OSAC Technical Series 0004

Human Factors in Validation and Performance Testing of Forensic Science

https://doi.org/10.29325/OSAC.TS.0004

OSAC Human Factors Committee

OSAC
Organization of Scientific Area Committees for Forensic Science

OSAC Technical Series 0004

Human Factors in Validation and Performance Testing of Forensic Science

Prepared for
The Organization of Scientific Area Committees (OSAC) for Forensic Science

Prepared by:
Human Factors Committee
Organization of Scientific Area Committees (OSAC) for Forensic Science

March 2020

https://doi.org/10.29325/OSAC.TC.0004

Document Disclaimer: This publication was produced using a consensus process, as part of the Organization of Scientific Area Committees (OSAC) for Forensic Science and is made available by the U.S. Government. Consensus for the purposes of the OSAC Technical Series publications means that all OSAC members had an opportunity to comment on the document and provide suggestions for revisions. Consensus does not mean that all OSAC members are in complete agreement with the contents of this publication. The views expressed in this publication and in the OSAC Technical Series publications do not necessarily reflect the views or policies of the U.S. Government. The publications are provided "as-is" as a public service and the U.S. Government is not liable for their contents.

Certain commercial equipment, instruments, or materials are identified in this publication to foster understanding. Such identification does not imply recommendation or endorsement by the U.S. Government, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Table of Contents

I. Introduction....1
II. Scope of Application....2
III. Definition and Explanation of Key Terms....3
IV. Distinguishing Consistency from Accuracy....6
V. Key Issues in Designing, Conducting, and Reporting Validation Research....8
Preliminary Considerations....8
Creating and Selecting Test Specimens: Variety and Number....11
Study Participants and Procedures....13
Analyzing Data and Reporting Results....17
Disseminating Results....23
VI. Internal Validation and Quality Assurance....24
VII. Concluding Note: The Importance of Validation in Forensic Science....26
VIII. References....27

List of Tables

Table 1. Data from a Hypothetical Validation Experiment for a Source Determination Method....18

Table 2: Data from a Hypothetical Validation Experiment for a Source Determination Method with a Five-Point Reporting Scale....20

I. Introduction

This publication offers advice on designing, conducting and reporting empirical studies on the accuracy of forensic examinations.[1] By offering suggestions on research that might be done and practices that might be developed in the future, this publication aims to help OSAC subcommittees develop and refine statements about the research needs of their disciplines. More broadly, it aims to help forensic scientists enhance their vision of ways forensic science might develop in the future and thereby facilitate continuing incremental improvements in forensic science standards and practice.

This document is an OSAC Technical Series Publication[2] rather than a standard or guideline. It establishes no requirements for current or future practice; it merely provides advice and suggestions. The information provided here was distilled from an extensive scholarly literature on human performance testing and the science of evaluation research[3] as well as from the practical experience of the HFC[4] and other OSAC members.

[1] For additional discussion of the same issues, readers should consult Martire and Kemp (2018).

[2] OSAC Technical Publications are commentaries designed to provide background and perspective on issues relevant to the standards development process. The Forensic Science Standards Board (FSSB) described the purpose and requirements of a Technical Series publication in a document titled "OSAC Technical Series Publication Process," September 28, 2018:

> The purpose of this series is to share information that was gathered during the analysis and development of documentary standards. The OSAC Technical Series publications are not intended to be used as standards documents and do not receive the same level of review as consensus standards that go through a standard developing organization (SDO).

This publication was prepared by the OSAC Human Factors Committee (HFC). Drafts of this publication were twice posted for public comment and were revised in light of comments received. It was also reviewed and vetted by OSAC's Scientific Area Committees (SACs) and Forensic Science Standards Board.

[3] Evaluation research uses social science methods to evaluate the performance of individuals or organizations at specific tasks. It sometimes employs special techniques to mitigate potential biases and distortions that arise when human beings know they are being studied (Powell, 2006).

[4] Members of the HFC have expertise in social and behavioral science disciplines that involve the study of human decision making and assessment of human performance, including the performance of experts. The need for social science expertise is widely recognized by scientists who study expert performance (Bozeman & Youtie, 2017; National Research Council, 2015). Major studies in clinical medicine, for example, are often performed by interdisciplinary teams that include psychologists and statisticians as well as physicians (see, e.g., Connors et al., 1995).

II. Scope of Application

The research strategies discussed here are helpful for establishing the range of validity of new forensic science methods and for demonstrating the range of validity of older methods. We discuss ways to test the accuracy of forensic science practitioners when they perform 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).[5] The research strategies described here may also be useful for other purposes beyond validation, such as assessing the effectiveness of training, identification of strengths and weaknesses of individual examiners, and even assessment of the strengths and weaknesses of laboratory systems. We discuss some of these additional purposes toward the end of this publication.

We focus primarily on assessment of practitioners' accuracy when performing analytic tasks that require the exercise of human judgment and expertise. Some of what we say about research design and reporting may also be relevant to assessing the performance of automated systems, but a full discussion of the validation of automated systems is beyond the scope of this publication.[6]

This publication does not address the testing of examiner performance on other tasks (beyond source determination or classification of items by type). Among the tasks that are not addressed here are:

- quantitation
- tasks that do not entail reaching a reportable finding on source or type (e.g., sample collection; sample preparation; instrument set-up and calibration)
- tasks that involve recognition of relevant evidence rather than reporting results about source or type of specific items (e.g., identification of relevant evidence at a crime scene)
- tasks that involve causal analysis (e.g., cause of death; cause of a fire)
- 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; assessment of manner of death)

It may be important to test examiner performance on such tasks, and some of the commentary offered here may be relevant to such assessments, but that is not the focus of this publication.

The way in which forensic science practitioners report their findings must be considered when researchers design and report studies of the accuracy of those findings. Because forensic scientists in the United States have traditionally reported most of their findings categorically, using reporting categories like "identification," "inconclusive" or "exclusion," our primary focus in this publication is on ways to test the accuracy of categorical findings. This requires research designed to estimate rates at which items of known source or type are correctly and incorrectly categorized. For example, a validation study might examine the rate of true and false identifications, and of true and false exclusions, that occur when a method is employed for making source determinations.
Our primary focus in this publication is on studies of this type.

[5] We recognize that testing the accuracy of a method is only one aspect of method validation. For a broader discussion of the validation of forensic science methods, *see* Forensic Science Regulator (2014).
[6] For discussions of the validation of automated systems, see, Ramos, Gonzalez-Rodriguez, Zadora & Aitken (2013); Meuwly, Ramos & Haraksim (2017); Haned, Gill, Lohmueller, Inman & Rudin (2016).

In recent years, forensic scientists in some disciplines have adopted non-categorical approaches to reporting, such as presenting likelihood ratios (LRs) and offering other statements about the strength of evidence (Aitken, Berger, Buckleton et al. 2011). To assess the accuracy of these kinds of results researchers must design their studies and report their findings a bit differently.

We discuss special issues researchers face when evaluating the accuracy of LRs toward the end of this publication, in a section titled: "Issue 10—Special problems in assessing the accuracy of likelihood ratios." It is important to note, however, that much of what we discuss in this document applies broadly to studies of practitioner performance, regardless of the reporting format or analytic framework. [7]

Finally, the focus of this publication is on how empirical research might ideally be designed and carried out to assess the validity and reliability of methods, assess performance, and meet quality assurance goals. This publication does not consider the costs of such research, nor does it attempt to balance the benefits of such research against the costs and difficulties of conducting it. This publication does not attempt to assess when or whether studies should be mandatory rather than optional. The goal of this publication is to provide information and insights that will assist OSAC subcommittees, and forensic scientists more generally, as they consider those important issues.

III. Definition and Explanation of Key Terms

Accuracy—The OSAC Lexicon defines accuracy as: "closeness of agreement between a test result or measurement result and the true value." In this document we will say that a method for determining source or type of an item is accurate (or has accuracy) when the result produced by the method corresponds to the ground truth regarding source or type. When assessing the accuracy of a method for source determination, it is important to distinguish accuracy when comparing items of same source (*see* Sensitivity) and accuracy when comparing items of different source (*see* Specificity).

Black-Box Study—A black-box study assesses the accuracy of examiners' findings without considering how the findings were reached. The examiner is treated as a "black-box" and the researcher measures how the output of the "black-box" (examiner's finding) 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.

Consistency—According to the definition of consistency in the OSAC Lexicon "consistent measures are those where repeated measurements of the same thing produce the same results." In this document, the terms consistency and reliability are used as synonyms (see Reliability).

