Blinding at Forensic Laboratories: A Meeting Report
April 23, 2019
By Robin Mejia, Maria Cuellar, and Jeff Salyards

Introduction

On November 1 and 2, 2018, CSAFE-CMU and the Allegheny County Office of the Medical Examiner hosted a workshop on blinding in forensic science. The event brought together researchers, laboratory directors, and laboratory quality managers to discuss issues involved in implementing blinding in forensic laboratories: what benefits blinding could provide and what challenges implementation could present. Attendees and presenters included researchers from Carnegie Mellon University and the University of Virginia, and laboratory directors and managers from labs in Houston, Texas and across the eastern and southeastern US, as well representatives from the Association of Forensic Quality Assurance Managers and a member of the NIST Forensic Science Standards Board.

Currently, blinding is not the norm in forensic laboratories in case work or in proficiency testing Research has long shown that blind proficiency tests can provide different results and information than open proficiency tests (see, for example, Lamotte et al, 1977). In 1992, the National Academy of Sciences issued a report opining that laboratories routinely engage in blind proficiency testing. [1] And in 2016 the National Commission on Forensic Science issued a Recommendation to the Attorney General to "require all DOJ FSSPs [forensic science service providers] to seek proficiency testing programs that provide sufficiently rigorous samples that are representative of the challenges of forensic casework."[2] However, blind proficiency testing has still not become standard procedure (Koehler, 2013) even though it has successfully been implemented and proven to be beneficial in other fields.

This meeting sought to understand the level of interest in implementing blinding, and particularly blind proficiency testing, in forensic laboratories among those in a position to implement it, and to identify the obstacles associated with implementing blinding that laboratory managers could expect to encounter. The meeting closed with an open session in which participants discussed these issues and began to synthesize a list of best practices for laboratories beginning (or considering) blind proficiency testing.

The authors are extremely grateful to all participants and the candor of the discussions that occurred at the meeting. In particular, the authors would like to thank Blythe Toma and Callan Hundl for their reviews and feedback on this document, which attempts to capture the range of experiences and opinions that arose. However, it's important to note that this is an emerging

[1] National Academy of Science Strengthening Forensics Science in the United States: A Path Forward
[2] National Commission on Forensic Science Recommendation to the Attorney General Proficiency Testing, 09/12/2016

area, and consensus was not reached on all issues. The views in this report are the authors' own.

Attendees

The event was held at the Allegheny County Office of the Medical Examiner in Pittsburgh, PA. Eight quality managers, four laboratory directors, a chief medical examiner, and a director of education came from seven forensic laboratory systems ranging in size from a single laboratory with fewer than 50 employees to a seven-laboratory system with over 200 employees. Two of the quality managers were representing AFQAM (the Association of Forensic Quality Assurance Managers.) Nine professors from three universities attended as well, representing fields ranging from statistics to psychology. Two graduate students and a post-doctoral researcher from statistics also attended.

Summary of Key Discussions

A) What is Blind Testing — Defining Terminology

Statistician Maria Cuellar, assistant professor of criminology at the University of Pennsylvania, gave an introductory talk called Blinding in Forensic Proficiency Testing and Casework. In it, she discussed three main concepts: blinding to task-irrelevant information, blind proficiency testing, and the goals of proficiency testing: testing the procedure and testing to find errors.

Blinding to task-irrelevant information is an approach that can be used to reduce contextual bias. Task-relevant and task-irrelevant information can be defined by using the National Commission on Forensic Science and ISO standards draft definition: "Information is task-relevant for analytic tasks if it is necessary for drawing conclusions: i) about the proposition in question, ii) from the physical evidence that has been designated for examination, iii) through the correct application of an accepted analytic method by a competent analyst."[3] It is task-irrelevant if it does not satisfy i, ii, or iii. To reduce contextual bias, examiners should be blinded to task-irrelevant information.[4]

Contextual bias is the error that occurs when well-intentioned experts are unconsciously vulnerable to making erroneous decisions due to task-irrelevant information. Contextual bias is an issue in all fields, not just forensics, which is why blinding is an essential part of experimental protocols in many fields of science. Experts cannot detect whether they are being influenced by information or choose to ignore certain pieces of information.

