JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, January 2014, Vol. 59, No. 1
doi: 10.1111/1556-4029.12275
Available online at: onlinelibrary.wiley.com

PAPER

GENERAL

Angi M. Christensen,[1] *Ph.D.; Christian M. Crowder,*[2] *Ph.D.; Stephen D. Ousley,*[3] *Ph.D.; and Max M. Houck,*[4] *Ph.D.*

Error and its Meaning in Forensic Science*

ABSTRACT: The discussion of "error" has gained momentum in forensic science in the wake of the *Daubert* guidelines and has intensified with the National Academy of Sciences' Report. Error has many different meanings, and too often, forensic practitioners themselves as well as the courts misunderstand scientific error and statistical error rates, often confusing them with practitioner error (or *mistakes*). Here, we present an overview of these concepts as they pertain to forensic science applications, discussing the difference between practitioner error (including mistakes), instrument error, statistical error, and method error. We urge forensic practitioners to ensure that potential sources of error and method limitations are understood and clearly communicated and advocate that the legal community be informed regarding the differences between interobserver errors, uncertainty, variation, and mistakes.

KEYWORDS: forensic science, error, limitation, forensic anthropology, *Daubert*, mistake

Discussion regarding "error" in forensic science analyses gained momentum following the *Daubert* ruling (1) and has intensified with the National Academy of Sciences' National Research Council Report *Strengthening Forensic Science in the United States: A Path Forward* (2). The role of science within the judicial system is nothing novel; however, the focus has shifted to include the evaluation of methods and techniques rather than simply the expert's interpretation of the results. Establishing scientific validity is challenging within the forensic sciences considering that the concept of error has different meanings and functions in the courtroom compared with the research setting (3). Estimating method validity and understanding error are important, however, regardless of whether conclusions end up in court.

The concept of error has been problematic, and too often, the courts as well as forensic practitioners misunderstand the meaning of error as it relates to forensic science research, procedures, and techniques. Error can be defined in a number of ways including the following: an act, assertion, or belief that unintentionally deviates from what is correct, right, or true; the condition of having incorrect or false knowledge; the act or an instance of deviating from an accepted code of behavior; or a mistake. Mathematically and statistically, error may refer to the difference between a computed or measured value and a true or theoretically correct value.

Considering these definitions, it is apparent that error in the forensic science realm can result from a number of different causes, contributing to the complexity in understanding the potential source(s) of error. The convergence of science and law has made the identification and interpretation of error in the courtroom an even greater challenge, especially as officers of the court typically lack a scientific background and specific knowledge regarding error analysis. Thus, the concept of error is often vague and subject to a variety of interpretations.

Admissibility criteria for expert testimony in the United States were redefined in the 1993 Supreme Court *Daubert* decision (1) resulting in significant changes in and interpretation of the Federal Rules of Evidence Rule 702 (4). The *Daubert* criteria were intended to provide guidelines for admitting scientific expert testimony to ensure its reliability and validity. In federal cases and in states that have adopted the *Daubert* criteria, trial judges might consider the following factors to assess the admissibility of scientific or technical expert testimony: (i) whether the theory or technique in question can be (and has been) scientifically tested, (ii) whether it has been subjected to peer review and publication, (iii) its known or potential error rate, (iv) the existence and maintenance of standards controlling its operation, and (v) whether it has attracted widespread acceptance within a relevant scientific community (1:593-94).

While the tumult surrounding the potential impact of the *Daubert* ruling on the forensic sciences seemingly began to dissipate over the years, the challenge to the forensic science community was renewed with the release of the National Academy of Sciences' National Research Council Report *Strengthening Forensic Science in the United States: A Path Forward* (2). This document outlined the scientific and technical challenges that must be met in order for the forensic science enterprise in the United States to operate at its full potential. In the Council's opinion, some disciplines were found to lack scientific rigor, leading

[1]George Mason University, Fairfax, VA.
[2]Office of Chief Medical Examiner, New York City, NY.
[3]Mercyhurst University, Erie, PA.
[4]Department of Forensic Sciences, Consolidated Forensic Laboratory, Washington, DC.

