Draft Guidance on Testing the Performance of Forensic Examiners

OSAC Human Factors Committee

May 8, 2018

I. Overview

This document provides guidance on designing, conducting and reporting empirical studies of the performance of forensic examiners on routine analytical tasks, such as comparing items to determine whether they have a common source, or classifying items by category (e.g., determining the caliber of a bullet or the size of shoe that made a shoeprint). Studies of examiner performance may be undertaken for a variety of purposes, including assessment of the validity of methods, identification of strengths and weaknesses of individual examiners, and assessment of the strengths and weaknesses of laboratory systems. There is an extensive scientific literature on assessing human performance. This document distills key points from that literature that forensic scientists may need to know.[1]

II. Scope of Application

The guidelines in this document are non-mandatory. This document does not require any individual or organization to study the performance of forensic examiners, nor does it require that such studies be conducted in any particular manner. The advice provided here does, however, reflect the considered judgment of the OSAC Human Factors Committee regarding the strengths and weaknesses of various research methods, based on principles and insights that are widely accepted in the scientific community. It is designed to help forensic scientists better understand the trade-offs entailed in choosing among various possible ways to design, conduct and report research on examiner performance.

This document applies both to analytic tasks performed in full by human examiners (e.g., fingerprint, bite mark, and tool mark comparisons) and to tasks performed in part by machine but interpreted by human examiners (e.g., interpretation of DNA electropherograms; interpretation of audio spectrographs). It is not intended to apply to readings provided entirely by machine and simply reported by humans. Our focus here is on methods that rely, at least in part, on human judgment.

The performance testing methods discussed here use known-source or known-type test specimens to assess examiners' accuracy when reaching conclusions about an item's source or

[1] For a detailed and insightful discussion of the same issues, readers should also consult Martire and Kemp (2016).

type. This document does not address the testing of examiner performance on other tasks. Among the tasks that are not addressed here are:

- quantitation
- tasks that do not entail reaching a reportable conclusion on source or type (e.g., sample collection; sample preparation; instrument set-up and calibration)
- tasks that involve recognition of relevant evidence rather than drawing conclusions about source or type of specific items (e.g., identification of relevant evidence at a crime scene)
- tasks that involve generation or evaluation of activity-level or crime-level hypotheses or theories (e.g., crime scene reconstruction; assessment of intent or motive)
- tasks that involve causal analysis (e.g., cause of death; cause of a fire).

It may well be important to test examiner performance on such tasks but that is not the focus of this document.

III. Definitions of Key Terms

Validation—is the process of evaluating a system or component, through objective evidence, to determine that requirements for an intended use or application have been fulfilled (*see* OSAC Lexicon). To validate methods for source determination, or for classifying physical items by type, it generally is necessary to conduct empirical studies of the method's accuracy under various circumstances, and to determine associated error rates. Empirical testing is particularly important when validating methods that rely, in part, on human judgment. [2]

Valid/Validity—As used in this document, validity is a quality or property of a forensic science method that is used for source determination or for classifying items by type. A method is valid (has validity) to the extent it produces accurate results—i.e., correct source determinations or correct classifications. The validity of such a method can be assessed by testing whether it produces accurate results when applied to test specimens of known source or type.

[2] As explained in the 2016 report of the President's Council of Advisors on Science and Technology (hereafter, PCAST, 2016):

> *Scientific validity and reliability require that a method has been subjected to empirical testing, under conditions appropriate to its intended use, that provides valid estimates of how often the method reaches an incorrect conclusion…. Without appropriate estimates of accuracy, an examiner's statement that two samples are similar—or even indistinguishable—is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact. Nothing—not training, personal experience nor professional practices—can substitute for adequate empirical demonstration of accuracy. (PCAST, 2016, p. 46)*

Reliability—As used in this document, and in most scientific disciplines, reliability refers to the consistency of results as demonstrated by reproducibility or repeatability (OSAC Lexicon). Reliability can be a property either of a method, instrument, or examiner. There are many dimensions of reliability. Test-retest reliability is a property of a method that produces the same results (consistency) when used repeatedly to test the same items. Intra-examiner reliability is a property of an examiner who produces the same results (consistency) when repeatedly asked to examine or compare the same items. Inter-examiner reliability is a property of a group of examiners who reach the same conclusion (consistency) when asked to examine or compare the same items. The reliability of a measurement instrument (i.e., its consistency over repeated measurements on the same items) is sometimes referred to as its precision.

Black-Box Study—A black box study assesses the accuracy of examiners' conclusions without considering how the conclusions were reached. The examiner is treated as a "black-box" and the researcher measures how the output of the "black-box" (examiner's conclusion) varies depending on the input (the test specimens presented for analysis). To test examiner accuracy, the "ground truth" regarding the type or source of the test specimens must be known with certainty.

White-Box Study—A white-box study is similar to a black-box study but also allows assessment of the thought process or methodology of the examiner. It allows the researcher to look inside the "black-box" and gain insight into how examiners reach conclusions.

Error—In this document we use the term *error* to refer to inaccurate conclusions arising from limitations of a test method when properly applied. Examples of error include the failure to distinguish items that differ with regard to source or type, but are indistinguishable at the level of analysis applied in the examination (e.g., failure of a DNA test to distinguish biological samples of two different individuals who happen to have the same genotypes at the loci examined by the DNA test).

Mistake—An instance in which an examiner does something wrong when applying a test method, such as failure to follow proper procedure, failure to record results correctly, failure to notice relevant information, or failure to interpret information in the proper manner. Mistakes can cause examiners to reach a wrong conclusion. Note that in this document we distinguish inaccuracy arising due to mistake from inaccuracy arising due to inherent limitations of a test method, which we call "error."

