Inconclusive Decisions and Error Rates in Forensic Science

Henry Swofford, Steven Lund, Hari Iyer, John Butler, Johannes Soons, Robert Thompson, Vincent Desiderio, J.P. Jones, and Robert Ramotowski

National Institute of Standards and Technology

International Association for Identification
Annual Educational Conference
Reno, NV USA

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

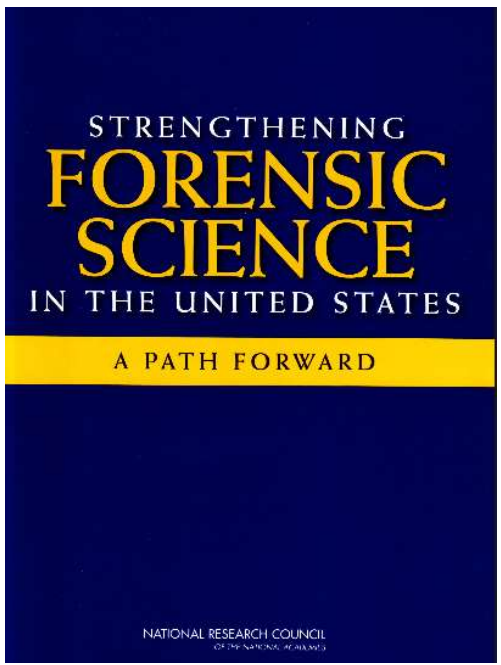
August 2024

Disclaimer

The opinions or assertions contained herein are views of the authors and do not necessarily reflect the views of the National Institute of Standards and Technology or the Department of Commerce.

Certain commercial entities might be identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that any of the entities identified are necessarily the best available for the purpose.

Background



REPORT TO THE PRESIDENT

Forensic Science in Criminal Courts:
Ensuring Scientific Validity
of Feature-Comparison Methods

Executive Office of the President
President's Council of Advisors on
Science and Technology

September 2016

Legal and the scientific communities have called for empirical evidence demonstrating the validity and reliability of forensic results.

Background

- Error rates (e.g., false positive or false negative rates) are satisfactory to represent performance when experts opine using a binary scale, such as “identification” or “exclusion”; however, few disciplines operate using a binary scale—most have “inconclusive” as an option.