*Presented at the 63rd Annual Meeting of the American Academy of Forensic Sciences, February 20–26, 2011, in Chicago, IL. The research presented in this manuscript was not conducted under the auspices of the New York City Office of Chief Medical Examiner (NYC-OCME) or the Department of Forensic Sciences (DFS). The opinions expressed herein are those of the authors and do not reflect the opinions of the NYC-OCME or the DFS.

Received 2 June 2012; and in revised form 26 Sept. 2012; accepted 27 Oct. 2012.

to recommendations that emphasized the need for increased and improved research. In particular, Recommendation three states that research is needed to establish the validity of forensic methods, develop and establish quantifiable measures of the reliability and accuracy of forensic analyses, and develop quantifiable measures of uncertainty in the conclusions of forensic analyses.

While the NAS report specifically calls attention to the lack of scientific testing and development of known error rates for many forensic methods, more generally, the report reminds us that as scientists *we must do science well*. Furthermore, as forensic practitioners, we must be cognizant of the concerns of the legal community. This includes understanding how the Courts view and evaluate scientific evidence, being proactive in educating the legal community about the scientific process, and being prepared to mitigate misinterpretation and misunderstanding of scientific results.

Understanding "Error"

Prior to the *Daubert* decision, method reliability and validity were not specifically required to be considered; the admissibility of scientific evidence was a matter of the general acceptance test under *Frye* (5). In the *Daubert* decision, "reliability" was used repeatedly to mean "dependability," which incorporates both scientific reliability and validity. "Reliability" in the scientific sense is often expressed in how different observers measure or score the same phenomenon differently, and there are a variety of statistical measures to quantify reliability depending on the type of observation (6). Highly reliable observations show very low or no interobserver variability and high repeatability. Reliability, however, is not sufficient for establishing validity.

Validity can best be thought of as the overall probability of reaching the correct conclusion, given a specific method and data. Methods that are considered "valid" give us the correct conclusion more often than chance, but some have higher validity than others and will give us the wrong answers less often. In this sense, "validity" is clearly what the Court had in mind when it emphasized "reliability." Indeed, since the *Daubert* decision, questionable method reliability/validity has been the most frequently cited reason for excluding or limiting testimony on forensic identification sciences (7).

While there are many aspects of error that will influence validity, the known rate of error provides a scientific measure of a method's validity and that is likely why it was incorporated as part of the *Daubert* guidelines. Of course, error rates are not known, but estimated; the *potential* error rate of any method is 100%. The error rate guideline, however, has often created more confusion than clarification. We attempt here to simplify the discussion by describing the differences between several generally recognized potential sources of error: practitioner error, instrument error, statistical error, and method error.

Practitioner error refers to a mistake or operator (human) error. It may be random or systematic, may be related to negligence or incompetence, and is, for the most part, unintentional and unquantifiable (8,9). Practitioner error may refer to blunders such as transposing numbers when recording data, incorrect instrument use, selection of inappropriate methods, or improper method application. Practitioner error may also be intentional, such as fraudulent behavior. While practitioner error is certainly a concern of the courts, it is not error in the scientific sense. Practitioner error is exceedingly difficult to estimate but can be reduced through quality assurance systems, training, proficiency testing, peer review, and adhering to validated protocols and discipline best practices.

Instrument (or technological) error can be defined as the difference between an indicated instrument value and the actual (true) value. Instruments should be calibrated against a standard (i.e., a certified reference material), but even when properly calibrated, they typically have a prescribed, acceptable amount of error which has been determined by the instrument's manufacturer. Instrument error is measured in various ways statistically and can be minimized by proper maintenance and calibration of instruments as a part of a laboratory quality assurance program, but some acceptable amount of error is understood and therefore recognized to exist.

Statistical error is the deviation between actual and predicted values, generally estimated by the standard error or other measure of uncertainty in prediction, for example when a prediction interval with an explicit probability is specified. Statistical error often merely expresses normal variability and is inherent in measurements and estimates because they are based on the properties of a sample. It is possible that the actual value of a measurement or estimate may fall outside of the prediction interval.