Context Management Procedure—A procedure designed to limit or control what a forensic examiner knows about the background or circumstances of a criminal investigation in order to reduce the potential for contextual bias. These procedures are designed to assure that the examiner has access to "task-relevant" information needed to perform the examination in an appropriate manner, while limiting or delaying exposure to information that is unnecessary or that might be biasing if presented prematurely.

Test Specimen—An item of known source or type that is submitted for forensic examination in order to test whether the examiner draws the correct conclusion regarding the item's source or type.

IV. Distinguishing Reliability from Validity

When designing studies of examiner performance, it is crucial to distinguish reliability from validity. People often confuse or conflate the concepts and mistakenly assume they are studying validity when actually studying only reliability.

Why isn't reliability a good indicator of validity, or stated equivalently, why isn't reliability a good indicator of accuracy? Reliability assessment does not require knowledge of ground truth. Reliability assessment measures only agreement among and within examiners or instruments, not whether the examiners or instruments agree with ground truth. While reliability assessment is generally easier to carry out, the absence of known ground truth means that a reliability assessment by itself is insufficient to test validity, as two examiners may agree yet both be wrong. Hence, asking examiners to replicate each other's work (a common approach in assessing inter-examiner reliability) is not a test of validity.

Reliability is nevertheless important. Studying the reliability of examiners' judgments on casework samples can provide laboratory managers with valuable information about laboratory and examiner performance. While the exact causes of disagreements may be difficult to determine, unreliable examiner judgments should alert managers to areas where improvements in performance are possible. If instruments are found unreliable, it may signal a need for repair or replacement or for use under more tightly controlled conditions (e.g. temperature/humidity). If examiners are making different judgments about the same physical specimens it may signal training deficiencies or, if all examiners are adequately trained, it may indicate that the method being applied needs greater refinement or specification, that there are limitations of the underlying method that are not fully understood, or possibly even that the method is invalid. Consequently, for quality assurance purposes, when a comparison or examination of casework samples is conducted independently by more than one examiner laboratories should routinely collect and retain data on how often the results agree and, if they disagree, on the extent of divergence. The same should be done if a single examiner has conducted the same examination or comparison of the same samples on more than one occasion.

Possible causes of less than perfect reliability. If managers find less than perfect reliability in the conclusions drawn by examiners who evaluate items known to be the same in relevant respects, they should try to determine the reasons for the inconsistency. While it may be difficult or impossible to assess which examiners are right and which are wrong in a particular case when ground truth is not known, the existence of the disagreement should alert

managers to a possible problem, which could trigger a review, carefully conducted retests, or additional validity testing with known-source test specimens in order to help pinpoint the reasons for the inconsistency. Possible reasons that should be considered include the following:

- ***Training deficiencies/mistakes:*** Low reliability of examiners' evaluations can sometimes signal a failure of one or more examiners to execute the forensic procedure correctly and therefore indicate a need additional (or better) training.

- ***Inconsistent decision thresholds:*** Examiners might also disagree because they have different thresholds for making decisions. One examiner might require stronger or clearer evidence to reach a particular conclusion than another. Empirical studies with known-source test specimens can be extremely useful for determining whether examiners are applying inconsistent decision thresholds and, more importantly, for assessing which decision threshold is better for maximizing the accuracy of the procedure. For example, such studies might show that some examiners being unduly conservative, judging samples unsuitable for comparison or judging comparisons inconclusive, while other examiners are reaching conclusions that are accurate. Alternatively, it might show that some examiners are drawing inaccurate conclusions, and thus that a more conservative approach is warranted. Table 1, to be described in a later section, will illustrate the kind of data needed to make these evaluations. These evaluations can be valuable for detecting mistakes, refining training procedures and for helping examiners improve their skills.

- ***Limited validity.*** If examiners are well-trained and are following the same method, and yet inter-examiner reliability is low, it may also call into question the validity of the method itself. That is, the inconsistencies may arise from error that is inherent in the method, rather than from mistakes in applying the method. From a statistical perspective, the reliability of conclusions across multiple examiners limits how accurate the examiners collectively can be. If half of the examiners conclude that two items are from the same source, and the other half conclude that the same items are from different sources, then only half of the examiners can be correct, which means that the examiners collectively reach the correct conclusion only half of the time.

- ***Limited validity as applied.*** It is important to keep in mind that a method may be highly accurate for some applications but less accurate for others. For example, a method might be highly accurate when used to examine high-quality specimens but less accurate when dealing with low-quality or marginal specimens. Inconsistency across examiners when evaluating a particular specimen, or making a particular comparison, may indicate that the method is less accurate when applied to items of that type, or may identify circumstances in which special caution is needed to avoid errors.

Reliability of judgments of sample adequacy. Laboratories should also monitor the reliability of examiners' judgments when assessing the suitability of forensic samples for testing

or comparison. If there is wide variability in these assessments it raises such questions as: (a) whether some examiners are exercising too little or too much caution in determining that items are suitable for analysis, and (b) whether mistakes in assessing the suitability of items for analysis are affecting examiners' accuracy. Lack of reliability may arise from correctable deficiencies in training, but it could also signal the need for additional research on how best to distinguish items that are suitable and unsuitable for analysis and comparison with existing methods.

V. Guidelines for Validity Testing

The following guidelines are designed to help forensic scientists design and carry out proper empirical studies of the validity of methods for making source determinations and for classifying items by type.

Guideline #1: Validation studies should involve test specimens of known-source or known-type.

Forensic scientists generally cannot know with certainty the "ground truth" regarding the nature or origin of the questioned items recovered in criminal investigations. Hence, those items generally cannot be used as specimens for validation studies. To obtain specimens of known-source or known-type, researchers generally must either create the samples themselves (e.g., obtain latent print impressions from known people; create footwear impressions with known shoes) or obtain specimens from a reliable vendor (e.g., samples of a particular chemical, compound or drug). As a general rule, the specimens used in a validation study must be specifically created, developed, or obtained for that purpose.