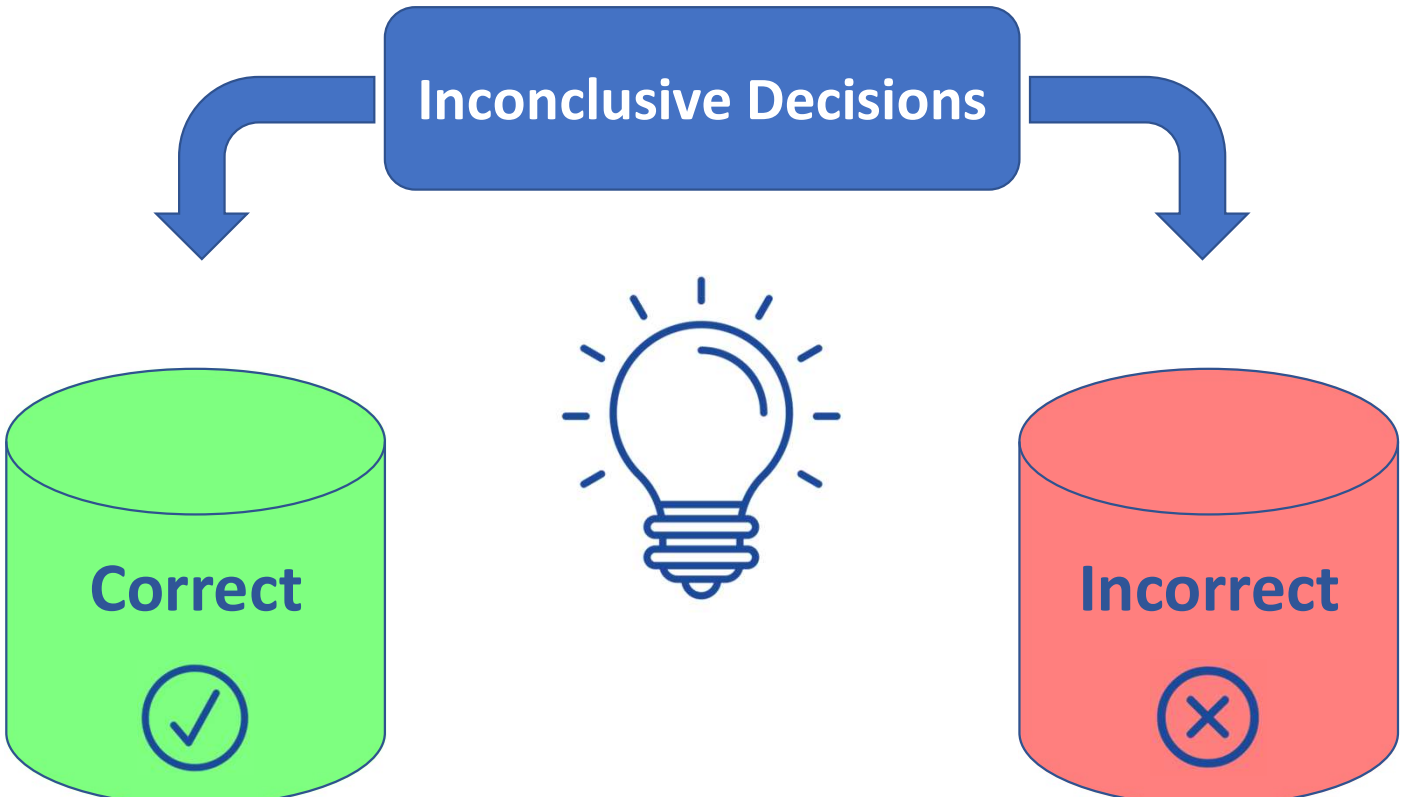
- When a conclusion scale is not binary, false positive and false negative rates alone are incomplete and can be misleading. Consider the following:

Method 1	Identification	Inconclusive	Exclusion
Mated Comparisons	0%	100%	0% ←
Non-mated Comparisons	→ 0%	100%	0%

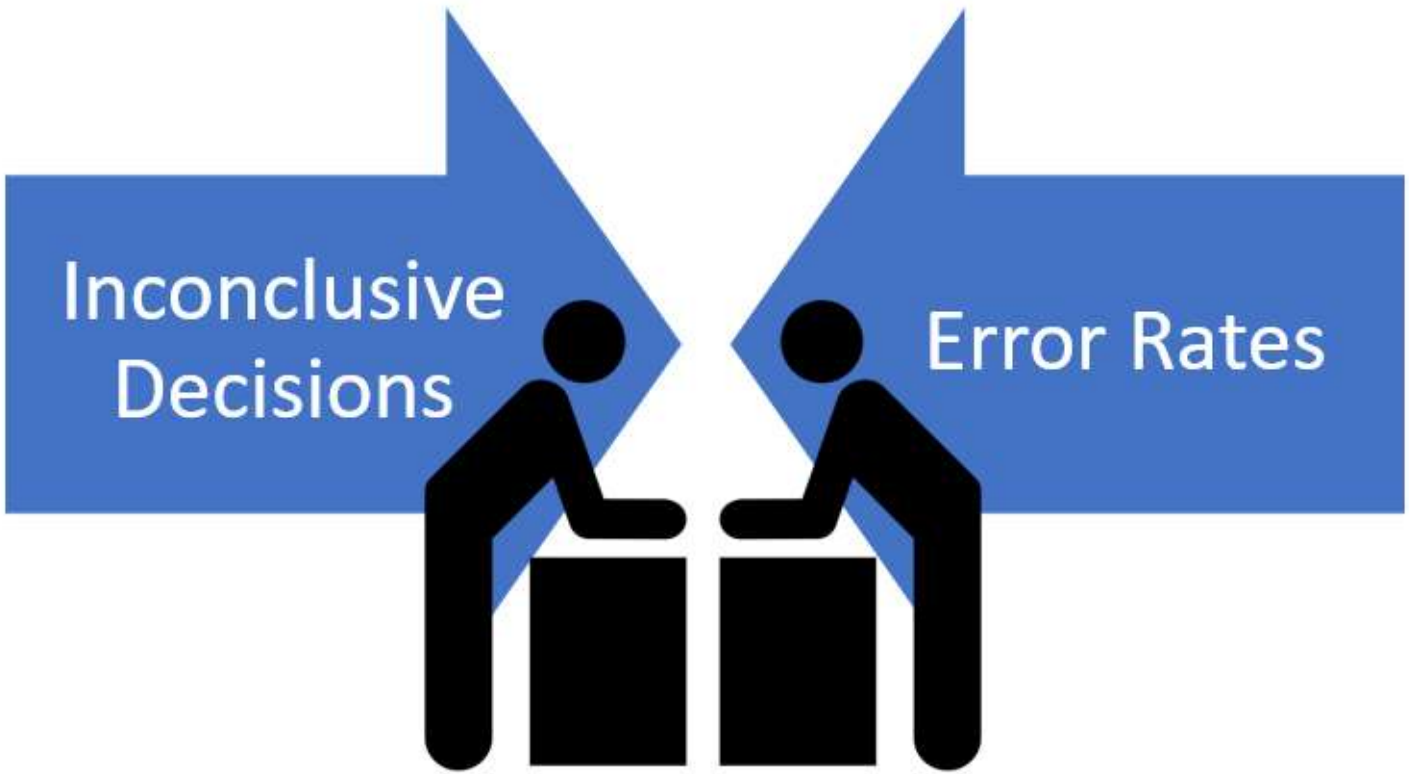
Method 2	Identification	Inconclusive	Exclusion
Mated Comparisons	100%	0%	0% ←
Non-mated Comparisons	→ 0%	0%	100%