Lastly, method (or technique) error relates to inherent limitations that have nothing to do with practitioner error or breakdowns in technology (8–10). Method error is often a function of how measurements or traits overlap among different groups or to the frequency of the observed trait(s) in the population at large. While these limitations are not themselves "errors," they affect the sensitivity or resolving power, probative value, and ultimately validity of the method. The more rare that a trait or suite of traits is in a population, the more sensitive that method is for associating the trait(s) to a particular individual, item or group. For example, nuclear DNA has greater resolving power for determining identity than mtDNA because the same mtDNA occurs more frequently in the population at large. The pelvis has greater resolving power than the skull in determining sex from skeletal remains because there is greater overlap between the sexes in features of the skull (because the pelvis is more sexually dimorphic than the skull).

Estimations of method errors are the most familiar. A frequentist estimation of an error rate is based on previous results alone: a 99% valid method, for example, would have a 1% error rate in the long run. A Bayesian estimation of an error rate is based on previous results and the specific observations or data being analyzed, so the estimated error rate applies only to the case at hand. There is no way to minimize method error (with, for example, additional training or calibration)—it simply exists as a function of inherent variation in the material itself. Such limitations, however, should be acknowledged and communicated in reports and testimony.

The estimated (known or potential) rate of error to which the *Daubert* guidelines refer can include a number of things such as the confidence interval, the statistical significance of a result, or the probability that a reported conclusion is incorrect. This may involve any or a combination of the factors discussed above, but most often largely involves statistical error and method error. The selection of good research designs and appropriate statistical models is imperative to produce valid scientific methods with low estimated rates of error.

The importance of implementing measures to minimize and account for error and limitations in forensic sciences methods should now be apparent. Understanding and appropriately communicating these issues can resolve confusion over the significance of results and can prevent intentional or unintentional misuse of error.

Misunderstanding "Error"

The misuse of error has serious ramifications in both legal and scientific communities. We have identified three sources regarding the misuse of error: (i) claiming a "zero error rate" for a technique, (ii) claiming that an error rate cannot be estimated, and (iii) attempting to calculate error rates *post facto* from activities that were not intended for that purpose such as proficiency tests, professional exercises, or other studies. Reasons behind these misuses range from not understanding the meaning of an "error rate," to improper training in statistics and the scientific method, to concerns that current methods will be exposed as lacking an empirical basis.

Some forensic practitioners have claimed that the error rate for their technique or method is zero (see (11) for one example). For example, the following testimony was provided by a fingerprint examiner in *People v Gomez* (12) explaining the reasoning behind the zero error rate claim:

> And we profess as fingerprint examiners that the rate of error is zero. And the reason we make that bold statement is because we know based on 100 years of research that everybody's fingerprints are unique, and in nature it is never going to repeat itself again.

The fallacy in the expert's reasoning is two-fold. First, the notion of uniqueness in forensic science is probabilistic and impossible to prove in a scientific sense, and this form of logic follows inductive reasoning (13). Second, this practitioner fails to understand that despite the strength of the basis for fingerprint association (that there is a low probability for two identical fingerprint patterns to exist), error, or limitations may still exist in the comparison methodology. Error rates relate both to the *frequency* of a particular trait(s), as well as to how *accurate* the methods of comparison are in determining an association or exclusion. Even if a feature is "unique," it does not mean that comparison methods can infallibly determine whether two samples originated from the same source. Such reasoning often results from the belief that humans may err but forensic techniques do not; a suggestion that is both erroneous and unfalsifiable (10). Admittedly, in disciplines (such as fingerprint comparison) where the method is primarily the judgment of the examiner, it can be impossible to disentangle method error from practitioner errors (10). Even so, there is always a nonzero probability of error, and to claim an error rate of zero is inherently unscientific.

Alternatively, some practitioners claim that an error rate simply cannot be estimated. This misguided philosophy likely results from a lack of proper testing to determine what the known method limitations or potential rate of error may be, insufficient statistical training of practitioners, or misunderstanding the meaning of error altogether. For forensic practitioners that make this claim, there is likely a fear that by acknowledging method limitations and potential error rates, the power of the analysis in the courtroom would be diminished. Other tactics involve carefully selected language to avoid the issue of error. In reality, however, if a method can be applied, error may exist and should be acknowledged.