Guideline #2: Participants' judgments about the source or classification of test specimens should be independent of one another.

The PCAST report reached the following conclusion about how to test the validity of forensic science methods that rely, in part, on examiners' subjective judgment:

> … the foundational validity of subjective methods can be established *only* through empirical studies of examiner's performance to determine whether they can provide accurate answers; such studies are referred to as "black-box" studies…In black-box studies, many examiners are presented with many independent comparison problems—typically, involving "questioned" test specimens and one or more "known" test specimens—and asked to declare whether the questioned test specimens came from the same source as one of the known test specimens. The researchers then determine how often examiners reach erroneous conclusions (PCAST, 2016, p. 49)

The PCAST report was critical of validation studies that employ "set-based analysis" in which examiners are asked to perform all pair-wise comparisons within or between small sets of test specimens (PCAST, 2016, p. 106-107). In set-based studies, complex dependencies arise among the answers, which interferes with the ability of researchers to make appropriate estimates of examiner accuracy. For example, when an examiner is asked to determine which of five questioned test specimens has the same source as a reference sample, the examiner's determination that some of the questioned test specimens have a different source has implications for (and may ultimately dictate) which test specimen is determined to have the same source, making it impossible for researchers to distinguish examiners' accuracy for making identifications (same-source determinations) from the examiners' accuracy for making exclusions (different-source determinations).

We agree with the PCAST report that validation studies should involve independent comparisons of known and questioned items, where the examiners' conclusions about any one comparison provide no information about the correct result of other comparisons.

Guideline #3: When reporting the accuracy of examiners' performance in validation studies, it is necessary to distinguish various types of errors.

To understand human performance at a source determination task, it is necessary to distinguish accuracy with same-source test specimens from accuracy with different-source test specimens.

- When an examiner says two items have the same source, but they in fact have different sources, the examiner has falsely reached a positive conclusion. This is referred to as a "false positive," or "false inclusion."

- When an examiner says, two items have different sources, but they in fact share a common source, the examiner has falsely reached a negative conclusion. This is referred to as a "false negative," or "false exclusion."

- Both kinds of errors (false positives and false negatives) need to be considered in assessing the accuracy of the test.

- Rates of other findings, such as "inconclusive" or the determination that a sample is not suitable for testing are also important. These rates must also be reported.

The rates of accuracy revealed by a validation study generally should not be reduced to a single number. It is misleading to say something like "the study showed that examiners were 90% accurate" because accuracy is likely to vary for same-source and different source comparisons, and because the overall rate of accuracy of a method will depend on the base-rate of same-source and different-source comparisons examiners are asked to make, as well as their rates of false negatives and false positives. Consequently, researchers (and forensic scientists generally)

should always separate same-source and different-source comparisons when providing data about the accuracy of examiner performance.

A further note on terminology: sensitivity and specificity. When assessing the accuracy of human performance on a task that involves classification or source determination, researchers sometimes report the "sensitivity" and "specificity" of the method or procedure used to perform the task.

Consider the following simplified chart:

		Actual Status	
		Same Source	**Different Source**
Examiner's Decision	**Same Source**	**A**	**B**
	Different Source	**C**	**D**

"Sensitivity" refers to the extent to which an examiner deems an item to be from the same source as another item when the two are actually from the same source. Thus sensitivity is equal to A/(A+C). It is sometimes also called the "hit rate" or the "true positive rate" of the method or procedure.

"Specificity" refers to the extent to which an examiner deems an item to be from a different source when the two are actually from a different source. Thus specificity is equal to D/(B+D). It is sometimes called the "true negative rate" or the "correct rejection rate." The specificity of a test is directly related to the false positive rate of the test, which is B/(B+D). As the specificity of the method increases, the false positive rate will necessarily decrease. For example, if the examiner, when comparing items from different sources, correctly decides they are different 95% of the time, then the rate at which she incorrectly decides they are the same (a false positive) cannot exceed 5%. If the specificity increased to 99%, then the false positive rate could not exceed 1%.

Thus there are two dimensions to validity: the accuracy of examiners in judging same-source comparisons (sensitivity), and the accuracy of examiners in judging different-source comparisons (specificity).

- For example, if examiners are given 20 trials for which the correct answer is "same source," and they concluded "same source" 17 times, their sensitivity is 85%.

- If examiners were given 40 trials for which the correct answer is "different source," and they said "different source" 36 times, their specificity is 90%.

Data about both sensitivity and specificity are needed to understand examiners' performance.

Guideline #4: Rates at which examiners reach conclusions of "no value" or "inconclusive" must also be reported.

The number and rate of "no value" and "inconclusive" determinations are vital pieces of data that researchers must record and report when studying the accuracy of forensic examiners. These data help place information about the performance of examiners in a particular study in the proper context.

The rates at which examiners reach "no value" and "inconclusive" determinations are important metrics for laboratory managers to monitor even when not testing examiner accuracy. If two examiners with similar case loads differ markedly in the rate at which they are able to reach conclusive determinations, it may signal that one or both are processing cases in a less than optimal manner. One of the examiners may be too cautious about drawing conclusions, failing to draw conclusions that could be reached with accuracy. Or, one of the examiners may be overconfident, reaching inaccurate conclusions based on inadequate evidence. In either case, the situation may warrant managerial attention to assure that both examiners are adequately trained and are handling casework in an optimal manner. It may also signal a need for further research to determine the best way to handle those cases on which examiners may disagree.