- … Both methods have 0% error yet perform very differently!

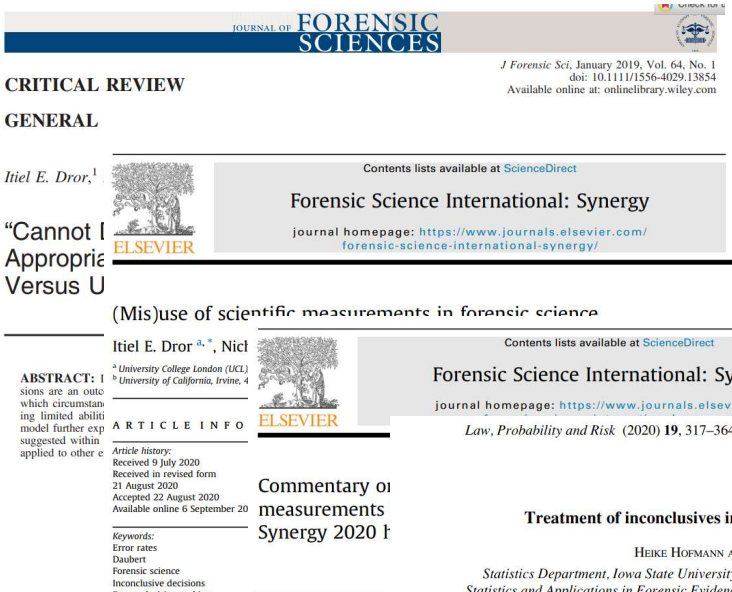
Background



Background



Background



JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, January 2019, Vol. 64, No. 1
doi: 10.1111/1556-4029.13854
Available online at: onlinelibrary.wiley.com

CRITICAL REVIEW

GENERAL

Itiel E. Dror,[1]

"Cannot I
Appropria
Versus U

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: https://www.journals.elsevier.com/forensic-science-international-synergy/

(Mis)use of scientific measurements in forensic science

Itiel E. Dror [a,*], Nicl

[a] University College London (UCL)
[b] University of California, Irvine, 4

ABSTRACT: I
sions are an outc
which circumstan
ing limited abiliti
model further exp
suggested within
applied to other e

ARTICLE INFO

Article history:
Received 9 July 2020
Received in revised form
21 August 2020
Accepted 22 August 2020
Available online 6 September 20