Context Management Procedure—A procedure designed to limit or control what a forensic examiner knows about the background or circumstances of a criminal investigation at a point in time or stage of analysis 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 (*see*, Risinger et al. 2002; Thompson, 2011; Found & Ganas, 2013; Stoel et al., 2015; Dror et al., 2015; National Commission, 2015).

[7] This publication does not address the issue of how forensic scientists should present their findings; it neither endorses nor recommends any particular reporting language, most notably with regard to the use of categorical reporting scales in source attribution or use of verbal predicates with likelihood ratios (see Issues 9 and 10 below).

Ground Truth—The actual or true state of affairs concerning the source or type of items submitted for evaluation—e.g., whether fingerprints submitted for comparison were made by the same finger or not; whether a shoeprint submitted for evaluation of its size and tread pattern was made by a shoe of given size and tread pattern**.**

Reliability—The OSAC Lexicon offers two definitions of the term reliability. "Reliability, evidentiary/legal" refers to "credibility and trustworthiness of proffered evidence." In this publication we adopt the second definition, referenced in the Lexicon as "reliability, statistical." The Lexicon defines this type of reliability as "consistency of results as demonstrated by reproducibility or repeatability." This document treats the terms reliability and consistency as synonyms. As we use these terms, reliability (consistency) 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 two or more examiners who reach the same result (consistency) when asked to examine or compare the same items.[8]

Sensitivity—Forensic scientists sometimes use the term sensitivity to refer to a threshold of detection, for example the level of concentration necessary to obtain a positive result in a test procedure designed to detect the presence of a specific substance. In statistics, by contrast, the term sensitivity is typically used to refer to the rate of true positives in a classification task—for example, the rate at which an examiner determines that same-source specimens have the same source. This publication uses the statistical definition.

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

The chart above is useful in explaining the meaning of the term sensitivity, as used here. It shows the accuracy of examiners' decisions in a hypothetical binary classification task: deciding whether two items have the same source or a difference source. (Correct decisions are noted in bold).[9]

Sensitivity refers to the probability that examiners will deem two items to be from the same source when they are from the same source. Thus, the proportion **A**/(A+C) provides an estimate of sensitivity. For example, if 100 examiners, all applying the same method, are each given 10 trials for

[8] The reliability of a measurement instrument (i.e., its consistency over repeated measurements on the same items) is sometimes referred to as its precision, but we elected not to use the term precision in this document because the term is sometimes used differently by others in the scientific community.

[9] Our use of the term "sensitivity" in this document should also be distinguished from "sensitivity analysis," which is the analysis of how the uncertainty in the output of a mathematical model or system can be divided and allocated to different sources or inputs—for example, an analysis of how much the output of a probabilistic genotyping system might be affected by uncertainty about specific modeling parameters, such as peak height variation.

which the correct answer is "same source," and they concluded "same source" 850 times and "different source" 150 times, their sensitivity, as measured in this experiment, would be 850/(850+150)=0.85, or 85%.

Sensitivity is sometimes also called the "hit rate" or the "true positive rate." The sensitivity of a method for source determination is the accuracy of the method when it is used to compare items having the same source. A decision that two items have a different source when actually they have the same source is sometimes called a "false exclusion." [10] In the simplified situation shown in the chart, in which the examiner has two possible decisions, the rate of false exclusions is equal to 1 minus sensitivity.

Specificity refers to the probability that examiners will deem two items to be from a different source when they are actually from different sources. Thus, in the chart above, specificity is equal to **D**/(B+D). For example, if 100 examiners were each given 10 trials for which the correct answer is "different source," and they said "different source" 900 times and "same source" 100 times, their specificity, as estimated by this sample of decisions, would be 900/(100+900)=0.90 or 90%.

Specificity is sometimes called the "true negative rate" or the "correct rejection rate." The specificity of a method for source determination is the accuracy of the method when it is used to compare items having a different source. Specificity is directly related to the false inclusion rate of the test, which is B/(B+D).[11] As the specificity increases, the false inclusion rate will decrease because together they add to 100% (for simple, binary decisions). For example, if the examiner, when comparing items from different sources, correctly decides they are different 95% of the time, then the rate of incorrect decisions that they are the same (false inclusions) will be 5%. If the examiner's specificity increased to 99%, then the false inclusion rate would have to be 1%.

Test Specimen—An item that is submitted for forensic examination to test the performance of an examiner or a test method.

Valid/Validity—The OSAC Lexicon defines validity as "the extent to which a conclusion, inference or proposition is accurate." 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.

---

[10] In statistical hypothesis testing, the failure to reject the null hypothesis, when that hypothesis is false, is called a "Type 2 error." Most forensic science disciplines treat the hypothesis of "different source" as the null hypothesis. Consequently, a mistaken report that two items have a different source, when they have the same source (a false exclusion) is sometimes called a Type 2 error. However, in some disciplines (e.g., forensic glass comparison) the hypothesis of "same source" is treated as the null hypothesis, which means that (in those disciplines) a false inclusion (rather than a false exclusion) is a "Type 2 error." To prevent potential confusion, forensic scientists can avoid using the terms "Type 1 error" and "Type 2 error" when discussing methods for determining whether items have the same source. Instead, they can use the more transparent terms "false inclusion" and "false exclusion." If "Type 1 error" and "Type 2 error" are used, then the null hypothesis must be stated to avoid ambiguity.

[11] When the null hypothesis is that the items being compared have a different source, a mistaken report that the items have the same source (a false inclusion) constitutes an erroneous rejection of that null hypothesis. For that reason, false inclusions in forensic science are sometimes called "Type 1 errors." As explained in the previous footnote, however, when the null hypothesis is that the items have the same source, a "false inclusion" is no longer a "Type 1 error."

The validity of such a method can be assessed by testing whether it consistently produces accurate results when applied to test specimens of known source or type.

Validation— The OSAC Lexicon defines validation as: "A process of evaluating a system, method, or component, to determine that requirements for an intended use or application have been fulfilled." This document focuses on validation of methods for determination of source or type through empirical studies to determine their accuracy and limitations.

There are several types of validation. According to the OSAC Lexicon, **developmental validation** refers to "the acquisition of test data and determination of conditions and limitations of a new methodology; this generally occurs while the conditions and parameters are being worked out prior to the establishment of a defined assay, procedure or product." Developmental validation helps to establish the conditions under which a method is repeatable, reproducible, and accurate. According to the OSAC Lexicon, **internal validation** refers generally to "the accumulation of test data within the laboratory for developing the laboratory standard operating procedures and demonstrating that the established protocols for the technical steps of the test and for data interpretation perform as expected in the laboratory." Internal validation helps establish that the method has been applied *in practice* in a manner that produces accurate results.

White-Box Study—A white-box study is like a black-box study, but it allows assessment of the thought process or methodology of the examiner. For example, a researcher might observe an examiner and ask why a particular action was taken at each step in a procedure. Studies of this type allow the researcher to look inside the "black-box" and gain insight into how examiners make findings. In one such study, researchers sought to determine how various factors affect latent print examiners' judgments about the sufficiency of prints for comparison as a way of better understanding a key process entailed in their evaluation of latent prints (Ulery, Hicklin, Roberts & Buscaglia, 2014).

IV. Distinguishing Consistency from Accuracy

Consistency and accuracy are different dimensions of examiner performance. While the two dimensions are related, and both are worthy of careful study, it is important not to confuse one dimension with the other.

Why consistency isn't necessarily a good indicator of accuracy. It is possible to assess the consistency of results without knowing whether they are correct. For example, one might study whether different examiners reach the same finding when assessing whether two fingerprints have a common source. To assess the accuracy of those findings, however, the researcher must know the ground truth (i.e., whether the fingerprints were made by the same finger or not). Multiple examiners may agree yet still be wrong. Asking examiners to replicate each other's work (an appropriate approach in assessing inter-examiner reliability) is not a test of accuracy. By itself, it cannot establish that the method they are applying is valid.

Consistency is nevertheless important. Studying the consistency of examiners' judgments on casework samples can provide valuable information about laboratory and examiner performance. Although the exact causes of disagreements may be difficult to determine, inconsistent examiner judgments can reveal areas where improvements in performance are possible.

Consequently, consistency assessment can be a valuable part of quality assurance standards and guidelines. Laboratories can collect and retain data on how often disagreements occur when more than one examiner independently performs a comparison or examination of casework samples. Having different examiners occasionally replicate the same comparisons can demonstrate that findings are consistent (reliable) across examiners.

Retest programs are a way to assess laboratory performance when evaluating complex sample types that cannot be duplicated by mock samples, such as a controlled dangerous substance that has been injected or ingested. Retesting may be the best option for assessing laboratory performance in evaluating drug metabolites and other challenging samples. As the original sample is from a forensic case, the retest is testing the consistency of laboratory performance under true case testing conditions.