Knowing the rates of "no value" and "inconclusive" determinations in casework can also be helpful in assessing the results of validation studies. If the rates in the study are very different than the rates in casework, it may indicate one or more of the following: (1) that the examiners in the study were being more (or less) cautious about drawing conclusions than they would be when doing casework; (2) that the comparisons presented to examiners in the study were more (or less) difficult than casework comparisons; (3) that other differences between the circumstances of the study and the circumstances in which casework occurs affected examiners' decisions. These are matters that researchers must consider when evaluating the merits of particular studies and the implications of research results.

Guideline #5: Researchers should report error rate data separately for presentations, total comparisons, and comparisons that led to conclusive results.

The table below shows hypothetical data from a black-box study assessing the accuracy of forensic examiners when comparing impressions to determine whether they were made by the same item or different items. We do not specify what type of items are involved, as the reporting format is generic and could be used in a wide variety of disciplines, including latent prints, tool marks, footwear impression, bite marks, etc. The reporting format is similar to that used to report the FBI's black-box study of fingerprint examiners (Ulery, 2011). It shows how findings of a black-box validation study can be reported, in order to display all relevant error-rate data.

Table 1: Data from a Hypothetical Validation Experiment for a Source Determination Method (Showing Error Rate Calculated Three Ways for Same-Source and Different Source Comparisons)

Examiners' Finding	Source of Sample Pair							
	Same Source				Different Source			
	#	% PRES	% COMP	% CALLS	#	% PRES	% COMP	% CALLS
No value (not compared)	300	30			100	10		
Inconclusive	300	30	42.8		100	10	11.1	
Exclusion	40	4	5.7	10	790	79	87.7	98.75
Inclusion	360	36	51.4	90	10	1	1.11	1.25
Total Calls	400				800			
Total Comparisons	700				900			
Total Presentations	1000				1000			

In this hypothetical study 100 examiners were each presented 20 pairs of impressions and were asked to determine whether each pair of impressions was made by the same item or different items. For each examiner, half of the pairs were same-source (made by same item) and half were different-source (made by different items). Examiners first determined whether the impressions were suitable for comparison; if they found that either impression was of no value, then no comparison was made (the Table shows that 300 of the same-source pairs and 100 of the different source pairs were found to be of no value). Examiners carefully compared all pairs determined to be "of value" and reported their conclusion as either an inclusion (same source), exclusion (different source), or inconclusive.

The data in the table show the number of pairs presented to examiners; the number of those that were from the same source (1000) and from different sources (1000); and a breakdown of examiners' findings for each type of comparison. The key findings regarding accuracy are the number and percentage of same-source pairs that examiners mistakenly excluded (false negatives; marked in yellow) and the number and percentage of different source pairs that examiners mistakenly included (false positives; marked in turquoise). Importantly, false positives and false negatives are reported three ways: (1) as a percentage of all presentations (% PRES); (2) as a percentage of all comparisons, i.e., excluding those comparisons where the impressions were deemed to be of no value (% COMP); and (3) as a percentage of all conclusive calls, i.e., excluding both no value comparisons and inconclusive (% CALLS). PCAST advocates reporting error rate data as a percentage of conclusive calls (ignoring no value and inconclusive comparison), on grounds that cases where examiners reached a conclusion are those likely to be used in a criminal proceeding, and hence the rates of error for those conclusions are most relevant. Our view is that forensic scientists should be prepared to present error rate data for their methods in a variety of ways. An advantage to presenting data in the tabular form shown

here is that it allows interested parties to easily see the differences in different error rates and to focus on whichever they deem most relevant. [3]

The data in this hypothetical study show a relatively low rate of false positives (about 1%) and a somewhat higher rate of false exclusions (4-10%, depending on how calculated), which indicates the level of accuracy of the method. Accuracy data of this kind would clearly be helpful in assessing the validity of a forensic method for source determinations. By studying the pattern of results, researchers can also gain clues about the adequacy of the study and how well the findings are likely to generalize to casework. For example, the higher rates of false exclusions than false identifications may indicate that participants in the study were being more cautious about declaring "inclusions" than "exclusions." Researchers would need to consider whether decision thresholds applied by participants in this study are likely to be the same or different than the thresholds applied in routine forensic practice. Another finding warranting consideration is the higher rate of "no value" determinations for same-source than different source test specimens. This might indicate a bias in the selection of same-source and different-source specimens used in the study (which could raise concerns about the representativeness of those specimens), or it could arise from a systematic tendency in examiners' decision-making about sample suitability, which could be important for understanding and potentially improving examiners' decision making process. In any event, by reporting validation data in the manner illustrated in Table 1 researchers can display their findings in a way that allows a fair and complete assessment of the validation study and its findings.

The example above concerns validation of a method for source determination. A similar approach can be taken to display data for a validation study of a method for classifying samples by type. The reporting table should show the types and numbers of the known-type test specimens that were presented for examination, broken down by the examiner's determinations of the types, including data on the rates at which examiners found the test specimens unsuitable for analysis or reported the results as inconclusive.

Guideline #6: Validation studies should measure and report the level of difficulty entailed in the tasks participants are asked to perform.

It is very likely that the error rates (and rates of inconclusive findings) of many forensic methods and procedures will differ with sample difficulty and sample types. For example, the AAAS report on latent fingerprint examination recently found that error rates of fingerprint examiners "were higher in studies for which the comparisons were more difficult." (AAAS, 2017, p. 45). In light of this variation, the AAAS report declared that:

[3] Some authorities (e.g., PCAST, 2016) have recommended that forensic scientists compute confidence intervals around these estimates and report false positive rates as the 95% upper limit of the confidence interval for comparisons that led to a conclusive determination of exclusion or individualization (i.e., excluding presentations that led to a determination of no value or inconclusive). For information on how to compute a confidence interval using data like that in this table, see the appendices to the PCAST report.

