JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, March 2012, Vol. 57, No. 2
doi: 10.1111/j.1556-4029.2011.01996.x
Available online at: onlinelibrary.wiley.com

Authors' Response

Sir,

We would first like to express our thanks for your valuable comments on our study about age estimation in a sample of Mexican prisoners. Second, we would like to answer your comments to better clarify for the readers the importance of the *Daubert*'s guidelines in the relationship between science and law.

We do not totally agree that the legal literature does not support the concept that statistical error rate is an absolute requirement under *Daubert*'s (1) or any other legal standard. Rather, we prefer to address this issue from a different point of view. In reality, the Court, in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, proposed four criteria for determining the admissibility of expert opinion under the Federal Rules of Evidence: testability of the underlying methodology, peer review and publication of that methodology, known or potential rate of error, and general acceptance by the relevant scientific community. Unfortunately, in the last two decades, these standards have been difficult to implement in American legal system (2). More specifically, the nonexclusive factor of known or potential error was especially problematic (2).

To understand generally how the courts have managed the known or potential rate of error (KPRE), Haug (2) and Haug and Baird (3) collected all federal trial and appellate cases prior to October 2008, which satisfied the criteria of citing *Daubert* and, specifically, using either the term "error rate" or "KPRE." His research produced 1585 trial cases with several thousand experts involved. At trial court level, he examined several elements, including the frequency with which experts were scrutinized under each factor and the frequency with which they were admitted to testify. For example, only 33 of 200 randomly selected experts were analyzed according to KPRE, and none of those 33 experts were analyzed according to KPRE alone. Not surprisingly, KPRE was the factor considered least among the four *Daubert*'s standards. This suggests that it is perhaps a "last resort" factor when trial courts prepare their support by not admitting an expert. This should be interpreted to mean that trial courts tend to avoid *Daubert*'s error rate. Whenever they considered this factor, they tended to use it to support their decisions not to admit an expert.

You state: "… a failure to disclose a numerical error rate has never been singularly fatal to the admission of expert testimony in any federal case to date in the United States." You also say: "The judges are not specifically required to assess error rates in order to determine expert scientific evidence admissible, although this may be one of several indicators they choose to pay heed."

We are inclined to believe rather that the United States Supreme Court does not know exactly what to do with the "known or potential rate of error" factor. If in the legal literature, one of the most important *Daubert*'s standards is not commonly considered, this is because of the considerable ignorance of the American judges about this matter (2). Whenever the error rate was applied in conjunction with other factors, its application was inconsistent with scientists' concept of error (2). The Court did not provide a specific definition for rate of error. Even theoretically, this lack of specificity is problematic for two reasons. First, while *Daubert* does say that the judge is to evaluate the underlying methodology, the Court has not defined what "underlying methodology" means. The decision does not specify whether error rates are to be measured at a *general* level or at the specific level. Second, the Court gave no guidance as to whether error rate analysis should be limited to a single type of error rate or which type of error rate would be preferred (2). Thus, a judge who does not have expertise in dealing with scientific uncertainty, agree with a particular interpretation, understand the full value or limit of currently used methodologies, or recognize hidden assumptions, biases, or the strengths of scientific inferences may reach an incorrect decision on the reliability and relevance of credible evidence.

Nearly two decades have now passed since the Supreme Court made judges the arbiters of scientific validity through *Daubert v. Merrell Dow Pharmaceuticals, Inc.* (1993). Although this decision was intended to improve how courts use science, recent empirical evidence reveals that judges continue to struggle with scientific evidence and that *Daubert* has failed to yield accurate or consistent decisions (2).

In October 2001, the first comprehensive national study assessing the scientific acumen of 400 state court judges was published. The results of the new study are dramatic. Focusing on how judges use the Daubert criteria to make legal decisions about scientific evidence, the study revealed that, although "judges overwhelmingly support the [Daubert] 'gate-keeping' role, … many of the judges surveyed lacked the scientific literacy seemingly necessitated by Daubert" (4, p. 433). In fact, 96% of the judges failed to demonstrate even basic understanding of two of the four *Daubert* criteria. For example, while 88% of the judges reported that "falsifiability" is a useful guideline for determining the quality of proffered scientific evidence, 96% of these same judges lacked even basic understanding of this core scientific concept. Scientific methodology today is based on generating hypotheses and testing them to see whether they can be falsified; indeed, this methodology is what distinguishes science from other fields of human inquiry. Surveyed judges were not expected to demonstrate even this level of comprehension. In fact, responses as simple as: "I would want to know to what extent the theory has been properly and sufficiently tested and whether or not there has been research that has attempted to prove the theory to be wrong," or "if it is not possible to test the evidence then it would weigh heavily with me in my decision" were deemed to be accurate statements. Only 14 of 352 judges demonstrated even this level of understanding. Similarly, 91% of the judges reported that they found error rates useful for determining the quality of proffered scientific evidence (3). Here again, judges do not seem to understand the scientific concepts they routinely employ. The *Daubert* Court cautioned that "in the case of a particular scientific technique, the court ordinarily should consider the known or potential rate of error and the existence and maintenance of standards controlling the technique's operation" (1, pp. 593-594). However, judges have misunderstood the definition of error rates and, therefore, their significance. When error rates are used to assess the validity of a scientific method, they may include false-negative errors (when an experimenter misses a real effect), false-positive errors (when an experimenter perceives an effect that did not in fact occur), and sampling errors (e.g., when an experimenter extrapolates from a small sample to a large population). Only 4% of the judges who reported that error rates were useful demonstrated fundamentally accurate understanding of the definition of error rates. As with falsifiability, the researchers did not expect a highly sophisticated level of comprehension. Responses defined as accurate included: "It would seem that if a theory or procedure has too high an error rate it would have to be rejected because the risk is too high of being wrong" and "I would want to