From a statistical perspective, the consistency of findings 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 reached the correct finding on only half of the items examined.

Possible causes of inconsistency. If examiners reach inconsistent findings, it may be important to determine the underlying reasons. Even if it is impossible to determine which examiners are right and which are wrong (because ground truth is not known), the existence of the disagreement may signal an underlying problem that needs to be addressed as part of quality assurance efforts. Possible causes to consider include the following:

- ***Inconsistent testing conditions:*** Factors such as environment, equipment, controls, and reference materials might be the source of the problem.
- ***Training deficiencies/mistakes:*** Inconsistency across examiners may signal a failure of one or more examiners to execute the forensic procedure correctly, indicating a need for additional (or better) training, or suggesting that the procedure itself is unclear or has not been explained in enough detail.
- ***Inconsistent decision thresholds:*** Examiners might disagree because they have different thresholds for making decisions. One examiner might require stronger or clearer evidence to reach a finding. Empirical studies with known-source test specimens can be useful both for determining whether examiners are applying inconsistent decision thresholds and for assessing which decision threshold is better for maximizing the accuracy of the procedure. For example, such studies might show that some examiners are being unduly conservative, judging samples unsuitable for comparison or judging comparisons inconclusive, when the samples could be correctly evaluated. Alternatively, research might show that some examiners are making errors by basing results on unsuitable samples, and thus that a more conservative approach is warranted. The kinds of data needed to make these evaluations will be described in a later section. These evaluations can be valuable for detecting mistakes, refining training procedures and helping examiners improve their skills.
- ***Operating beyond the limits of validity.*** If examiners are well-trained and are following the same method, discovering that they reach inconsistent results may call into question the validity of the method itself as applied in those instances. 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. A breakdown in the consistency of examiners' evaluations may signal that the examiners are working outside the range in which the method is valid and fit-for-purpose, or it may identify circumstances in which special caution is needed to avoid errors.

OSAC subcommittees can contribute to the continuing improvement of forensic science by developing standards and guidelines both for assessing consistency and for investigating and responding to evidence of inconsistency. The best approach is likely to vary across disciplines, so this is an issue for each subcommittee to consider.

Consistency of judgments of sample adequacy. Guidelines and standards also can be created for monitoring the consistency (inter-examiner reliability) of judgments about the suitability of forensic samples for testing or comparison. The discovery of wide variability in assessments would raise 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 consistency 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. In any case, active monitoring of the degree of consistency could be an important element of quality assurance and hence is an appropriate procedure to consider as standards are developed (Dror & Langenburg, 2019).

V. Key Issues in Designing, Conducting, and Reporting Validation Research

In this section we discuss a variety of issues that forensic scientists face when they design, conduct, and report studies on the accuracy of a method used for source determination or other types of classification. The information provided here should be helpful to OSAC members involved in the development of standards for validation and quality assurance as well as to forensic scientists more generally as they study the accuracy of their methods. For more details on creating a forensic-science validation study, see Martire and Kemp, 2018. For information about how and why to use Open Science practices when conducting such as study, see Chin, Ribeiro, & Rairden, 2019.

Preliminary Considerations

Issue 1: Institutional Review Board Review

Forensic scientists planning to do validation research should familiarize themselves with federal regulations regarding the treatment of research participants (human subjects). Under what is known as the "Common Rule," projects using federal monies for "research" involving human subjects must be reviewed and approved in advance by an Institutional Review Board (IRB).[12] The federal rules apply to research that is "conducted, supported, or otherwise subject to regulation by any Federal department or agency … ."[13]

[12] 45 CFR 46; 28 CFR 46.
[13] 45 CFR 46.101(a).

IRB review is also necessary if the research may result in the disclosure of information that could be used to identify individuals.[14]The requirement of IRB review does not apply to many data gathering activities conducted by forensic laboratories. Data collection for the purpose of quality control, quality assurance, personnel assessment or other administrative purposes generally does not constitute "research" within the meaning of the federal rules. "Research" is defined as a "systematic investigation … designed to develop or contribute to generalizable knowledge."[15] Hence, a federally-funded study designed to create generalizable knowledge about the accuracy of a pattern-matching method (e.g., a black-box study) would require IRB review unless it is exempt, but a test of the accuracy of pattern matching examiners conducted for the purpose of training or quality assurance (and without the intention to create a publishable finding) would not require IRB review.[16] IRB review has generally not been needed for proficiency testing and was not seen as necessary when blind test specimens were introduced into the flow of casework at the Houston Forensic Science Center.[17]

Even if a project constitutes research under the federal rules, it may fall under Exemption 3 of the Common Rule, which says that a research project need not go before an IRB if it involves only a "benign behavioral intervention" to which the subject prospectively agrees and "the information obtained is recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained, directly or through identifiers linked to the subjects."[18] There is also an exemption for secondary studies of existing data when the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.[19] This exemption may apply to studies in which data collected for administrative purposes is later compiled and analyzed in an effort to create generalizable knowledge (so long as the identity of the individuals involved is protected from disclosure).

To verify that a research study is indeed exempt from IRB review, universities and some funding agencies may require the researcher to consult with a Human Subjects Protections Office or a Certified IRB Professional. Even if an independent determination is not required by local rules, it is generally good practice to consult with knowledgeable professionals before deciding a project is exempt from review.

Researchers who need IRB review but work for agencies that do not have an IRB may be able to obtain an exemption determination, or a review, from a commercial IRB. Another option is to collaborate with a researcher, such as a university faculty member, who has access to an IRB.

---

[14] Researchers must avoid doing harm either to individuals who participate as research subjects (e.g., forensic scientists whose performance is being tested) or to individuals who provide test specimens (e.g., DNA; fingerprints) that might be used to identify them. These risks are generally managed by aggregating or otherwise de-identifying the data, and by withholding details about the test specimens that could allow identification of individuals. The procedures used to manage such risks should be reviewed and approved by an IRB.

[15] 45 CFR 46.102(l).

[16] If data on human performance that is collected for training and quality assurance purposes (without IRB review) is later compiled by a research who intends to create generalizable knowledge (e.g., through publication or presentation of the findings), the researcher who compiles the data should seek IRB review of the study.

[17] The Director of the Houston Forensic Science Center told the Human Factors Committee that he had not sought IRB review before initiating a blind testing program because he viewed the program as part of the laboratory's quality assurance effort.

[18] 45 CFR 46.104(d)(3).

[19] 45 CFR 46.104(d)(4).

Where multiple institutions or agencies are involved in a study, it usually is possible to arrange an IRB Authorization Agreement in which one IRB takes oversight responsibility for the research project.

Many organizations have rules in place regarding protection of human subjects. Researchers who are uncertain about the need for IRB review, or other legal requirements surrounding the protection of human subjects, should consult with legal counsel before beginning their research.

Issue 2: Study administration general issues.

Validation research should be conducted as objectively as possible. Experience has shown that the motives and interests of researchers can influence how they conduct studies (often unintentionally). To minimize this risk, steps should be taken to assure that the motives, desires, or perspectives of the research staff do not create biases in how the study is designed, how the study is conducted, or how the results are interpreted and reported. For example, 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 by, or at least informed by, such disinterested individuals.

Another step to consider in assuring the scientific rigor of the project is pre-registration of the research materials, design and analyses (see Chin, Ribeiro, & Rairden, 2019). Pre-registration is a relative recent trend among social scientists designed to assure the research is carried out in an open and transparent manner. It requires researchers to disclose publicly their research design, hypotheses, research materials, data analysis plans, and other aspects of the study before collecting data. It guards against certain practices that may undermine the credibility of a study, such as changing the proposed hypotheses or analytic methods after seeing the data, and partial or selective reporting of study findings.

It is also good practice for researchers and research participants to be "blind" to some aspects of the study. (This comprises a "double-blind" study.) Research staff who interact directly with research participants should 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 research participants from even subtle hints regarding the correct result.

We emphasize that the problem to be addressed is not that study administrators or laboratory administrators will *intentionally* release biasing information. Rather, it is the risk that research participants can be influenced unintentionally, even when the administrators are acting in good faith and appear to be doing nothing aimed at influencing test results. Our focus here is on keeping the research staff and participants blind to the expected results of the study. In a later section we discuss studies in which the research participants are blind to the existence of the study.

Creating and Selecting Test Specimens: Variety and Number

A key issue in validation is whether the test specimens adequately represent the range and difficulty of the items encountered in ordinary casework. If the study is designed to test the accuracy of a method for casework in general, then the samples should represent the full range and distribution of types and difficulty normally seen in casework. If the research is designed to test the accuracy of a method for a particular type of case (e.g., mixed DNA samples or 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.