Keywords:
Error rates
Daubert
Forensic science
Inconclusive decisions
Expert decision making

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: https://www.journals.elsevier.com

Commentary o
measurements
Synergy 2020 h

Keywords:
Error rates
Daubert
Forensic science
Inconclusive
Expert decision marking
Study design

Law, Probability and Risk (2020) **19**, 317–364

Treatment of inconclusives in the AFTE range of conclusi

HEIKE HOFMANN AND ALICIA CARRIQUIRY

Statistics Department, Iowa State University, 2438 Osborne Dr, Ames, IA 50011
Statistics and Applications in Forensic Evidence (CSAFE), Iowa State University, 6
Ames, IA 50011, USA

AND

SUSAN VANDERPLAS[†]

Statistics Department, University of Nebraska Lincoln, 340 Hardin Hall North Wing, Lincoln, NE 68583-0963, USA

[Received on 9 September 2020; revised on 2 February 2021; accepted on 9 September 2020]

In the past decade, and in response to the recommendations set forth by the National Research Council Committee on Identifying the Needs of the Forensic Sciences Community (2009), scientists have conducted several black-box studies that attempt to estimate the error rates of firearm examiners. Most of these studies have resulted in vanishingly small error rates, and at least one of them (D. P. Baldwin, S. J. Bajic, M. Morris, and D. Zamzow. A Study of False-Positive and False-Negative Error Rates in Cartridge Case Comparisons. Technical report, Ames Lab IA,

Law, Probability and Risk (2021) **20**, 153–168
Advance Access publication on June 28, 2022

https://doi.org/10.1093/lpr/mgac005

Inconclusives and error rates in forensic science: a signal detection theory approach

Ohio Sta

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: https://www.journals.elsevier.com/forensic-science-international-synergy/

Forensic science and the principle of excluded middle: "Inconclusive" decisions and the structure of error rate studies

[Recei

Alex Biedermann

[a] University of Lausanne, Scho
[b] Northumbria University, Sch

There are
whether a
come fror
'inconclu
tion of we

ARTICLE INF

Article history:
Received 23 February 2021
Received in revised form
20 March 2021
Accepted 29 March 2021
Available online xxx

Keywords:
Inconclusive
Decision
Error rate
Principle of excluded middle

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: www.sciencedirect.com/journal/forensic-science-international-synergy

Inconclusives, errors, and error rates in forensic firearms analysis:Three statistical persp

Alan H. Dorfman [a,]

[a] National Center for Health Stati
[b] Universities of Michigan & Mar

ARTICLE INFO

Keywords:
Likelihood ratio
Virtual comparison microscopy
Nonresponse
Cognitive bias
Test-blind

PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES OPEN ACCESS

Validity of forensic cartridge-case comparisons

Max Guyll[a,1], Stephanie Madon[a], Yueran Yang[a], Kayla A. Burd[a], and Gary Wells[a]

Edited by Timothy Wilson, University of Virginia, Charlottesville, VA; received June 20, 2022; accepted March 6, 2023

This article presents key findings from a research project that evaluated the validity and probative value of cartridge-case comparisons under field-based conditions. Decisions provided by 228 trained firearm examiners across the US showed that forensic cartridge-case comparison is characterized by low error rates. However, inconclusive decisions constituted over one-fifth of all decisions rendered, complicating evaluation of the technique's ability to yield unambiguously correct decisions. Specifically, restricting evaluation to only the conclusive decisions of identification and elimination yielded true-positive and true-negative rates exceeding 99%, but incorporating inconclusives caused these values to drop to 93.4% and 63.5%, respectively. The asymmetric effect on the two rates occurred because inconclusive decisions were rendered six times more frequently for different-source than same-source comparisons. Considering probative value, which is a decision's usefulness for determining a comparison's ground-truth state, conclusive decisions predicted their corresponding ground-truth states with near