[3] National Commission on Forensic Science; Views of the Commission Ensuring that Forensic Analysis is based Upon Task-Relevant Information

Blind proficiency testing is an approach to reduce the bias due to knowing you are being tested. The problem with open testing, in which analysts know they are being tested, is that analysts might behave differently than they do in casework. An open test does not necessarily test the analysts' performance in casework. Furthermore, it can add additional bias beyond taking extra care with a test. Whereas in case work analysts make decisions based on the case scenario, in open proficiency test programs, analysts report factoring into their decisions historical knowledge of proficiency test providers' previous test offerings and the known limitations of proficiency test manufacturing.

The goals of proficiency testing are testing the procedure, which helps verify the laboratory's technical procedures are followed, and testing to find errors, which means determining whether the analyst provided the right answer. This initial presentation laid the groundwork for the workshop participants to speak the same language.

After the talk there was a discussion among the members of the workshop. A key point emerged about the difference between two types of errors: an error due to a mistake or misconduct by an analyst, and an error due to the limitations of method being employed. Sometimes when statisticians talk about a process having an error, forensic practitioners think they are being accused of carelessness or malicious intent. Thus, it is important to specify which type of error is being discussed. Some suggested calling the first type of error a "mistake" and the other type just "error" but there was no consensus about this. A related point that arose was the need to distinguish between proficiency tests, which are designed to assess the performance of an examiner or a laboratory process, and error rate studies, which are also needed in many forensic disciplines. Error rate studies were not the focus of the meeting, but participants wanted to emphasize that both proficiency testing and error rates studies are needed.

B) Implementing Blind Testing – The Houston Forensic Science Center Experience (HFSC)

The Houston Forensic Science Center (HFSC) has one of the most well-known blind proficiency testing programs in a forensic laboratory in the United States. HFSC CEO and president Dr. Peter Stout described Houston's blind proficiency test program, highlighting challenges they had overcome, as well as the unique features of the Houston Forensic Science Center that enable blind testing at a large scale. Key points raised include:

1) Stout described his experience and understanding of labs that do workplace drug testing, noting that many have implemented blind proficiency tests at scale for some time. Stout focused on the military, noting that the Navy has 2 labs and the Army has 1 lab, which between them process 8-10 million samples per year. They conduct approximately 3000 double-blind tests per year. Blind tests are sent to the commands, which package them to look like cases and send them to the

laboratories. He noted that Health and Human Services workplace drug testing programs and Department of Transportation drug testing programs also required participating laboratories to do blind proficiency testing. Those labs also have four sets of open proficiency tests per year. He noted that in his experience, those labs could generally turn around a negative drug test in a day and a positive drug test in a day and a half, but when they did the open proficiency test, they would "uniformly get them back on day 9.5", suggesting that analysts were taking extra care with the tests when compared to their behavior with regular casework.

2) This experience was part of what led Stout to believe that HFSC should augment its open proficiency testing program with blind proficiency tests. Open proficiency tests do provide useful information and meet accreditation requirements, but blind proficiency tests provide a better assessment of how analysts perform casework, and, importantly, enable the laboratory to assess the entire pipeline of evidence handling from submission to final report. He also has a large lab with a team of quality managers that report directly to him. Currently HFSC has a goal that an average of percent of completed casework each month be proficiency tests. At the time of the meeting, the submission schedule was:
   - Toxicology: 14/month
   - Seized Drugs: 30/month
   - Biology: 4/month
   - Firearms BQC: 1/month
   - Firearms blind verification: 1/month
   - Latent print processing: 3/month
   - Latent print comparison: 10/month
   - Digital forensics: 1/month
   - Forensic multimedia (A/V): 1/month

3) Stout then discussed several issues that arise when implementing blind proficiency testing, addressing both issues that are general to the field and the specifics of the HFSC experience.
   a. There are cultural challenges to changing how things are doing in a law enforcement setting with established norms and in a field with a history that has included claiming error rates of zero in many disciplines. These are discussed more below, but Stout noted that he is in a privileged position in part because Houston had one of the more epic set of failures in its crime lab in the early 2000s, which resulted in wrongful convictions being discovered, a large legal settlement the county had to pay, and the temporary shuttering of facilities. These problems led to wholesale reforms that included removing the laboratory from a law enforcement chain of command and hiring a scientist (Dr. Stout) to lead the new lab. So, he has unusual autonomy and has been able to establish a new set of norms in Houston.