When reporting results, researchers should be careful to disclose all that is known about the nature of the test specimens and how they were selected.

Issue 3: The source of test specimens: Created versus Casework

When possible, the test specimens used in a validation study should be specifically created, developed, or obtained for that purpose, so that the "ground truth" regarding their origin will be known with certainty. Because the true origin of casework samples generally cannot be known with certainty, the use of casework samples as test specimens for validation research raises concerns.

Casework samples may nevertheless be the best option when it is not practical to create suitable test specimens for which ground truth is known. Creation of suitable test specimens might be unethical or even illegal in some circumstances (e.g., toxicological test specimens from individuals exposed to controlled substances or poisons). In such instances, research on whether examiners evaluate casework samples consistently, and in a manner thought to be correct,[20] may be the only feasible method of validation. As already noted, casework samples can be very useful as test specimens for studying the consistency of examiners' judgment. In addition, injecting samples from completely unrelated cases into the flow of casework may be a useful way to study the rate of false inclusions in source determination tasks.

In its Views document on *Facilitating Research on Laboratory Performance*, the National Commission on Forensic Science (2016) commented on the need to develop sets of test specimens that could be used by multiple forensic laboratories to test the accuracy of their methods and called for governmental assistance in developing sets of test specimens for research purposes.[21]

---

[20] It is often possible to marshal evidence on the probable origin of casework samples, even if the true origin cannot be known with certainty.

[21] The National Commission explained its position as follows: "Development of suitable sets of research samples is a time-consuming and expensive task that will exceed the resources of many forensic laboratories. Laboratories may be able to cooperate to share this burden. Known-source latent print images, for example, could be prepared by one laboratory and shared electronically with other labs—creating efficiencies through inter-laboratory cooperation. Nevertheless, it is unrealistic to expect forensic laboratories themselves to bear the entire burden of creating such samples. Assistance from governmental agencies is needed.

"It is the view of the Commission that a government agency, such as the National Institute of Standards and Technology (NIST), should play a leading role in creating test sets for research on laboratory performance. This is a function that will be most efficient if handled in a centralized manner by an agency with expertise in testing. NIST has made valuable contributions to forensic DNA testing by providing mixed biological samples to laboratories for proficiency testing. In the view of the Commission, it would be desirable for NIST to expand its efforts in this arena to include the creation of test sets for other types of research on laboratory performance. NIST (and other government agencies) should also consider funding the creation of research test sets by private vendors and research organizations." (National Commission, 2016).

The recommendations of the National Commission may be helpful as OSAC subcommittees develop lists of research needs for their disciplines. More generally, inter-laboratory cooperation in the development of sets of known-source test specimens for research purposes can advance the field.

Issue 4: Evaluating test specimens regarding suitability and level of difficulty

A test specimen that originates from a known source may be unsuitable for analysis or may lack sufficient distinguishing features to allow it reliably to be identified or associated to the source through forensic analysis. Researchers who conduct validation studies often include a mix of test specimens of varying quality. Disagreement among examiners about what constitutes suitability and sufficiency can be an important part of such a study. In some instances, however, researcher may find it helpful to conduct "pre-assessment" to identify in advance test specimens that are sufficient in quantity and quality, so that they can focus on examiners' performance when evaluating test specimens that are known to be sufficient to permit identification or association. Both approaches have merit although they are designed to answer different questions.

Lawyers and judges often want to know "the error rate" of a forensic method or procedure (e.g., *Daubert v. Merrell Dow Pharmaceuticals, Inc*., 509 U.S. 579 (1993)). However, the error rate of a given procedure is likely to vary based on factors that affect the difficulty of the analysis in a particular case.[22] It is therefore helpful to assess the level of difficulty posed by the test specimens. In some disciplines, quantitative measures of sample complexity and sample quality already exist or are being developed.[23] These tools are useful for assessing the difficulty posed by test specimens and can help clarify the relevance of error rate data for forensic casework. 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.

OSAC subcommittees developing lists of research needs for their disciplines may wish to consider including research on how to assess in a rigorous manner the difficulty of the analytic results examiners must routinely reach.

Issue 5: Adequacy of sample size

- ***Numbers of Samples and Examiners:*** The accuracy of a method or procedure cannot be tested adequately with small numbers of test specimens, or with small numbers of examiners. Results obtained with small samples often vary greatly due to random factors. A larger sample of examiners will tend to better represent the underlying population of examiners.

---

[22] The AAAS report on latent fingerprint examination noted 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:

> …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)

[23] Researchers have made substantial progress in developing measures of the difficulty of latent print comparisons (Hicklin, Buscaglia, & Roberts, 2013; Kellman et al., 2014; Yoon et al., 2013). Researchers in other fields can work to develop such measures as well. These measures could then be incorporated into studies of examiner accuracy so that the implications of the findings will be better understood and easier to apply.

- ***Number of Test Specimens***. To properly assess the accuracy of a method at any difficulty level, it is important to include adequate numbers of test specimens at that level. To take account of sampling variability, researchers often report confidence intervals (or other interval estimates) for error-rate estimates. The confidence intervals will generally become tighter, and hence the error-rate estimates will be more precise, as the number of tests increases.

- ***Number of Examiners.*** Validation studies designed to establish the accuracy of a method should also involve multiple examiners. As noted above, there may be inconsistencies across examiners in the results they obtain using a method. 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.

Statisticians or other experts familiar with statistical power and sample size requirements for experimental research can assist in determining the appropriate numbers of test specimens and examiners for validation and reliability studies. This is particularly important if examiners will be evaluating different numbers or types of test specimens, as this can create unwanted correlations that will complicate statistical analysis of the findings.

It may take considerable time to collect enough data to make a meaningful assessment of the accuracy of a method in general. Data collection can be facilitated by enlisting assistance from multiple, cooperating laboratories, conducting multiple smaller studies, and seeking grant funding to offset the administrative and operational costs of the research. Validation research at multiple sites permits researchers take account of differences in local practices and culture and address concerns about variation in performance across sites.

Study Participants and Procedures

Issue 6: How to present the test specimens to study participant.

Test specimens should be presented to study participants in a way that avoids hints or clues as to the correct interpretation. It is important to avoid providing any suggestions as to the number or proportion of test specimens that will fall into categories. For example, if study participants know (or can infer) that half of the test specimens they see will be of one type, and half of another type, then the findings they reach about some of the specimens are likely to influence the findings they reach on others. Imagine, for example that a participant is asked to classify ten specimens as either Type A or Type B. If the participant knows that half of the specimens will be Type A, and is able to identify those Type A specimens, then the participant will know (without even doing an examination) that the other specimens must all be Type B. Hence, the study will not do a good job of assessing whether the participant can correctly identify Type B test specimens based solely on the physical characteristics of those specimens.

The most straightforward way to study how accurately examiners can classify items is to present a series of test specimens to them one at a time, asking them to decide about each test specimen before presenting the next one. For example, the researcher might present a series of footwear impressions to

an examiner, asking the examiner to determine the size and, if possible, the type of shoe that made each impression. The researcher should avoid providing information that would allow participants to draw any inferences about the probability of seeing shoes of a given size or type. When test specimens are presented in this manner, each determination can be regarded as an independent test of the examiner's accuracy.

Similarly, the most straightforward way to study how accurately examiners can make source determinations is to present test specimens as a series of pairs, asking the examiner to judge whether each pair of specimens has a common source, before presenting the next pair. Again, the researcher should avoid providing any information that might allow inferences or create expectations about the proportion of same-source and different source pairs that will be presented in the study, or about the probability that any particular pair will be same-source or different source.[24] Participants should be instructed to make their determination in the same manner they would for routine casework. The advantage of this approach is, again, that each determination can be regarded as an independent test of the examiner's accuracy.

OSAC members have asked whether it makes sense to present test specimens in groups or clusters. Those who favor this approach argue that samples are often presented this way in casework. For example, a firearms examiner might be asked to evaluate associations between a group of bullets or shell casings and particular firearms. The problem with presenting test specimens in this manner is that it may provide subtle cues to participants about the expected results. For example, if 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. It would be better for the instructions to state that the number of matches may range from 0 to *n*, where *n* is the number of questioned items. Even if the number of items that have the same source is not specified, however, the manner of presentation may create expectations that influence examiners' findings. For example, participants in the study may expect that at least one same-source test specimen will be presented and might rely on that expectation in making determinations (e.g., by inferring that the most similar pair presented must be a same-source pair).[25]

This concern does not mean that researchers should never present test specimens in groups or that research studies that use that approach are worthless. That approach requires special care to avoid any hint or suggestion about the number of test specimens within any group that are likely to be of a particular type or source. Researchers should also consider randomly varying the number of same-source and different source comparisons within groups to reduce any systematic effects of participants' expectations. Including an occasional "blank set" also is important.