> …it is unreasonable to think that the "error rate" of latent fingerprint examination can meaningfully be reduced to a single number or even a single set of numbers [ref omitted]. At best, it might be possible to describe, in broad terms, the rates of false identifications and false exclusions likely to arise for comparisons of a given level of difficulty (AAAS, 2017, p.45)

Consequently, researchers who are designing studies to measure the accuracy of forensic examiners on various tasks should measure the difficulty of the comparisons examiners are asked to perform. If tests have varying levels of difficulty then results that control for these differences should be presented. Having adequate metrics for difficulty will help assure that the relevance of error rate data for casework can be assessed in an apples-to-apples manner. It would be misleading, for example, to conclude that high error rates from a study designed to be extremely challenging for examiners reflect the likelihood of error in cases where examiners make more straightforward, easy comparisons; and vice-versa. However, for these distinctions to be truly useful, objective measures of what makes for a level of difficulty should be specifically defined.

Researchers who study latent fingerprint analysis have made substantial progress in developing measures of the difficulty of latent print comparisons (Hicklin, Buscaglia and Roberts, 2013; Kellman et al., 2014; Yoon et al., 2013). Researchers in other fields should develop such measures as well, and these measures should be incorporated into studies of examiner accuracy so that the implications of the findings will be better understood and easier to apply.

Ideally, to determine validity researcher will study the accuracy of forensic examiners in a particular discipline using a wide range of samples that are representative of those encountered in actual casework. In reporting the results of validity studies, the nature of the test samples should be reported along with how levels of difficulty were ascertained, and level of difficulty should be taken into account in interpreting a validation study's results and in evaluating its implications.

Guideline #7: To accurately estimate a method's operational error rate, examiners should be "blind" to the fact they are being tested—that is, they should not know that samples are test samples rather than ordinary casework.

Psychologists who study human performance in domains other than forensic science have long noted that test performance is affected by whether test takers know they are being tested. Knowing that one is being tested changes people's responses (Orne, 1962). Problems may be approached in different ways than they would be if the problems were thought to be routine, and test takers may shift their decision thresholds in ways designed to produce desirable outcomes (Paulhus, 1991). Hence, error rates observed in open testing (where people know or can easily figure out that they are being tested) may not reflect error rates in ordinary practice. For additional discussion of this point, see the AAAS report on latent fingerprint examination (AAAS, 2017, at pp. 46-51).

One way around this problem is to construct "blind" tests in which examiners do not know their performance is being tested. This can be done by incorporating research specimens into the routine flow of casework in a manner that makes them indistinguishable from other items examined by the laboratory. A number of authorities have urged that blinded studies of examiner accuracy be conducted (*see*, National Commission on Forensic Science, 2016; AAAS, 2017, at pp. 47-51; PCAST report, at p. 59).

We recognize that blind studies are difficult to conduct in laboratories where examiners communicate directly with detectives and have access to police reports and other information. To conduct blind tests in these settings laboratory managers will need to enlist the support of law enforcement in preparing simulated case materials that are sufficiently realistic to pass as real cases. Elaborate simulations of this type are feasible, although they are burdensome and expensive.[4] Their value, however, justifies the expense of periodically paying for such evaluations.

Fortunately, blind studies are easier to conduct in laboratories that employ context management procedures to shield examiners from task-irrelevant contextual information.[5] In such laboratories, it is possible for laboratory managers to insert research test specimens into the flow of casework in a manner that is undetectable without the need to involve personnel outside the laboratory in a deception. The case manager knows which items come from actual casework, and which are items are test specimens prepared for research, but (if care is taken) the examiners do not know.

Successful blind testing programs of this type have been implemented in forensic laboratories (Found & Ganas, 2013; Kerkhoff, Stoel, Berger, Mattijssen, Hermsen, Smits & Hardy, 2015; Mattijssen, Kerkhoff, Berger, Dror & Stoel, 2016).[6] The ability to implement blind testing in this way is a secondary benefit that arises when laboratories adopt context management procedures (National Commission on Forensic Science, 2016). Forensic scientists who are interested in conducting blind studies of examiner performance are advised to contact laboratories that have attempted such studies and academic experts who have participated in such research, in order to learn from their practical experience.

[4] See, Peterson, et al, 2003 (discussing a pilot test of blind testing of forensic DNA laboratories). Blind studies of this type have been conducted successfully by the U.S. Army Defense Forensic Science Center.

[5] The Human Factors Committee regards the implementation of context management procedures as good practice apart from the way it facilitates blind testing, for reasons explained by the National Commission on Forensic Science (2015); Stoel et al. (2014); Dror et al., (2015); and Found & Ganas (2013).

[6] For example, the Houston Forensic Science Center has been conducting blind testing in three of its disciplines (controlled substance, blood alcohol, and firearms analysis) and is planning to expand the blind testing program to latent print analysis and DNA analysis. Similar programs have been adopted by the Netherlands Forensic Institute and (under the direction of Bryan Found) by the document examination section of the Victoria Police Forensic Services Department in Victoria, Australia.

Non-blind studies can be useful for some purposes. While blind studies (in which examiners do not know they are being tested) have important advantages and should be conducted where possible as part of forensic science validation, non-blinded studies in which examiners know they are being tested, but do not know the correct answer, also have legitimate uses.

- Proficiency testing, which often is non-blind, is vital for establishing that examiners have gained the minimal level of proficiency needed to perform competently.

- Non-blind tests may be useful for providing feedback to examiners during training on their level of performance; it may help them identify strengths and weaknesses as they develop their analytic skills. These tests may also help examiners maintain and improve skills during their professional careers.