Significance

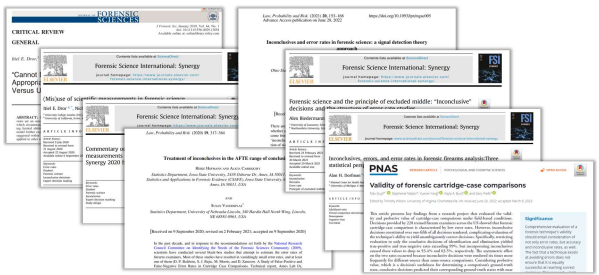
Comprehensive evaluation of a forensic technique's validity should entail consideration of not only error rates, but accuracy and inconclusive rates, as well. The fact that a technique excels at avoiding errors does not ensure that it is equally successful at reaching correct

Background

There is a desire to focus solely on error rates as a means of representing reliability. Consequently, several different perspectives and definitions for error rates have been proposed based on different treatments of inconclusive decisions, e.g.:

- Inconclusives should be ignored altogether
- Inconclusives should be considered always “correct”
- Inconclusives should be considered always “incorrect”
- Inconclusives should be considered sometimes “correct” and sometimes “incorrect”
- Inconclusives should be considered neither “correct” nor “incorrect”

Three Issues



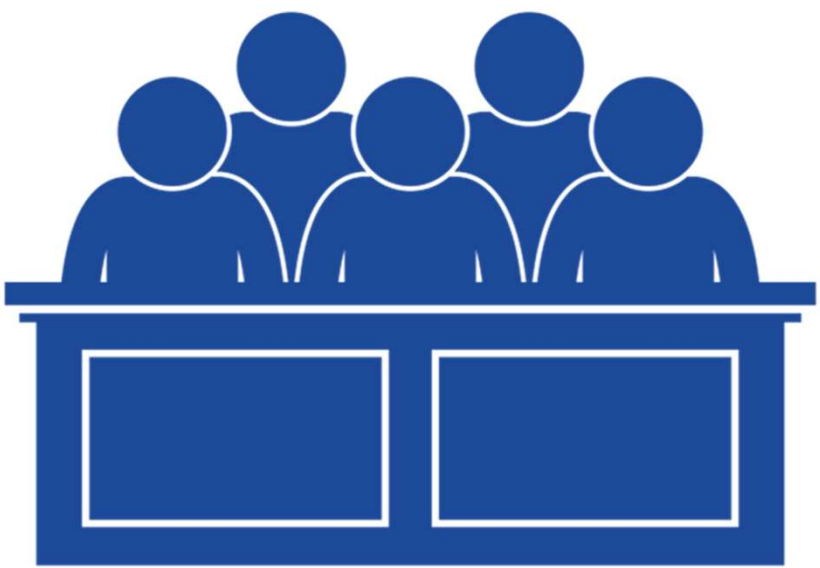
1. Error rates alone (i.e., false positive and false negative rates) have been used as the measure of method performance despite being unsuitable with non-binary conclusion frameworks.

2. Measures of reproducibility (or other factors that do not consider decision outcomes in relation to ground-truth) have been conflated with measures of discriminability (e.g., use of consensus opinion or decision rules to label results as “correct” or “incorrect”).

3. Assessments of method conformance have not been fully considered as a necessary factor for determinations of reliability for a particular case.

Context

Users of forensic results are presented with the outcome of an examination conducted by a particular analyst and tasked with discerning between two propositions of interest (e.g., two patterns were made by the same source). To properly interpret that result, three questions need to be considered:



(1) What method did the analyst apply when conducting the forensic examination?

(2) How effective is that method at discriminating between propositions of interest (i.e., mated vs. non-mated sources)?

(3) How relevant is the data reflecting the discriminability (i.e., diagnostic capacity) of that method (generally) to the examination in the case at hand (specifically)?

Two Concepts

Method *Conformance*

Relates to assessments of whether the outcome of a particular method is the result of the analyst's adherence to the procedure(s) that define that method.