[24] Researchers sometimes design black-box studies in such a way that half of the item-pairs presented to participants are same-source and half are different-source. Participants who are aware of this practice might use this information to help them make findings about the test specimens they are asked to evaluate, which would tend to undermine the validity of the study. To mitigate this problem, researchers should consider randomly introducing some variation among participants in the proportion of same-source and different source item pairs they see in the study and informing participants that this randomization will occur.

[25] The authors of the PCAST report were so concerned about this problem that they decided to ignore validation studies that employed "set-based analysis" in which examiners are asked to perform all pair-wise comparisons within or between small sets of test specimens (PCAST, 2016, pp. 106-107).

Experienced researchers can sometimes detect and find solutions to design problems that novice researchers miss. Novice researchers who set out to design validation studies can develop better studies and enhance their own research skills if they consult in advance with more experienced scientific colleagues. Most major universities have academic researchers with extensive experience designing research involving human participants, and these individuals are often willing to share their expertise freely.

Issue 7: Assuring that the method being tested for validity is followed by participants in the study.

Black-box studies assess error rates among a group of examiners who may differ in how they do their work. Although these studies are valuable for assessing error rates among practitioners, the lack of a standardized practice may make it difficult to determine exactly what "method" is being tested. If the goal of the study is to assess the validity of a method, then researchers should take additional steps to assure the method is fully characterized and is being followed:

Fully Characterize the Method: The researcher should first identify what steps and procedures the method entails. This must be done in enough detail to allow assessment of whether individual examiners are following the method in the intended manner, and in the same way as other examiners do.

Ensure That Examiners Use the Method Properly and Follow All Required Steps During the Validation Study. Researchers should check that the examiners, during the study, 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 study, monitoring the examiners during the study, providing aids like check lists to be sure that all proper procedures are being followed, and reviewing the performed procedures post-test to ensure than all required steps were taken.

Issue 8: Should some validation studies be conducted in a manner that leaves participants "blind" to the fact they are being studied—that is, should they not know that they are evaluating test specimens for a research study rather than ordinary casework?

Psychologists have long noted changes in the behavior of people who know they are being studied (Orne, 1962). People who know they are being studied may approach problems differently than they otherwise would. On categorical or classification tasks, they may consciously or unconsciously shift their thresholds for making decisions to produce more desirable outcomes (Paulhus,1991). Hence, error rates observed when people know (or can easily figure out) that they are being studied may not reflect error rates in ordinary practice.[26] People who know they are being studied may also make different judgments than they otherwise would regarding strength of evidence or might express different levels of confidence.

[26] 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 evaluated. This can be done by incorporating test specimens into the routine flow of casework in a manner that makes them indistinguishable from other items examined by the laboratory. Several authorities have urged that blinded studies of examiner accuracy be conducted as part of a broader effort to establish the range of validity of forensic science methods (National Commission on Forensic Science, 2016; AAAS, 2017, at pp. 47-51; PCAST, 2016, at p. 59). The added value of blind studies is that they can determine whether the level of performance observed in open studies is comparable to actual forensic practice.

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. Although elaborate simulations of this type are burdensome and expensive,[27] they have considerable scientific value and may be feasible in some settings.[28]

Blind studies and blind quality assurance programs are easier to conduct in laboratories that employ context management procedures to shield examiners from task-irrelevant contextual information.[29] In these laboratories there is a division of labor between bench-level examiners, who examine and interpret the physical evidence, and case managers, who communicate with submitting agencies and investigators (Mattijssen, Kerkhoff, Berger, Dror & Stoel, 2016). In such laboratories, it is easier for laboratory managers to insert research test specimens into the flow of casework in a manner that is undetectable because there is no need to involve personnel outside the laboratory. The case manager knows which items come from actual casework and which items are test specimens prepared for research, but (if care is taken) the examiners do not know. Blind testing programs of this type have been implemented in a few forensic laboratories (Kerkhoff et al., 2015; Kerkhoff et al., 2018).[30]

In blind studies, the test specimens and their presentation should give no hint of whether they are test specimens or routine evidentiary items. It is good practice to institute procedures for checking whether the supposedly blind samples are being detected as test specimens. Examiners are often 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.[31]

---

[27] See Peterson et al., 2003 (discussing a pilot test of blind testing of forensic DNA laboratories).

[28] Blind studies of this type have been conducted by the U.S. Army Defense Forensic Science Center.

[29] For background information on context management procedures, see Risinger, et al. 2002; Thompson, 2011; Found & Ganas, 2013; Stoel et al, 2015; Dror et al., 2015; Mattijssen et al., 2016).

[30] The Houston Forensic Science Center has been conducting blind testing in three areas (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 were implemented for a time by the document examination section of the Victoria Police Forensic Services Department in Australia.

[31] Information about the incentive program was provided to the Human Factors Committee by the Director of the Houston Forensic Science Center.

Blind studies will be most practical in disciplines in which:

- The items examined are relatively uniform
- Examiners typically evaluate a single item or a small number of items
- Examiners have little or no need for contextual information

In other circumstances, blind studies are more difficult, and retesting program may be a more viable alternative. Blind studies will also be impractical, obviously, for the developmental validation of new methods that have not yet been adopted in forensic laboratories. In such instances, other approaches to validation, such as non-blind black-box and white-box testing, may be the best available avenue for validation.

Analyzing Data and Reporting Results

Different disciplines and different laboratories collect different types of data and assess and report them differently. Below we deal with issues having to do with reporting results, first with categorical results (with few versus many categories) and then using likelihood ratios.

Issue 9: How to report the results of validation studies on methods used to reach categorical results

Methods Using only Inclusion/Exclusion (and Inconclusive) Judgments

When findings are reported categorically, validation studies are typically designed to produce data on rates of correct and incorrect categorizations. The simplest example is a method for source determination in which the practitioner reports either an inclusion (meaning the items being compared could have the same source) or exclusion (meaning the items could not have the same source). For such a method, there are two kinds of errors that the practitioner might make: reporting an inclusion (i.e., that two items have the same source) when they in fact have different sources (a false inclusion); and reporting an exclusion when the items in fact have the same source (a false exclusion).

Both kinds of errors (false inclusions and false exclusions) should be reported when presenting the results of a validation study. It is widely known and understood that efforts to decrease the number of false inclusions may increase the number of false exclusions, and vice-versa. We also realize that forensic scientists have generally been willing to tolerate more false exclusions in order to minimize the number of false inclusions. Nevertheless, it is important to measure and document both types of errors when assessing the range of validity of a method.

If other findings are possible, researchers should also report the rates of those other findings. For example, if study participants are allowed to reach a finding of "inconclusive" or to make the determination that a sample is not suitable for testing, then the rates at which they make those determinations should also be reported. These data help place information about the participants' performance in the proper context. Whether determinations of this type should ever be regarded as erroneous is a subject to ongoing discussion in the scientific literature (see, e.g., Dror & Langenburg, 2019), but in any event the rate of such findings should be reported.

Table 1 illustrates a way of presenting error rate data that takes account of participants' determinations of the suitability of items for comparison, or if they found the data inconclusive. The table 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, and bite marks.

The reporting format is like that used to report the FBI's black-box study of latent fingerprint examiners (Ulery et al., 2011). It shows how findings of a black-box validation study can be reported, in order to display all relevant error-rate data.

In this hypothetical study 100 examiners were each presented 20 different pairs of impressions, leading to 2,000 presentations. They were asked to determine whether each pair was made by the same item or different items. For each examiner, half of the pairs were made by same item, and half were 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. Table1shows that 300 of the presentations of same-source pairs and 100 of the presentations of different source pairs were found to be of no value. Examiners compared all presentations determined to be "of value" and reported their findings as either inclusion (same source), exclusion (different source), or inconclusive.

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)

	Source of Sample Pair							
Examiners' Finding	**Same Source**				**Different Source**			
	#	% PRES	% COMP	% CALLS	#	% PRES	% COMP	% CALLS
No value (not compared)	300	30			100	10		
Inconclusive	300	30	43		100	10	11	
Exclusion	40	4	6	10	790	79	88	99
Inclusion	360	36	51	90	10	1	1	1
Total Calls	400				800			
Total Comparisons	700				900			
Total Presentations	1000				1000			

The table breaks down the examiners' findings by the type of the pairs (same source, difference source). Accuracy can be measured in terms of the proportions of true exclusions (specificity) and true inclusions (sensitivity)—or their complements, false exclusions and false inclusions. The table highlights (in yellow) three percentages for same-source presentations in which examiners made false exclusions, and (in turquoise) three percentages for different-source presentations in which examiners made false inclusions. Specifically, the proportions for false inclusions and false exclusions are reported as: (1) a percentage of all presentations (% PRES); (2) a percentage of all comparisons, i.e., excluding those comparisons where the impressions were deemed to be of no value (% COMP); and (3) a percentage of all conclusive calls, i.e., excluding both "no value" and "inconclusive" comparisons, and including only cases where the examiner reached a conclusive result (% CALLS).