- As the PCAST report has noted, non-blind black-box studies, such as the well-known studies of latent print examiners, are valuable for establishing the foundational validity of forensic disciplines—i.e., for demonstrating that examiners are capable of discriminating accurately between same-source and different-source prints. Error rates in such studies must, however, be viewed with caution; error rates in blinded tests and in actual casework may well be different. Research in other fields indicates that error rates are sometimes lower when people know their performance is being assessed.

Although we are discussing testing in which examiners do not know the correct answer, testing can be of some value even if the correct answer is known. For example, it can be a way for examiners to self-check their performance. Of course, such tests should not be used by management to measure how well examiners are performing.

Guideline #8: The conditions examiners confront in validity studies should mirror ordinary conditions that examiners face when performing the type of casework being studied.

To obtain an accurate estimate of the error rate of a forensic procedure, the test specimens used in the study must be similar to items typically examined in casework and the examination must be conducted in the same manner as normal casework. To accomplish this, the following guidelines are recommended:

- ***Representativeness of Test Specimens.*** The test specimens must be representative of the range and difficulty of the items that come to the laboratory as ordinary casework, for the type of casework being studied. If the study is designed to test the performance of a laboratory for casework in general, then the samples should represent the full range and distribution of types and difficulty normally seen in the laboratory. If the research is designed to test the accuracy of the laboratory for a particular type of case (e.g., mixed DNA samples; low-quality latent prints), the range of test items can be limited to items of that class, but the test items should still be representative of the range and difficulty of the items within that class.

- ***Study Administration.*** As indicated above, there are important advantages to inserting test specimens into the normal flow of casework in a manner that leaves examiners unaware of which items might be test specimens (and, ideally, to keep them unaware that they are being tested). If that cannot be done, researchers should nevertheless make their best efforts to keep the conditions of the study as similar as possible to ordinary casework in all other respects.[7]

- ***Characteristics of Test Specimens.*** In blinded studies, the test specimens should give no hint of whether they are test specimens or routine evidentiary items, which means not only that the two cannot be distinguished but also that they way they are presented should give no hint that they differ.

- ***Checking that "blind" studies are truly blind.*** When conducting a blinded study, it is good practice to institute procedures for checking whether the supposedly blind samples are being detected as test specimens. Examiners are often quite perceptive about the source of the items they examine and may occasionally suspect or know that an item is a test specimen despite the best efforts of the researchers to make it "blind." The Director of the Houston Forensic Science Center has been dealing with this problem by offering an incentive (a Starbucks gift card) to examiners who correctly identify test specimens in the lab's blind testing program, while charging a small fee to examiners who guess incorrectly that an item is a test specimen. Feedback from this incentive process has allowed laboratory managers to improve their blinding procedures and reduce the chances that blind test specimens will be recognized.[8]

- ***Manner of Reporting:*** Conclusions drawn during reliability and validity testing should be reported using the same categories as are normally used in practice. "Unblinding," to give examiners feedback on their performance, should not occur until all results are finalized and securely reported.

[7] Researchers may sometimes have good reasons for asking examiners to follow different procedures than they normally would. For example, a researcher might wish to determine whether a variation in approach would improve performance. Or, a researcher may need examiners to record more, or different, judgments than they ordinarily would in order to gain greater insights into examiners' mental process, confidence, decision threshold, etc. (For an interesting example of such a study involving latent print analysis, s*ee,* Thompson, Tangen & McCarthy, 2014). Studies of this type can be valuable for gaining insights that will lead to improvement of forensic science as a discipline, but error rates observed in such studies probably do not reflect error rates in ordinary practice. Hence, studies of this type are not the best way to measure the accuracy of existing methods.

[8] Information about the incentive program was provided to the Human Factors Committee by Peter Stout, the Director of the Houston Forensic Science Center, who is a strong advocate of blind testing programs in forensic laboratories.

Guideline #9: Adequacy of Sample Size

Numbers of Samples and Examiners: The accuracy of a method or procedure cannot be adequately tested with small numbers of test specimens, or with small numbers of examiners.

- ***Number of Test Specimens***. Most validation studies will involve test specimens that vary in how difficult they are to examine. In order to properly assess the accuracy of a method, it is important to include adequate numbers of test specimens at each level of difficulty. Results obtained with small samples often vary greatly due to random factors. Results with larger samples are more stable and tend to better represent results for the underlying population of similar items. Hence, it is important to include a sufficient number of samples of each type or difficulty level.

- ***Number of Examiners.*** To establish the accuracy of a method or procedure, it should be tested using multiple examiners. As noted above, the overall validity of a method is limited by its reliability, and testing more than one examiner is necessary to assess reliability across examiners (sometimes called inter-rater reliability). When ground truth is known, examining findings across multiple examiners helps distinguish mistakes that arise from deficiencies in the training or skills of a particular examiner from errors that arise from more general limitations of the method.

In determining the appropriate number of samples for a validation study, researchers would be well advised to consult with statisticians or other academic experts familiar with statistical power and sample size requirements for experimental research.

Guideline #10: Maintaining the objectivity of research personnel.

Validation research should be conducted in a manner that is as objective as possible. That may require steps to assure that the motives, desires or perspectives of the research staff do not create biases in the design of the study, how the study is conducted or how the results are interpreted and reported. Experience has shown that the motives and interests of researchers can influence how they conduct studies (often unintentionally) in ways that create bias. Steps should be taken to minimize this risk.

It is generally a good practice to involve disinterested experts (those with no stake in the outcome of the study) in the design and analysis of studies. Crucial design questions, such as the nature of the test specimens and how they will be presented, should be made (at least in part) by individuals who have no particular stake in the outcome of the study.