Method *Performance*

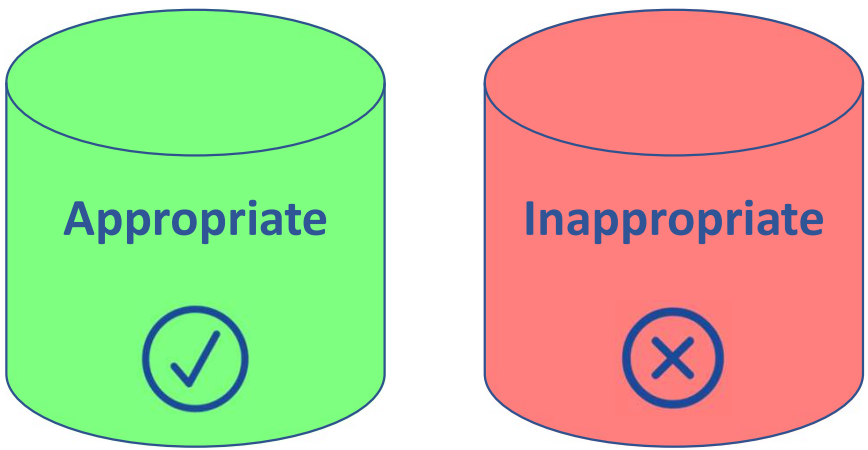
Relates to measures that reflect the extent to which the outcome of a particular method can effectively distinguish between different propositions of interest.

Method *Conformance*

Method *Conformance*

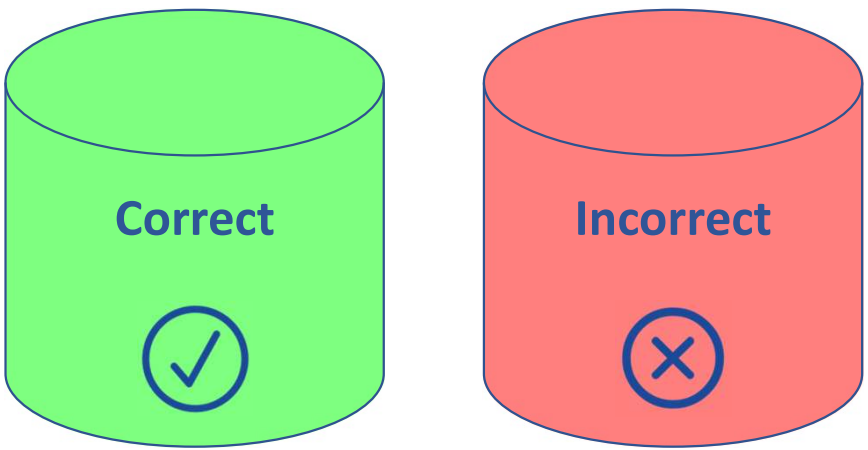
Relates to assessments of whether the outcome of a particular method is the result of the analyst's adherence to the procedure(s) that define that method.

- An "appropriate" decision is one that was produced by adhering to the established procedure.
- An "inappropriate" decision is one that was produced by deviating to the established procedure.



Method *Performance*

- A “correct” decision is one that accurately represents the true source-origin state of the items being compared.
- An “incorrect” decision is one that falsely represents the true source-origin state of the items being compared.