b. The primary logistical challenge is creating samples that mimic cases, as many forensic disciplines require types of evidence that are more challenging to create from scratch than a vial of blood or urine, such as fired bullets or used crowbars. Additionally, evidence submissions must come in as part of a realistic sounding case with realistic submission materials. He challenges examiners to try to find blind tests by offering $20 Starbucks gift card if they correctly identify a test sample and asking them to pay $1 if they think a sample is a test, but it is actually real casework. Analysts have a great deal of accumulated knowledge about what kinds of evidence come from which kinds of crimes and neighborhoods and can spot unusual samples from context. Stout found that certain characteristics alert analysts of the fact that it is a test sample, such as that the packaging is too neat for the form has nice handwriting (so it probably was not submitted by a police officer). Only once has an examiner reported a test incorrectly: the evidence came with an individual's name that sounded like a phrase. The examiner thought it was a joke, but it turned out the name was real, so the examiner paid Stout $1. Additional logistical hurdles are often discipline-dependent, such ensuring blind tests aren't uploaded into national databases such as the Combined DNA Index System (CODIS).
c. The structure of Laboratory Information Management Systems (LIMS) create challenges to implanting blind proficiency testing. All laboratories require an information management system to track evidence and monitor case work. Small laboratories may use homegrown programs or spreadsheets, but most large laboratories use a commercial LIMS from one of a handful of primary vendors: JusticeTrax, Porter Lee's Beast, and Starlims. Stout said that HFSC is switching to JusticeTrax in part because it is the only product that offered functionality for an internal flag that a case was a blind proficiency test that would be hidden to examiners. However, he noted that the migration is facing many obstacles, and much of HFSC's data was still in Porter Lee. In addition, LIMS were not designed with blinding in mind, and managers say that it can be difficult or impossible to blind analysts to parts of a case that would be necessary if they wanted to implement blinding to task-irrelevant information. Several attendees told horror stories of trying to migrate LIMS.

4) Discussion

There was a wide-ranging discussion after Stout's talk, in which managers from laboratories that are beginning to implement blind testing, or planning to do so, discussed their experiences. One conclusion that emerged was that each laboratory operates in a unique context that defines the challenges managers will face in implementing blind testing. However, the group also identified a number of key themes:

a. The cultural history of the field of forensics is an obstacle to the implementation of blind testing. As has been documented in a number of reports (Cole 2014, 2009, 2005), in many disciplines, examiners have historically claimed the ability to generate perfect identifications or error rates of zero. No scientific discipline actually achieves error rates of zero, and errors have been found in many disciplines in forensics. However, asking examiners and managers to embrace this fact—that errors happen—as a baseline truth and focus on identifying the problems that causes errors is harder than it sounds from the outside.
b. The cultural challenges are exacerbated by the fact that forensic analyses occur in an adversarial context in which test results can become part of legal proceedings, and examiners can have their reputation challenged in court based on test results, as proficiency tests can be requested through the discovery process and used to challenge an examiner's qualifications.
c. Buy-in from senior management is essential to the successful implementation of blind testing. When laboratory staff report into a law enforcement chain of command, obtaining such buy in can require significant education and effort, as training in scientific methods and norms is not part of the background of those positions.
d. In the near term, any implementation of blind proficiency testing programs will be in addition to existing open proficiency testing programs. This means that it's an additional cost that must be justified and covered.
e. Small laboratories will find it more challenging than larger laboratories to implement blind testing as they lack the staff and financial resources of a laboratory such as HFSC. There was discussion of how interested laboratories could partner with each other to share resources.