Presenting data in the tabular form shown here allows interested parties to easily see the differences in different error rates and to focus on whichever they deem most relevant.[32]

The data in this hypothetical study show a relatively low rate of false inclusions (about 1%) and a somewhat higher rate of false exclusions (4-10%, depending on how calculated). Accuracy estimates of this kind would clearly be helpful in assessing the validity of a forensic method for source determinations. For example, the higher rates of false exclusions than false inclusions 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 (an issue that could be addressed by blind studies).

Another finding of this hypothetical study 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. The latter could be important for understanding and improving examiners' decision-making process. In any event, by reporting validation data as illustrated in Table 1, researchers can display their findings to allow a fair and complete assessment of the accuracy proportions.

Methods Using Support Rating Scales, Confidence Ratings, or Type Classifications

When practitioners utilize a broader range of categorical findings, the design and reporting of validation research becomes more elaborate but can follow a similar format. Suppose that practitioners in a discipline decide to use a five-point scale for strength-of-evidence evaluations with the following categories: (1) strong support for same source; (2) moderate support for same source; (3) inconclusive or indeterminate; (4) moderate support for different source; (5) strong support for different source. Table 2 shows how data might be presented for a study designed to establish the range of validity of this method.

In this hypothetical study, 100 examiners were each presented 20 pairs of impressions and were asked to evaluate each pair using the five-point reporting scale. Table 2 shows the examiners' findings broken down by each of the five reporting categories, and also broken down by whether the sample pairs being compared were known to have the same source or a different source.

[32] Some commentators (e.g., PCAST, 2016) have recommended that forensic scientists compute one-sided confidence intervals around the third percentage and report that false inclusion rate together with the upper limit of its 95% confidence interval.

Table 2: Data from a Hypothetical Validation Experiment for a Source Determination Method with a Five-Point Reporting Scale

Examiners' Finding	Source of Sample Pair: Same Source #	Same Source % PRES	Same Source % COMP	Same Source % CALLS	Different Source #	Different Source % PRES	Different Source % COMP	Different Source % CALLS
No value (not compared)	300	30			100	10		
1. Strong support same source	400	40	57	80	10	1	1	1.25
2. Moderate support same source	90	9	13	18	40	4	4	5
3. Inconclusive	200	20	29	--	100	10	11	--
4. Moderate support different source	10	1	1	2	150	15	17	18.8
5. Strong support different source	0	0	0	0	600	60	67	75
Total Calls	500				800			
Total Comparisons	700				900			
Total Presentations	1000				1000			

As in Table 1, the percentages falling in each category 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" and "inconclusive" comparisons and including only comparisons on which the examiner reported a finding (% CALLS).

The same kind of table can be used to evaluate the accuracy of practitioners who report their source conclusions with varying levels of confidence. Suppose, for example, that examiners decide to use a seven-point scale to report their opinion about whether two items have the same or a different source, and the scale ranges from high confidence of same source (+3) to high confidence of a different source (-3). Researchers who study the accuracy of such a method will need to break down the number of responses falling in each of the seven response categories for known same-source and different-source items. The result will look like Table 2, with seven rather than five categories on the left side of the table. Researchers can then determine the rate of false positives (results that falsely point toward identification) and the rate of false negatives (results that falsely point toward exclusion) for each level of confidence. Studies of this type generally find lower rates of error for judgments made with higher confidence (e.g., Phillips et al. 2018, discussed further below).

The examples shown in Table 1 and Table 2 involve methods for evaluation of source determination. A similar approach can be taken to display data for a validation study of a method for classifying samples by type (e.g., determining the size or manufacturer of the shoe that produced a shoeprint, or the caliber of a bullet). In such a study, test specimens of known type would be presented to examiners, who would be asked to determine their type. The reporting table could break down the examiners' findings regarding type against the known types of the test- specimens. The table should also include data on the rates at which examiners found the test specimens unsuitable for analysis or reported the results as inconclusive.

What Should Be Reported

The accuracy of a method generally should not be reduced to a single percentage. It is misleading to say something like "the study showed that examiners were 98% 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 how many same-source and different source comparisons examiners make, the sensitivity and specificity of the method, and various other factors.

Suppose that a researcher conducted a study in which ten examiners were each asked to make source determinations about ten same-source item pairs. And suppose the examiners made no errors—that is, they correctly identified all the items as having the same source. The researcher might be tempted to report that the examiners were 100% accurate, with a false inclusion rate of zero, but there are two ways to be inaccurate, and this study only concerned with examiners' accuracy when comparing same-source items (sensitivity). It provides no information about their accuracy when comparing different-source items. Because examiners have no opportunity to compare different-source items, there is no way they could have made a false inclusion; therefore, the false inclusion rate was guaranteed to be zero. This is an extreme example, but it shows how misleading it can be to use a single percentage to characterize the accuracy of a forensic method, why it is necessary to distinguish same-source and different source comparisons, and why the sensitivity and specificity of the method must both be reported.

Combining the two proportions by averaging them does not solve this problem.[33] Other data analytic methods better characterize the accuracy of a forensic test method using a single number. A popular statistic in some fields is the AUC (Area Under the ROC Curve). Although the AUC has its own limitations, it is a more plausible way to compare the overall accuracy of different methods or groups. For example, (Phillips et al., 2018) used the AUC to compare the accuracy of trained facial examiners, reviewers, "super-recognizers," fingerprint examiners, students, and various statistical algorithms at a task that required them to determine whether pairs of photographs showed the same person or different people.[34]

This technical publication takes no position on how or whether error-rate data should be presented in case work reports or in courtroom testimony. The discussion is limited to the ways that researchers should report error rate data from research studies in scientific presentations and publications. There is debate about the extent to which the error rate found in any study reflects that error rate in practice or the probability of error in any case. Without joining into this debate, we simply note that collecting more and better data on the accuracy of forensic-science methods furthers our understanding of how well these methods can work and may suggest incremental improvements in them.

[33] The sample likelihood ratio (true-inclusion proportion divided by false-inclusion proportion) indicates how strongly, on average in the group studied, an inclusion supports the same-source hypothesis. A corresponding likelihood ratio can be computed for an exclusion. This combination of the statistics on the two type of accuracy expresses the diagnostic value of the average examiner's opinion better than either the false-inclusion proportion, the false exclusion proportion, or their average.

[34] Using data on the rates of false inclusions and false exclusions, Phillips et al. computed the AUC for each individual to conclude, among other things, that facial examiners and reviewers were statistically the same; that facial examiners and reviewers were more accurate than fingerprint examiners and students; and that the accuracy of the best automated face recognition system was comparable to the best human facial examiners.

Issue 10: Special problems in assessing the accuracy of likelihood ratios

A likelihood ratio (LR) is a statement about the relative probability of some observed data under two alternative hypotheses. A forensic scientist might state, for example, that the findings observed when comparing two items are 1,000 times more probable *if* the items have the same source than *if* the items have a different source. In this context, the LR can be viewed as a statement about the strength of the evidence for distinguishing the hypothesis that two items have the same source from the alternative hypothesis that the items have a different source (e.g., Robertson, Vignaux & Berger, 2016).

Because the LR is a continuous measure (ranging from zero to infinity), rather than a categorical source attribution or classification, assessing the accuracy of LRs for purposes of establishing the range of validity of a method poses special challenges. Helpful commentaries on this issue have been provided by Meuwly, Ramos & Haraksim (2017); Lund & Iyer (2017) and Morrison (2011). This issue has also been discussed in connection with the validation of probabilistic genotyping software (Bright, Evett, Taylor, Curran & Buckleton, 2015) and automated systems for forensic voice comparison (Morrison & Thompson, 2017). Validation studies on probabilistic genotyping systems offer excellent examples of methods that can be used to assess the accuracy of LRs (e.g., Bright, Richards, Kruijver et al., 2018; Moretti, Just, Kehl, et al., 2017). In this context, accuracy involves assessment of both the discriminating power of the LR—i.e., its ability to distinguish between the underlying hypotheses—and the calibration of the LR, which involves assessment of whether LRs produced by a method appropriately reflect the strength of the underlying evidence (*see* Meuwly et al., 2017).