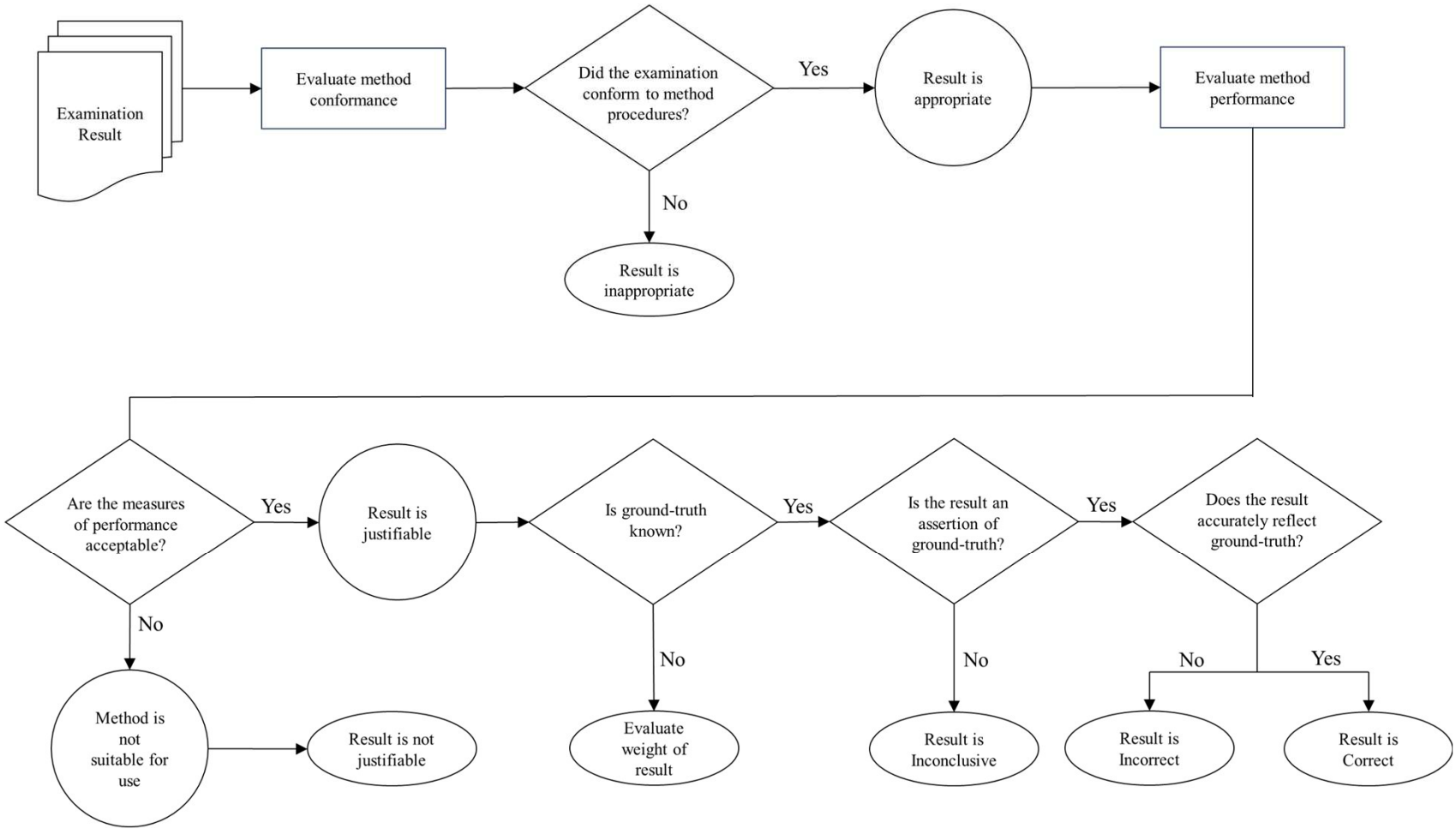
Method *Performance*

Relates to measures that reflect the extent to which the outcome of a particular method can effectively distinguish between different propositions of interest.

Inconclusive Decisions

- Inconclusive decisions are an important outcome of forensic examinations.
- Inconclusive decisions are not conducive to performance characterizations that require labeling each conclusion as “correct” or “incorrect.”
  - A “correct” decision is one that accurately represents the true source-origin state of the items being compared.
  - An “incorrect” decision is one that falsely represents the true source-origin state of the items being compared.
- An inconclusive decision is an outcome of the examination for which a conclusive assertion about the source-origin of the items being compared was not made; thus, inconclusive decisions are **neither “correct” nor “incorrect”** in the context of measuring performance.
- *Any* outcome of an examination (including inconclusive decisions) might be **“appropriate” or “inappropriate”** in the context of assessing conformance depending on whether the decision was produced as a result of adhering to or deviating from established procedures, decision criteria, or conditions for which the method has been deemed acceptable.