C) Lab-University Partnerships
   a. Researchers from CSAFE institutions, including Karen Kafadar and Sharon Kelley from the University of Virginia, and Jeff Salyards from Compass Scientific Consulting discussed academic partnerships with forensic laboratories, including work on assessing case management practices and understanding dispute resolution in subjective disciples that require secondary confirmation of findings, developing quality metrics that can be used to objectively assess image quality in disciples such as fingerprint analyses, testing new methods in disciplines including drug toxicology, and assisting with implementing blind testing. Key take-aways from the discussion included that:
      i. Academics can provide assistance in design of new procedures and especially assessing processes and changes.
      ii. Practitioners are invested in sharing results and best practices with colleagues and yet have limited to no dedicated time for research and

publication. Collaborations can help bridge that gap as academics are evaluated on their ability to publish and disseminate findings.

iii. Such partnerships could help speed the adoption of blind proficiency testing as a practice. Participants suggested that an academic institution or research consortium could aid laboratories in analyzing data from blind proficiency tests. They noted that researchers could help with analyses of specific lab tests and also argued that the field would benefit from the collection of results in a central location, enabling inter-lab studies and analysis of aggregated results.

Next Steps

On the final morning of the workshop, Robin Mejia, from the Department of Statistics and Data Science at CMU, and Jeff Salyards, Principle Analyst at Compass Scientific Consulting, led a group discussion and review issues that were raised in the sessions the previous day. Attendees brainstormed steps needed to facilitate the implementation of blind proficiency tests at more laboratories.

1) Define what blind testing and develop and explanation of benefits:
    a. Blind proficiency tests are tests in which the examiner does not know that she or he is being tested.
    b. Blind proficiency tests provide a better ability to assess how examiners work on cases. Important to note is that this is helpful data but not a substitute for error rate studies.
    c. Blind proficiency tests enable ability to assess the entire pipeline of case processing.

2) Define the goals of testing:
    a. Should tests mimic casework or stress test systems?
    b. Consensus was that initial tests should mimic casework, as what laboratory managers want to know is how their analysts perform on the kinds of cases they usually work.
    c. There was robust discussion about the concept of stress testing systems or "testing to failure" – providing very challenging blind tests to see where systems break down when stressed. Everyone agreed that this is useful information, and also that that is not the same thing as testing examiner proficiency in the kinds of cases they normally see. There was concern about how results from stress tests could be used by defense attorneys, and discussion about the history and culture of laboratories, where failure is viewed as unacceptable and a possibly career-ending event for an analyst, not as a learning opportunity. In particular, it

was viewed as important to address whether testing to failure programs would be subject to disclosure requirements as Brady (exculpatory) evidence.

3) Logistical challenges. The group revisited the logistical issues discussed in the previous day:

    a. Laboratory culture, reporting structures and the need to obtain buy-in from management. The need to obtain buy-in from senior management was considered an essential requirement to the successful implementation of blind proficiency testing, and attendees noted that in many contexts, obtaining support for senior management could require an education process, as blind testing requires additional resources and a change from a status quo that already meets legal requirements. Increasing external calls for blind testing or testing similar to casework coming from organizations and panel related to forensic science can help quality managers make a case, as could changing norms in the field—if blind testing becomes something that "all the good labs are doing." Other strategies discussed included raising the desire demonstrate quality work and concerns over the high cost of a mistake that such a testing program could help avoid. In addition, the issue of obtaining intra-agency buy in was discussed, as police partners are needed as submitting agencies, and legal communities and courts need to understand what the programs are and why the matter. HFSC staff noted that they publish test results on their public-facing website, but many attendees felt this level of openness might not be possible in other jurisdictions.
    b. Staff and money challenges for labs that do not have the economy of scale that HFSC has. Participants discussed developing a consortium, noting it would be helpful for an organization like CSAFE to serve as an organizer and clearing house for ideas and actual test cases. Near term goals included working with HFSC to buy additional toxicology samples when they put their orders in in order to obtain the same discounts. Longer term goals included sharing or collaboration on the development of test cases in disciplines that require complex fake cases and physical evidence like crow bars.
    c. Developing realistic cases remained a key concern, as well as managing the logistics of a program where most employees are blinded to the fact that a case is a test. Issues include:
        i. Acquiring and preparing materials. A key challenge to implementing blind proficiency tests is that crime scene evidence can be complex to mimic. While for some disciplines such as toxicology, materials can be purchased from vendors such as RTI, for others, such as arson or ballistics, evidence will need to be generated by the laboratory staff from locally purchased materials. In these cases, it is essential that someone involved in the evidence preparation understand the characteristics of evidence

retrieved from a crime scene. In an early proficiency test, one quality manager said the analyst looked at a crow bar submitted for latent print processing and said, no one holds a crow bar that way. As noted above, ensuring that packaging and submission materials look authentic is also crucial. Notably, authentic does not mean perfect. Analysts have spotted proficiency test because they say the handwriting is too neat or "too female." If spelling mistakes are common in real submissions, they should occur on proficiency test submissions, too. Additionally, making submission match what is expected from neighborhoods or even know law enforcement officers is important. If it's known and a senior officer only does large busts, he or she should not submit a proficiency test with a small drug sample.