It is also good practice that those who interact directly with research participants be blind to the expected results. Experience has shown that non-blind research staff can sometimes unintentionally provide subtle cues to study participants that may hint at or guide them to the

correct answer. Proper research procedures will insulate examiners participating in the study from even subtle hints.

In blind testing situations (where examiners do not know they are being tested), care must be taken to avoid providing subtle hints about when testing is underway and which samples are test samples. While laboratory managers must, of course, be aware of blind testing programs, it often is possible to find ways to limit the knowledge of those who interact most directly with examiners in order to minimize the potential for such problems. The experience of other managers who have undertaken such research can help guide new researchers toward possible approaches.

We emphasize that the problem to be addressed is not that laboratory administrators will *intentionally* release biasing information. Rather, it is the risk (revealed by numerous studies) that research participants can be influenced unintentionally by study administrators, even when the administrators are acting in good faith and appear to be doing nothing aimed at influencing test results. Subtle leakage of cues can nevertheless occur and must, therefore, be guarded against.

We therefore recommend the following as best practices for such research:

- **Disinterested individuals should be involved in as many aspects of the study as possible.**

- **Those who communicate directly with study participant should be kept as blind as possible to information about the correct answers and, when feasible, to information that a particular item is a test specimen.**

Laboratory managers who wish to conduct performance testing studies would be well advised to consult with academics on study design and analysis. Involving managers of other laboratories might also be helpful.

Guideline #11: To assess the accuracy of a method, the procedural steps associated with the method must be fully specified and care must be taken to assure that this method is followed.

It is difficult to draw conclusions that can be generalized from studies of examiner accuracy if the exact method that examiners use to make a comparison is unspecified and may differ across examiners in idiosyncratic ways. For example, it might be misleading to talk about the validity of the ACE-V method for tool mark comparison if different examiners implement the method in very different ways. It will not be clear whether inaccurate results arise from mistakes in applying the method or from the margin of error inherent in the method itself.

Hence, to rigorously test the validity *of a method* it is necessary to:

- ***Fully Characterize the Method:*** Practitioners must first specify what steps and procedures the method entails. This must be done in sufficient detail so that examiners can learn to reliably follow the method, and perform it as intended, and in the same way as other examiners do.

- ***Ensure That Examiners Use the Method Properly and Follow All Required Steps During the Test.*** Researchers must ensure that the examiners, during the test, are in fact following the specified method as intended. This may require testing examiners in advance to ensure that only examiners who know how to implement the method properly are included in the test, monitoring the examiners during the test or providing aids like check lists to be sure that all proper procedures are being followed, and/or reviewing the procedures followed post-test to ensure that all required steps were taken.

Guideline 12: ***Systematically document research procedures and results. Share information on study procedures and findings with the scientific community in an open and transparent manner.***

Researchers should develop a data collection plan that allows for systematic, comprehensive and transparent documentation of what takes place in the study, including preparation of test specimens, recruitment of participants, how test specimens were presented to participants, participants' judgments, and any processing or analysis of the resulting data.

It is appropriate to delay release of research findings in order to give those who conducted the study the first opportunity to publish. Once researchers publish their findings, however, they should openly share their study materials and data with interested parties, and particularly with academics and fellow researchers. The scientific obligation of transparency and sharing applies even (and perhaps especially) when it is suspected that those asking to examine the study materials are approaching the matter with a critical eye, or with a desire to challenge the findings. Openness and transparency are basic norms of science that should not be put aside in order to advocate for, or defend, a particular point of view. Secrecy is anathema to science.

In some instances the privacy interests of those who participate in such studies, either as research participants or by providing test specimens, may justify limiting the amount of information released. But individual privacy interests can usually be addressed by releasing information in anonymized form—that is, by removing details that could be used to identify particular individuals.

V. Guidelines for Performance Assessments of Individuals

This document has focused primarily on testing the performance of forensic examiners for the purpose of assessing the validity of commonly used methods. But the guidelines set forth in the previous section also apply to research studies designed to assess the performance of specific individuals. Generally, the methods for testing the performance of individual examiners should be the same as those used for validity testing, with a few exceptions.

When researchers focus on validation of a particular method, it is important to assure that examiners follow that method in a uniform and correct manner. Such uniformity is not necessary, and may even be undesirable, when the focus is on individual performance. For example, the goal of the research may be to identify individual examiners who are outstanding performers, and then to study how their techniques differ from those of less accomplished examiners. Hence, for such studies, the researcher may wish to encourage (rather than discourage) individuality of approach.

By focusing on the performance of specific individuals, rather than aggregate performance of a group of examiners, researchers can identify and provide feedback on the strengths and weaknesses of individual examiners relative to one another. Testing of this type may help managers determine which examiners are best for a particular type of examination (Found & Ganas, 2013). It can also provide valuable feedback to help examiners identify their strengths and weaknesses in order to improve their skills. As the National Commission on Forensic Science has noted:

> 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 their judgments are mistaken, incomplete, or otherwise suboptimal (National Commission, 2016).

The National Commission noted that forensic science practitioners will rarely know with certainty whether the judgments made in casework are accurate, and hence that: "There is only one consistently effective way to provide valid feedback to analysts on the accuracy of their performance—they must be tested using samples for which the ground truth is known." (National Commission, 2016). Implementing testing programs that allow such feedback is therefore an important component of laboratory quality assurance as well as an important tool of professional development.

What Can Be Tested?

Individual performance can be characterized in terms of accuracy (validity), reliability, and other factors (e.g., speed, efficiency). It is important to understand, however, that many trials may be needed to obtain stable estimates of performance (e.g., error rates) for an individual examiner.

- ***Accuracy (Validity).*** As is true for the test itself, an individual examiner's performance can be characterized with respect to overall accuracy, as well as with respect to correct inclusions versus exclusion, false positives, false negatives, and rate of inconclusiveness or inadequacy of the samples. These can also be characterized separately by examiner for types of examination and levels of difficulty.