Evaluation of Results



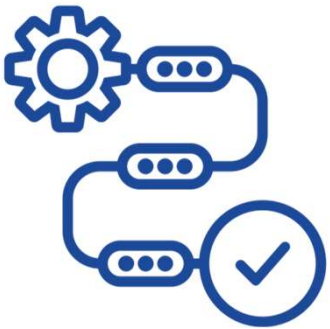
Implications to Practice

1. Reliability Determinations
2. Method Validation
3. Reporting Results
4. Performance Monitoring

#1: Reliability Determinations

- Determinations of the reliability of analysts' examination results require consideration of those results in the contexts of both method conformance *and* method performance – a result alone is not sufficient for one to assess its reliability.
- Performance data is *only* relevant to applications of the same step-by-step procedures (i.e., same method).
- Deviating from procedures does not mean the non-conforming analyst performed better or worse than those who did conform to procedures. … ***However …***
  - Performance data from other analysts who did conform to procedures (such as during validation studies) might not adequately reflect the performance of the non-conforming analyst for the examination in question.
  - There might be little to no information with which to assess the reliability of the outcome produced by the non-conforming analyst.

#2: Method Validation



- Method validation is the process of *verifying* that a particular method can be properly applied and produce results that achieve the required performance specified for its intended use.
    - Whether a method is suitable for use in a given case depends on whether enough data exists to characterize its performance in such cases and, if so, whether that performance is acceptable for use.
- Studies that purport to characterize the performance of a particular method (i.e., validation studies) are only relevant if conformance to that method can be demonstrated.
- Forensic service providers must have well documented and detailed step-by-step procedures that define their methods so that conformance can be assessed.
- Forensic service providers that do not have well documented and detailed step-by-step procedures that define their method, including relevant decision criteria that establish the conditions for which the application of the method and different outcomes are appropriate, are unlikely to be able to meaningfully support a claim that the outcome of their examination is the product of a valid and reliable method.

#3: Reporting Results

- Reporting results is the process of communicating a particular outcome of a method to users of that information to enable them to make inferences and decisions about the truth of various propositions in question.
- To properly interpret a result from a forensic examination, users need:

- Assurance that the outcome is an appropriate application of a method (i.e., method conformance);

- Information about the performance of that method (i.e., validation data) to understand the extent to which the examination result is predictive of the true source-origin state of the compared items under conditions relevant to the examination at hand (e.g., to consider the "predictive value" of the result by assessing the likelihood ratio of the decision).

Method	Identification	Inconclusive	Exclusion
Mated Comparisons	75%	25%	<1%
Non-mated Comparisons	<1%	50%	50%

- *This is particularly important for inconclusive decisions that might not be symmetrically distributed between mated and non-mated comparisons.*

#4: Performance Monitoring

- Performance monitoring activities include assessments of method conformance, measures of method performance, or both for a *particular* method or an aggregate of multiple methods within or across laboratories (e.g., through intralaboratory testing, proficiency testing, interlaboratory comparisons, or black-box studies).

- Aggregate measures of performance provide important information about a discipline overall but do not necessarily constitute as a validation or generalizable performance characteristics for any *particular* method unless it can be shown that the same method was used by all participants.

- The development and use of standard methods help reduce variability and ensure aggregate measures of performance can be used to support validation while reducing resource burdens that would otherwise be placed on laboratories to accomplish this independently.

Key Takeaways

- Laboratories should have well-documented and detailed step-by-step procedures that define their method, including conditions for method application and decision criteria for results.

- Laboratories should have a means for empirically demonstrating conformance of analysts' adherence to method procedures.
  - NOTE: Demonstrating consistency of outcomes (e.g., through verification or separate examinations of the same evidence) is not sufficient to serve as a basis for assessing or demonstrating conformance to a method or labeling a result as "appropriate."

- Laboratories should have data demonstrating the performance of their methods (e.g., validation data) which measure discriminability (i.e., 2x3 table) and reproducibility (i.e., 3x3 table) that reflect how often the outcomes produce the correct result and how often the outcomes are consistent when the method is applied by different analysts for the same items.

- Laboratories should include information in their reports (or casefile) that allow recipients to properly evaluate the weight of the result (i.e., information about conformance to the method and performance of the method).

Henry Swofford, Ph.D.
Lead Scientist
Forensic Science Research Program
Special Programs Office
National Institute of Standards and Technology
Henry.Swofford@nist.gov

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Act and Make an Impact!

Participate in OSAC's 2024 Registry Implementation Open Enrollment Event

July 1 Through September 2

September 15 Through 21

OSAC
Organization of Scientific Area Committees for Forensic Science

Visit booth #322 for more info!