ii. Determining who needs to know and be involved. An effective blind proficiency testing program requires balancing the need to involve enough people to make the process work, while minimizing the number of people who know which cases are tests. The exact balance will depend on the size of the lab but a successful blind proficiency test will generally involve knowledge and involvement of: laboratory quality manager(s), staff involved in uploads to national databases such as CODIS, individuals at a submitting police agency, both to do submissions and be aware of which cases were a test if a report is released, possibly someone at the District Attorney's office for similar reasons.

iii. Managing the LIMS system. The group discussed the possibility of drafting a document laying out exactly the options that laboratory managers want their LIMS to provide. For the purposes of this report, key features would involve the ability to have a flag that a case is a proficiency test only the quality management team could see but that would prevent reports from being issued to external agencies.

iv. Making sure that results are not released to police or District Attorney's offices as real cases.

v. Ensuring results not uploaded to systems like AFIS and CODIS. Staff involved in uploads will need to know which cases not to upload.

vi. Determining metrics to review. Blind proficiency tests allow quality managers to determine if the right answer was obtained, and also if the correct process was followed. Also, the question was raised of determining whether and how proficiency tests would affect overall metrics for the agency such as turn around time and case submission number. The ability to flag cases within the LIMS so that appropriate metrics could be created is an issue.

4) Drafting a paper reviewing these issues and presenting detailed case studies in successful implementation proficiency tests in forensic laboratories, starting with HFSC.

A writing group, including the authors of this memo, was formed at the meeting. We include our initial bibliography below.

Bibliography

Forensic Science References

Bozon, M. V., et al. "Error rate for HLA-B antigen assignment by serology: implications for proficiency testing and utilization of DNA-based typing methods." Tissue Antigens 50.4 (1997): 387-394.

Budowle, Bruce, et al. "A perspective on errors, bias, and interpretation in the forensic sciences and direction for continuing advancement." Journal of Forensic Sciences 54.4 (2009): 798-809.

Cole S.A. More Than Zero: Accounting for Error in Latent Fingerprint Identification. Journal of Criminal Law & Criminology, Vol. 95 (2005)., No. 3, pp. 985-1078.

Cole S.A. Forensics without Uniqueness, Conclusions Without Individualization: The New Epistemology of Forensic Identification. Law, Probability and Risk, Vol. 8 (2009), No. 3, pp. 233-255.

Cole S.A. Individualization is Dead, Long Live Individualization! Reforms of Reporting Practices for Fingerprint Analysis in the United States. Law, Probability and Risk, Vol. 13 (2014), No. 2, pp. 1-34.

Cowan, Everard James, and Roger Koppl. "An experimental study of blind proficiency tests in forensic science." The Review of Austrian Economics 24.3 (2011): 251-271.

DNA Advisory Board. "Quality assurance standards for forensic DNA testing laboratories." Forensic Science Communications 2.3 (2000).

Found, Bryan, and John Ganas. "The management of domain irrelevant context information in forensic handwriting examination casework." Science & Justice 53.2 (2013): 154-158.

Hagerman, Paul J. "DNA typing in the forensic arena." American journal of human genetics 47.5 (1990): 876.

ISO, EN. "IEC 17043: 2010 Conformity assessment–General requirements for proficiency testing." Bedömning av överensstämmelse—Allmänna krav för kompetensprövning (av laboratorium) (2010).

Jonakait, Randolph N. "Forensic science: The need for regulation." Harv. JL & Tech. 4 (1991): 109.

Kaye, David H. "The admissibility of DNA testing." Cardozo L. Rev. 13 (1991): 353.

Koehler, Jonathan J. "Proficiency tests to estimate error rates in the forensic sciences." Law, Probability and Risk 12.1 (2013): 89-98.

Koehler, Jonathan J. "Fingerprint error rates and proficiency tests: What they are and why they matter." Hastings LJ 59 (2007): 1077.