Validation studies for methods that produce LRs are essentially the same as validation studies for categorical methods regarding experimental design and the choice and presentation of test-specimens. Again, test specimens of known source or type that are sufficient in number, representativeness, and diversity to provide a fair test of performance are necessary. For a source-determination task, that means that pairs of same-source items and pairs of different-source items are required. The key difference is that the outcomes are more easily presented and summarized for categorical judgments. In those situations, researchers can report the number or percentage of categorical findings for each value of the ground truth, as in Tables 1 and 2. When studying a LR method, researchers must record the LR that was reported for each comparison to form more complex sample distributions of LR values for the same-source and different-source comparisons.

A rough way to assess the difference between these distributions is to break them into somewhat arbitrary categories based on the value of the LRs, and then to compare the number of same-source and different-source comparisons falling in each category. This binning procedure throws away potentially valuable information, but it allows the researchers to analyze and report findings in simpler tables. For example, a researcher might decide to break the LRs into three categories based on whether each LR supports the same-source hypothesis, the different-source hypothesis, or is neutral (an LR value at or near one). Or, the researcher could decide to break the LR values into five categories analogous to the strong-moderate-neutral-moderate-strong scale in Table 2.[35]

[35] Let X be a LR value chosen by the researcher as the threshold for distinguishing moderate and strong evidence. Then the categories could be (1) strong support for same-source ($LR \geq X$); (2) moderate support for same source ($X > LR > 1$); (3) neutral or inconclusive ($LR = 1$); (4) moderate support for different source ($1 > LR > 1/X$) ; (5) strong support for different source ($LR \leq 1/X$).

The results then can be analyzed and reported as we discussed for these tables.

This categorizing of LRs seems particularly appropriate if forensic scientists themselves break LRs into categories when discussing the meaning of LRs in reports and testimony, though this practice is controversial. Under ENFSI Guidelines, for example, a LR may be "expressed by a verbal equivalent according to a scale of conclusions." (ENFSI, 2015, p. 16). ENFSI endorsed no particular scale, but it provided, "for illustration purposes only," seven categories for LRs. The Association of Forensic Service Providers (AFSP, 2009) proposed a similar categorical scale for use in explaining the value of LRs. In 2018, the Executive Board of SWGDAM authorized reporting a "qualitative statement that conveys the degree of support indicated by a likelihood ratio" in addition to the numerical value of the LR. Under the SWGDAM guidelines the qualitative statement must be taken from a five-point scale of verbal qualifiers like the AFSP scale (see, Recommendations of the SWGDAM Ad Hoc Working Group on Genotyping Results Reported as Likelihood Ratios, 2018).

More sophisticated procedures for assessing the accuracy of LRs are discussed by Meuwly et al. (2017) and Morrison (2011). Examples can be found in validation studies on probabilistic genotyping systems, such as Moretti et al. (2017) and Bright et al. (2018). These include Tippett plots (Tippett et al., 1968), the Log-Likelihood Ratio Cost, and Empirical Cross-Entropy. Graphical presentations may also be helpful. For example, researchers studying DNA mixture analysis have displayed the probability density of LRs, based on validation data from black-box studies, which allows display of the full range of observed LRs with known contributors and known non-contributors (e.g., Bright et al., 2018). Additional studies are needed to determine the extent of lay jury understanding of LRs, categorical scales and verbal predicates.

Disseminating Results

Issue 11: Sharing research findings in an open, transparent manner

A data collection plan should provide 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 to give those who conducted the study the first opportunity to publish. Once researchers publish their findings, however, they should freely share their study materials and data with academics, fellow researchers, and all other interested parties to the extent possible under IRB restrictions and privacy laws.[36]

In some instances, the privacy interests of those who participate in such studies, either as research subjects or by providing test specimens, may justify limiting the information released. Privacy interests usually can be addressed by releasing information in anonymized form—that is, by removing details that could be used to identify individuals. As noted in the discussion of IRB review, test specimens or raw data that might be used to identify individuals, such as fingerprints or DNA profiles, warrant special care.

[36] *See* Chin, Ribiero, & Reirdan, 2019 (regarding open science in forensic science); Nosek, et al. (2015) ("Transparency, openness, and reproducibility are readily recognized as vital features of science"); Mnookin, et al. (2011) ("A research culture, we argue, must be grounded in the values of empiricism, transparency, and a commitment to an ongoing critical process.").

The appropriateness of the measures taken to protect privacy is usually addressed during IRB review.

VI. Internal Validation and Quality Assurance

This technical publication has focused primarily on testing the performance of forensic examiners to assess the validity of commonly used methods, especially those that depend critically on human judgment. Validation studies of this type are often designed to assess the accuracy of a method in general, and hence do not need to be conducted in every laboratory.[37] Once published, studies demonstrating the circumstances under which a method produces accurate results can be relied upon by the entire discipline. In some disciplines, however, additional internal validation studies may be needed to assure that a method also works well in each laboratory.[38]

There are steps that could be taken in every laboratory to help verify that methods are being implemented in an appropriate manner and to assure that the laboratory is producing high quality results. As one laboratory director explained, there are three question about the accuracy of laboratory performance that need to be addressed by research: (1) Is there a valid method? (2) Are examiners applying that method properly to produce accurate results? And (3) as time passes, are they *still* applying the method properly in a manner that produces accurate result? The first question concerns developmental validation and can be addressed for the discipline.
The second question concern internal validation and needs to be addressed in each laboratory. The third question concerns ongoing quality control and needs to be addressed in each laboratory.

We have already discussed why it is important for researchers seeking to validate and improve methods to fully characterize the method being examined and to take steps to assure that study participants are following that method. Once a method has been validated, laboratory managers can take steps to assure that examiners are faithfully following that method. If examiners implement a method in idiosyncratic ways that deviate from what was validated, the findings of the validation research might not apply, and accuracy could suffer. Whether examiners are correctly applying validated methods is therefore a quality assurance issue worthy of careful consideration.

As discussed earlier, it is also important for laboratory managers to monitor the consistency (reliability) of examiners' judgments and to take steps to investigate evidence of inconsistency. Where consistency is a concern, laboratories occasionally can have an examiner re-examine an item seen previously or have items evaluated by different examiners to assess consistency (reliability). A retesting program also is a means of assessing reliability. A program of consistency checks can detect problems that might be missed by other important quality control procedures, such as Technical Review.

Laboratory managers also can collect data on the rates at which examiners reach different possible findings. For source determinations that are reported categorically, managers might collect data on the rates at which each examiner finds samples submitted for analysis to be suitable for comparison, and the rates of identifications, exclusions, and inconclusive findings. (In disciplines that employ

[37] Foundational studies of this type could be conducted cooperatively in a small number of laboratories, or government agencies. Such studies might also be conducted by a university or by an independent agency. Cooperation by laboratories and practitioners obviously is critical to the success of these efforts.

[38] In DNA analysis, for example, laboratories often have extensive internal validation protocols to establish laboratory-specific parameters (e.g., peak-height variance, stutter percentages) needed for interpretation.

more extensive reporting categories, data could be collected on the rates at which samples are placed in each reporting category.) Information of this type is useful because it may call attention to sub-optimal performance. For example, some examiners might be too cautious about reaching findings, or be failing to reach findings that could be reached with accuracy. Alternatively, some examiners may be overconfident, reaching inaccurate findings based on inadequate evidence. While there may be very good explanations for discrepancies across examiners in some cases—for example, more experienced examiners may be taking on harder cases—without tracking rates across examiners managers have no systematic basis for making such assessments.

Data on the rates at which examiners reach various findings also allow potentially valuable comparisons across laboratories. For example, if one laboratory found submitted samples of a type to be suitable for comparison at much lower rates than other laboratories, ascertaining the reasons for the disparity could be valuable. Is the discrepant laboratory dealing with more difficult cases? Do the examiners in that laboratory need additional training on how to evaluate difficult cases? Also, if laboratories are using different methods, this may be an opportunity for them to adopt the better performing methods.[39]

Finally, having examiners evaluate known-source test specimens is an important component of internal quality control testing programs. Thus, most laboratories do some known-source testing of individual examiners as part of training and proficiency testing. The National Commission on Forensic Science (in a Views document titled *Facilitating Research on Laboratory* Performance) called for the expansion of internal quality control testing, reasoning that testing programs can help assure the validity of analytic methods as applied by specific examiners in a particular laboratory (National Commission, 2016). These testing programs are also valuable for providing training and feedback to examiners, for estimating error rates for specific types of examinations or samples, and for quality assurance (National Commission, 2016).[40]

Regarding existing proficiency testing programs, the Commission noted that:

> Proficiency tests involve relatively few samples … and the tests are typically designed to be relatively easy for a competent analyst to pass. Hence, proficiency tests have limited value for establishing the limits of reliability and accuracy that analytic methods can be expected to achieve as the conditions of forensic evidence vary. These tests provide little useful feedback to forensic analysts on the limits of their expertise when dealing with difficult cases or marginal evidence and hence have little value for helping experienced analysts improve their skills (National Commission, 2016, p. 6)

The Commission called for additional government funding of research programs that involve testing examiners in the laboratory setting by having them process known-source test specimens, ideally by inserting the test specimens in the flow of casework in a manner that makes them indistinguishable from other samples—i.e., as a blind test. The Commission also emphasized the need to explore the

[39] Data collection efforts of the type discussed here are for analyzing the performance of the laboratory as a system, not for pressuring or sanctioning individual examiners (unless there is clear evidence that a particular examiner's work is deficient). When assessing consistency, managers must also take into account differences in laboratory protocols and procedures that may lead to inconsistent results, such as varying policy on when to call a given comparison inconclusive.