- ***Reliability.*** An individual examiner's performance can also be characterized in terms of reliability.

    - This can be done if the examiner tests the exact same sample on different occasions, but is likely to be meaningful only if the examiner does not realize that a test is a repeat examination, or if enough time has passed between two examinations so that the examiner is unlikely to recall past findings. Repeat testing of this sort assesses examiner-specific test-retest reliability and not inter-examiner reliability. Such assessments are useful although they are not always practicable.

    - If more than one examiner has independently tested the same sample without knowing what others have found, then the inter-examiner reliability of a given examiner's judgments may also be characterized, but without knowing ground truth this alone will not indicate whether a particular examiner is more or less accurate than others who have examined the same sample and reported different results.

Test Sample Requirements for Individual Examiner Performance Testing

As with all testing, there must be sufficient observations to ensure that a stable and reliable performance evaluation is obtained. To obtain meaningful results, it may be necessary to assess an examiner's accuracy for various item types and for various levels of difficulty within item type. Laboratories can consult with statistical experts to determine the number of tests that will strike a sensible balance between what a lab needs to learn and the cost of acquiring that knowledge.

VI. Concluding Note: The Importance of Validation in Forensic Science

Why should forensic scientists conduct empirical studies to assess the accuracy of their methods? While validation is necessary in all scientific disciplines it is particularly important in forensic science due to the consequences that often hang on the results of a single forensic science analysis or comparison. The judgments of a DNA analyst, latent print examiner or tool mark examiner, based on a single comparison, can have huge consequences for human lives. It

is the importance of forensic scientists' conclusions to the justice system that makes it vital to have data on their accuracy.

Which methods require empirical validation? We believe forensic scientists will conclude that, with few exceptions, the analytic methods used for comparing items to determine whether they have a common source, or for classifying items into categories, should be assessed for accuracy using test specimens of known source or type. These methods will also need follow-up reliability checking because, as discussed above, the overall validity of a method depends on its being reliably employed.

There will, of course, be debate about how extensive such research needs to be and about the best ways to conduct such research. This document is designed to help forensic scientists evaluate and decide those issues in an informed and thoughtful manner.

References

American Association for the Advancement of Science (AAAS)(2017). *Forensic Science Assessments: A Quality and Gap Analysis: Latent Fingerprint Examination*. Available at: https://www.aaas.org/report/latent-fingerprint-examination

Dror I.E., Thompson W.C., Meissner C.A., et al. (2015). Context Management Toolbox: A Linear Sequential Unmasking (LSU) Approach for Minimizing Cognitive Bias in Forensic Decision Making. *Journal of Forensic Sciences*, Vol. 60, No. 4, pp. 1111-1112.

Found, B., & Ganas, J. (2013). The management of domain irrelevant context information in forensic handwriting examination casework. *Science and Justice*, 53, 154-158.

Hicklin R. A., Buscaglia J., and Roberts M. A. (2013). Assessing the Clarity of Friction Ridge Impressions. *Forensic Science International*, 226(1), 106-117.

Kellman P. J., et al. (2014). Forensic Comparison and Matching of Fingerprints: Using Quantitative Image Measures for Estimating Error Rates through Understanding and Predicting Difficulty. *PLoS ONE*, 9(5), 1-14.

Kerkhoff, W., Stoel, R. D., Berger, C. E. H., Mattijssen, E. J. A. T., Hermsen, R., Smits, N., & Hardy, H. J. J. (2015). Design and results of an exploratory double blind testing program in firearms examination. *Science & Justice*, 55(6), 514-519.

Martire, K. A., & Kemp, R. I. (2016). Considerations when designing human performance tests in the forensic sciences. *Australian Journal of Forensic Sciences*, 1-17.

Mattijssen, E. J. A. T., Kerkhoff, W., Berger, C. E. H., Dror, I. E., & Stoel, R. D. (2016). Implementing context information management in forensic casework: Minimizing contextual bias in firearms examination. *Science & Justice*, 56(2), 113-122.

National Commission on Forensic Science, Views of the Commission: *Facilitating Research on Laboratory Performance* (adopted unanimously September 13, 2016). Available at: https://www.justice.gov/ncfs/page/file/909311/download

National Commission on Forensic Science, *Ensuring That Forensic Analysis is Based Upon Task-Relevant Information* (adopted December 8, 2015). Available at: https://www.justice.gov/archives/ncfs/file/818196/download

Orne, M. 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.

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, CA: Academic Press.

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.

President's Council of Advisors on Science and Technology (PCAST) (2016). *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Executive Office of the President, September 2016. Available at: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf

Stoel R.D., Berger C.E.H., Kerkhoff W., et al. (2014). Minimizing Contextual Bias in Forensic Casework, in *Forensic Science and the Administration of Justice: Critical Issues and Directions*. KJ Strom and MJ Hickman, Eds. Sage Publications, pp. 67-86.

Thompson M.B., Tangen J.M. and McCarthy D.J. (2014). Human Matching Performance of Genuine Crime Scene Latent Fingerprints. *Law and Human Behavior*, Vol. 38, No. 1, pp. 84-93.

Ulery B.T., Hicklin R.A., Buscaglia J. and Roberts M.A. (2011). Accuracy and Reliability of Forensic Latent Fingerprint Decisions. *Proceedings of the National Academy of Sciences*, USA, Vol. 108, No. 19, pp. 7733-7738.

Yoon S., et al. (2013). LFIQ: Latent Fingerprint Image Quality. *Biometrics: Theory, Applications and Systems* (BTAS), IEEE Sixth International Conference on Biometrics Compendium, pp. 1-8.