Koppl, Roger. "How to improve forensic science." European Journal of Law and Economics 20.3 (2005): 255-286.

Mnookin, Jennifer L. "The validity of latent fingerprint identification: Confessions of a fingerprinting moderate." Law, Prob. & Risk 7 (2008): 127.

Peterson, Joseph L., et al. "The feasibility of external blind DNA proficiency testing. I. Background and findings." Journal of Forensic Science 48.1 (2003): 1-11.

Peterson, Joseph L., et al. "Developing criteria for model external DNA proficiency testing: final report." Chicago, IL: University of Illinois at Chicago (2001).

Peterson, Joseph L., and Penelope N. Markham. "Crime laboratory proficiency testing results, 1978–1991, I: identification and classification of physical evidence." Journal of Forensic Science 40.6 (1995): 994-1008.

Peterson, Joseph L., and Penelope N. Markham. "Crime laboratory proficiency testing results, 1978–1991, II: Resolving questions of common origin." Journal of Forensic Science 40.6 (1995): 1009-1029.

Reeder, Dennis J. "Impact of DNA typing on standards and practice in the forensic community." Archives of pathology & laboratory medicine 123.11 (1999): 1063-1065.

Saks, Michael J., and Jonathan J. Koehler. "The coming paradigm shift in forensic identification science." Science 309.5736 (2005): 892-895.

Stockham, Rex A., Dennis L. Slavin, and William Kift. "Specialized use of human scent in criminal investigations." Forensic Science Communications 6.3 (2004).

Thompson, William C., Franco Taroni, and Colin GG Aitken. "How the probability of a false positive affects the value of DNA evidence." Journal of Forensic Science 48.1 (2003): 1-8.

Whitman, Glen, and Roger Koppl. "Rational bias in forensic science." Law, Probability & Risk 9.1 (2010): 69-90.

Zabell, Sandy L. "Fingerprint evidence." JL & Pol'y 13 (2005): 143.

Medical & Analytical Chemistry References

Analytical Methods Committee. "Proficiency testing of analytical laboratories: organization and statistical assessment." Analyst 117.1 (1992): 97-104.

Davis, Kenneth H., Richard L. Hawks, and Robert V. Blanke. "Assessment of laboratory quality in urine drug testing: A proficiency testing pilot study." JAMA 260.12 (1988): 1749-1754.

Hamlin, William. "Proficiency testing as a regulatory device: a CAP perspective." Clinical chemistry 38.7 (1992): 1234-1236.

Hansen, Hugh J., Samuel P. Caudill, and D. Joe Boone. "Crisis in drug testing: Results of CDC blind study." JAMA 253.16 (1985): 2382-2387.

Hofgärtner, Wolfgang T., and Jonathan F. Tait. "Frequency of problems during clinical molecular-genetic testing." American Journal of Clinical Pathology 112.1 (1999): 14-21.

Hofherr, L. K., et al. "Methods for a model blind proficiency testing system." Clinical laboratory science: journal of the American Society for Medical Technology 5.3 (1992): 160-164.

Jain, Naresh C., Thomas C. Sneath, and Robert D. Budd. "Blind proficiency testing in urine drug screening: The need for an effective quality control program." Journal of Analytical Toxicology 1.3 (1977): 142-146.

LaMotte Jr, Louis C., et al. "Comparison of laboratory performance with blind and mail-distributed proficiency testing samples." Public Health Reports 92.6 (1977): 554.

Mason, Morton F. "Some realities and results of proficiency testing of laboratories performing toxicological analyses." Journal of analytical toxicology 5.5 (1981): 201-208.

Parsons, Patrick J., et al. "Evaluation of blood lead proficiency testing: comparison of open and blind paradigms." Clinical chemistry 47.2 (2001): 322-330.

Shahangian, Shahram. "Proficiency testing in laboratory medicine: uses and limitations." Archives of pathology & laboratory medicine 122.1 (1998): 15.

Stull, Tina M., et al. "Variation in proficiency testing performance by testing site." JAMA 279.6 (1998): 463-467.

Walker, Ron, Mike Thompson, and Richard Lawn. Proficiency testing in analytical chemistry. Royal Society of Chemistry, 2010.