[40] The AAAS report on latent print examination agreed and called for ongoing testing of this type for latent print examiners (AAAS, 2017).

limitations of examiners' expertise with difficult samples.[41]

Whether such programs are practical given the resource limitations that laboratories currently face is a real question, and whether laboratories should be *required* to implement them is beyond the scope of this technical report. Our goal has been to discuss important facets of such programs from a human factors perspective. We close, however, with a more hortatory and policy-oriented note on the role of the research methods described here in forensic science and criminal justice system.

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

Why should forensic scientists conduct empirical studies to assess the accuracy of their methods? Validation is necessary in all scientific disciplines. It is particularly important in forensic science because of the consequences that may follow from 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 dramatic consequences for human lives—a fact that the forensic science and legal communities know and acknowledge. The manifest importance of forensic science findings to the justice system makes it vital to have data on their accuracy.

Which methods require empirical validation? With few exceptions, the analytic methods used for comparing items to determine whether they have a common source, or for classifying items into other categories, should be assessed for accuracy using test specimens of known source or type. Those methods will also need follow-up reliability checking to ensure that a method lives up to its potential as established through validity testing.

There will, of course, be debate about how extensive such research needs to be and about the best ways to conduct such research. We expect that many OSAC subcommittees will discuss these issues as they develop standards for validation and quality assurance. We hope that this document will enrich those discussions and help forensic scientists evaluate and decide those issues in an informed and thoughtful manner. And we hope that OSAC members will feel free to call on the Human Factors Task Group for assistance when we can be of help.

[41] "Studies that involve highly challenging samples will be particularly valuable for helping examiners improve their skills. For example, latent print examiners sometimes need to make critical decisions about whether a low-quality latent print (e.g., a print containing limited detail or distortions) can accurately be identified, or whether the comparison should be deemed inconclusive. Research on this issue will not only address general concerns about the reliability and accuracy of judgments in such cases but will also provide feedback that will help examiners improve their decision making in such cases" (National Commission, 2016, p. 5).

VIII. 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

Association of Forensic Science Providers. (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, 49(3), 161–164.

Bozeman, B. & Youtie, J. (2017). *The Strength in Numbers: The New Science of Team Science*. Princeton, N.J.: Princeton University Press.

Bright, J., Evett, I.W., Taylor, D., Curran, J.M., & Buckleton, J. (2015). A series of recommended tests when validating probabilistic DNA profile interpretation software. *Forensic Science International: Genetics*, 14: 125-131.

Bright, J., Richards, R. Kruijver, M. et al. (2018). Internal validation of STRmix—A multi laboratory response to PCAST. *Forensic Science International: Genetics*, 34: 11-24.

Chin, J. M., Ribeiro, G., & Rairden, A. (2019). Open forensic science. *Journal of Law and the Biosciences*, advanced access publication, 1-34. doi: 10.1093/jlb/lsz009

Connors, A.F., Dawson, N. V., Desbiens, N. A. et al. (1995). A controlled trial to improve care for seriously ill hospitalized patients: The study to understand prognoses and preferences for outcomes and risks of treatment (SUPPORT). *JAMA - Journal of the American Medical Association, 274*, # 20, 1591-1598.

Dror, I.E., & Langenburg, G. (2019). "Cannot decide": The fine line between appropriate inconclusive determinations versus unjustifiably deciding not to decide. *Journal of Forensic Sciences*, 64(1), Jan 2019. DOI: 10.1111/1556-4029.13854.

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.

European Network of Forensic Science Institutes. (2015). ENFSI Guideline for Evaluative Reporting in Forensic Science: Strengthening the Evaluation of Forensic Results across Europe. http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf

Forensic Science Regulator (2014). Guidance: Validation, FSR-G-201, Issue 1. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/375285/FSR-G-201_Validation_guidance_November_2014.pdf

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

Haned, H., Gil, P., Lohmueller, K., Inman, K., & Rudin, N. (2016). Validation of probabilistic genotyping software for use in forensic DNA casework: Definitions and illustrations. *Science & Justice*, 56:104-108.

Hicklin R. A., Buscaglia J., & 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.

Kerkhoff, W., Stoel, R. D., Mattijssen, E. J. A. T., Berger, C. E. H., Didden, F.W., & Kerstholt, J.H. (2018). A part-declared blind testing program in firearms examination. *Science & Justice*, 58: 258-63.

Lund, S.P. & Iyer, H. (2017). Likelihood ratio as weight of forensic evidence: A closer look. *Journal of Research of National Institute of Standards and Technology*, 122, No. 27: 1-32.

Martire, K. A., & Kemp, R. I. (2018). 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.

Meuwly, D., Ramos, D., & Haraksim, R. (2017). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International*, 276: 142- 153.

Mnookin, J.L., Cole, S.A., Dror, I.A., Fisher, B.J., Houck, M.M., Inman, K., Kaye, D.H., Koehler, J.J., Langenburg, G., Risinger, D.M., Rudin, N. Siegel, J., & Stoney, D.A. (2011). The need for a research culture in the forensic sciences. *UCLA Law Review*, 58: 725-779.

Moretti, T.R., Just, R.S., Kehl, S., Willis, L.E., Buckleton, J.S., Bright, J., Taylor, D.A. & Onorator, A.J. (2017). Internal validation of STRmix for the interpretation of single source and mixed DNA profiles. *Forensic Science International: Genetics*, 29: 126-144.

Morrison, G. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science and Justice*, 51: 91-98.

Morrison, G.S., & Thompson, W.C. (2017). Assessing the admissibility of a new generation of forensic voice comparison testimony. *Columbia Science & Technology Law Review*, 18: 326-433. http://www.stlr.org/download/volumes/volume18/morrisonThompson.pdf

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

National Research Council (2015). *Enhancing the Effectiveness of Team Science*. Washington, D.C.: National Academies Press.

Nosek, B.A., Alter, G. Banks, G.C., et al. (2015). Promoting an open research culture. *Science*, 348(6242): 1422-1425

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.

Phillips, P.J., Yates, A.N., Hu, Y. et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithems. Proceedings of the National Academy of Sciences, 115(24): 6171-6176. www.pnas.org/cgi/doi/10.1072/pnas.1721355115.

Powell, R.R. (2006). Evaluation research: An overview. *Library Trends*, 55(1): 102-120.

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

Ramos D., González-Rodríguez J., Zadora G., & Aitkin, C.G.G. (2013). Information-theoretical assessment of the performance of likelihood ratio computation methods. *Journal of Forensic Sciences*, 58: 1503–1518. http://dx.doi.org/10.1111/1556-4029.12233

Risinger, D.M., Saks, M.J., Thompson, W.C., & Rosenthal, R. (2002). The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. *California Law Review*, *90(1),* 1-56.

Robertson, B. Vignaux, G.A., & Berger, C.E.H. (2016). Interpreting Evidence: Evaluating Forensic Science in the Courtroom, 2d Ed. Chichester, U.K.: John Wiley & Sons.

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

Thompson, W.C. (2011). What role should investigative facts play in the evaluation of scientific evidence? *Australian Journal of Forensic Sciences. 43(2-3*): 123-134.

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

Tippett, C., Emerson V., Fereday, M., Lawton, F., Richardson, A., Jones, L., & Lampert, S (1968). The evidential value of the comparison of paint flakes from sources other than vehicles, Journal of the Forensic Science Society, 8(2): 61–65.

Ulery, B.T., Hicklin, R.A., Buscaglia J, &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.

Ulery, B.T, Hicklin, R.A., Roberts, M.A., & Buscaglia, J. (2014). Measuring what latent fingerprint examiners consider sufficient information for individualization determinations. *PLoS ONE* 9(11): e110179. doi:10.1371/journal.pone.0110179

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.