance and quality assurance testing discussed herein.

[21] The Commission recognizes that the results of performance testing may fall within the government's disclosure obligations under *Brady v Maryland, 373 U.S. 83 (196*3). But the right of defendants to examine such evidence does not entail a right to present it in the courtroom in a misleading manner. The Commission is urging that courts give careful consideration to when and how the results of performance testing are admitted in evidence, not that courts deny defendants access to evidence that they have a constitutional right to review.