In the absence of known or potential error rates for a method, it is not acceptable to derive error rates from practitioner proficiency tests, professional exercises, or studies that were not designed to estimate method error rates. Proficiency tests are typically designed to monitor performance and demonstrate that personnel produce reliable work and that analytical procedures are conducted within established criteria. Exercises designed to explore the ability of practitioners to perform certain analyses may not qualify as a true proficiency test, and caution should be exercised when results of such studies are employed outside of their intended use.

As one example, the American Board of Forensic Odontology (ABFO) implemented a study developed as an "educational exercise whose primary purpose was designed to survey the degree of agreement (or disagreement) between [board-certified ABFO] diplomates confronted with cases of varying amounts and quality of bitemark evidence" (14). Using data from this exercise, a forensic odontologist reported the following in an affidavit for a 2002 Supreme Court of Mississippi trial (15):

> On average, 63.5% of the examiners committed false positive errors across the test cases. If this reflects their performance in actual cases, then inculpatory opinions by forensic dentists are more likely to be wrong than right.

This claim misrepresents both proficiency testing and error as the results of the study do not provide a true measure of either. Neither the accuracy of the method nor the reliability of its application is illuminated from this study, yet the results have been extrapolated into an "error rate" for an entire method. The legal and scientific communities are left not knowing any more about the method's validity or reliability despite—or perhaps because of—interpretations like this.

Another example of incorrectly extrapolating an error rate is the use of results from a 2001 paper by Houck and Budowle (16). Although a novel study of the concordance between phenotype and genotype in human hairs by two forensic methods (microscopy and mtDNA), the results are often used in statements of "error rates" for microscopical hair examinations. The paper reviewed 170 cases in which microscopical and mtDNA examinations were conducted on human hair samples in forensic casework. Of 170 cases, 133 were sufficient for analysis; of these, nine cases were found where the hairs had a similar microscopic appearance (phenotype), but different mtDNA sequences (genotype). Interpretations of these results by commentators and practitioners demonstrate the mathematical gymnastics some go through to force the numbers into being a rate of error:

> One way to report such data is to say that of the 26 cases in which the mtDNA found an exclusion, the examiners using the visual approach called an association nine times. These data indicate a Type I false-positive error rate of 35% (9/26). Another way to look at the data is to report that nine times out of the 78 times that visual examiners declared an association (12%), the mtDNA technique showed an exclusion. (17)
>
> Houck and Budowle found a false positive rate of 11% or 35%, depending on how one calculates the false positive rate. (18)
>
> …for example, an FBI study that found that 1/8 of hair samples said to "be associated" based on microscopic comparison were subsequently found to come from different people based on DNA analysis. (19)

These "error rates" are calculations based on a misunderstanding of the nature of the original study and putting the numbers to a use for which they were not designed.

Conclusions

It is imperative that researchers and practitioners have a thorough understanding of the various concepts of error in order to design proper research to produce valid forensic science methods, as well as to be able to properly report and explain results. We strongly recommend that educational programs in forensic sciences as well as training programs for practitioners address error and error analysis. We must also consider the legal context as judges and lawyers typically do not understand how error rates are derived or the complexity in separating "mistakes" from uncertainty.

Studies have exhibited varied success in properly evaluating the reliability of certain traits used in analyses. Researchers are applying more sophisticated measurement techniques and statistical analyses to evaluate forensic evidence and are also increasingly finding ways of quantifying traits that have historically remained fairly subjective and thought to be unquantifiable (see (20) for an example from anthropology, (21) for an example from fingerprints, and (22) for an example in forensic hair comparison). For the most part, contemporary research presents error values, but the term is often not defined, and the potential effect on evidentiary examination is not addressed (23). In recent years, we have also seen disciplines working toward the development of standards and best practices and advancement of forensic standards and techniques through the formation of Scientific and Technical Working Groups (24). In addition to recommending best practices for techniques and quality assurance measures to reduce practitioner and instrument error, guidelines also often include recommendations for proper research design and the selection of appropriate statistical methods.

Too often, the term "error" is a source of confusion and even misused in the courtroom and in forensic science. This has occurred despite the increased profile of and reliance on the concept of error following the *Daubert* guidelines and the NAS Report. As forensic scientists, we must be concerned with the clarity, reliability, and validity of our methods. Due to our involvement with the legal system, we should also be proactive in educating the legal community about the differences between scientific error, method limitations, uncertainties, and mistakes and be prepared to mitigate issues related to error. This can best be accomplished by ensuring that we understand, acknowledge, and communicate method limitations and potential sources of error in our research and forensic analyses.

Acknowledgments

We would like to thank several of our colleagues who reviewed earlier versions of this paper.

References

1. Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).
2. Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council. 2009. Strengthening forensic science in the United States: a path forward. Washington, DC: National Academies Press, 2009.
3. Bird SJ. Scientific certainty: research versus forensic perspectives. J Forensic Sci 2001;46(4):978–81.
4. *Federal Rules of Evidence* (1975; 2000).
5. *Frye v. Unites States* 54 App. D.C. 46, 293 F. 1013 (1923).
6. Hand DJ. Measurement theory and practice: the world through quantification. London, U.K.: Arnold, 2004.
7. Page M, Taylor J, Blenkin M. Forensic identification science evidence since *Daubert*: Part I – a quantitative analysis of the exclusion of forensic identification science evidence. J Forensic Sci 2011;56 (5):1180–4.
8. Christensen AM, Crowder CM, Houck MM, Ousley SD. Error, error rates and their meanings in forensic science. Proceedings of the 63rd Annual Meeting of the American Academy of Forensic Sciences, 2011 Feb 21–26; Chicago, IL. Colorado Springs, CO: American Academy of Forensic Sciences, 2011.
9. Dror IE, Charlton D. Why experts make errors. J Forensic Ident 2006;56 (4):600–16.
10. Saks MJ, Koehler JJ. The coming paradigm shift in forensic identification sciences. Science 2005;309:892–5.
11. *United States v Mitchell* 145 F.3d 572 3d Cir. (1998).
12. *People v Gomez*, 99CF 0391 (2002).
13. Page M, Taylor J, Blenkin M. Uniqueness in the forensic identification sciences – Fact or fiction? Forensic Sci Int 2011;206:12–8.
14. American Board of Forensic Odontology (ABFO). Position paper on bitemark workshop 4. ASFO News 2003;22(2):5.
15. *Brewer v. State* 819 So.2d 1169 (2002).
16. Houck MM, Budowle B. Correlation of microscopic and mitochondrial characteristics in human hairs. J Forensic Sci 2002;47(5):964–7.
17. Saks M, Koehler J. Response Sci 2006;311(5761):606.
18. Cole S. More than zero: accounting for error in latent fingerprint identification. J Crim Law 2005;95(3):985–1078.
19. Lander ES. Testimony before the United States Senate Committee on Commerce, Science, and Space. The Science and Standards of Forensics, March 28, 2012.
20. Christensen AM. Testing the reliability of frontal sinus outlines in personal identification. J Forensic Sci 2005;50(1):18–22.
21. Neumann C, Champod C, Puch-Solis R, Egli N, Anthonioz A, Bromage-Griffiths A. Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. J Forensic Sci 2007;52 (1):54–64.
22. Brooks E, Comber B, McNaught I, Robertson J. Digital imaging and image analysis applied to numerical applications in forensic hair examination. Sci Justice 2011;51(1):28–37.
23. Crowder C, Ingvoldstad M. Observer error trends in forensic anthropology. Proceedings of the 61st Annual Meeting of the American Academy of Forensic Sciences; 2009 Feb 16–21; Denver, CO. Colorado Springs, CO: American Academy of Forensic Sciences, 2009.
24. National Institute of Justice. Scientific Working Groups; http://www.nij.gov/topics/forensics/lab-operations/standards/scientific-working-groups.htm (accessed September 14, 2012).

Additional information and reprint requests:
Angi M. Christensen, Ph.D.
George Mason University
4400 University Drive
Fairfax, VA 22030, USA
E-mail: achris12@gmu.edu