know about the probability of making a mistake" (4, p. 444). Only 15 of 364 judges had even this level of comprehension. This means that, two decades after *Daubert*, courts have systemic and ongoing problems in assessing the quality of scientific evidence. If judges themselves do not understand the *Daubert* criteria, they cannot hope to make meaningful, accurate, or consistent assessments of scientific evidence. If a large number of judges are clearly confused or ill-informed about basic scientific principles, *Daubert* cannot be accurately or consistently applied.

We agree that in *Kumho v. Carmichael*, the Justice Breyer stated that all of *Daubert*'s criteria do not necessarily apply in all circumstances, even when scientific evidence is being reviewed. However, according to the *Kumho* decision (5), the court also realized that there must be a flexible approach in assessing expert testimony, considering how the type of evidence may vary across disciplines. The judge just considered that not all of *Daubert*'s criteria may be applicable to the expert testimony and those which do apply should be used to evaluate admissibility (6). For example, depending on each case, the reporting of statistical error should be necessary and could be addressed to assess expert admissible scientific evidence. Currently, it is not necessary to show that every method we use is highly reliable; rather, it is imperative to demonstrate that we are sure (statistically and scientifically) of how reliable a technique is (6). The challenge is to employ research designs that adequately test the variable(s) of interest and give us proper measures of reliability (6). The *Daubert* validity inquiry needs to be reformulated, so that the forensic methodology's "error rate" factor is the primary (and if possible, the only) factor the court considers.

As regards the 95% confidence levels, we agree that "correctly classify an individual 95% of the time" is not (and has never been) a recognized legal standard. In fact, the use is of this standard is just considered by the statistical community as a historical artifact and again, using 95% is a purely arbitrary convention. This is the interval, computed from the study sample data, within which we can expect the population value to lie with a 95% level of probability (i.e., we can be 95% confident that the population value falls within this interval). Intervals can be calculated for any desired level of confidence and depend on the two factors that cause the main effect to vary: the number of observations and the spread in the data (commonly measured as a standard deviation). The 95% confidence level is the most popular, but some authors use 99% and 90% is seen on occasion; 95% is usually chosen because it conforms to the customary acceptance of a 5% *p*-value as the threshold for statistical significance. People use different *p*-values and confidence intervals in special situations, but 0.05 and 95% seem to work most of the time (7–10). With a bigger sample, the 95% confidence interval gets narrower, and the results of the study become more precise.

We agree that the 95% confidence range "is not an absolute requirement under *Daubert* or any other legal standard." However, in *Merrell Dow Pharmaceuticals v. Havner* (11,12), for example, the court selected the 95% statistical significance level as the minimum threshold in order for epidemiological studies to constitute "some evidence" of causation. The court explained the significance of this approach: "The generally accepted significance level or confidence level in epidemiological studies is 95%, meaning that if the study were repeated numerous times, the confidence interval would indicate the range of relative risk values that would result 95% of the time" (12, p. 723). In adopting this norm, the court cautioned that: "We should not widen the boundaries at which courts will acknowledge a statistically significant association beyond the 95% level to 90% of lower values" (12, p. 724).

Disciplines like Forensic Anthropology may be problematic in the eyes of the courts, because they involve some degree of subjectivity, but subjectivity does not necessarily equal unreliability (6).

Forensic Anthropology is an applied discipline and should be treated as such. In fact, forensic anthropologists can set statistical standards (precision, accuracy, and bias) for a theoretical and empirical validation process to guide researchers and practitioners as well as assist the courts. Some forms of anthropological testimony could be therefore subject to *Daubert*'s guidelines. For example, the Supreme Court noted that a judge should consider *Daubert*'s standards in situations where they are a reasonable measure of reliability of expert testimony (1,3). This type of testimony is based on methodology, which is quantitative and testable and has definable error rates, examples being methods used to estimate the sex and age of an unknown skeleton. These techniques have established biologically based categories and a limited number of variables that assess which category best describes an unidentified individual. Another example of a skill that falls under *Daubert*'s standards is the estimation of time since death. Some techniques used to establish the *postmortem* interval are more empirical and are subject to *Daubert*'s standards as they use well-defined stages and mathematical and statistical descriptions.

Most forensic researchers recognize that in many cases, the probability of misidentification for "that particular case" would be difficult to estimate, but this is precisely why scientists assess this error using experimental studies. Based on proper hypothesis testing and statistical analysis of collected data, researchers can put a probability or confidence interval on their likelihood of correct assessment (6).

The importance of individual experience and the need for intuitive decision making suggest that it is unlikely that precise, standardized protocols can be developed for assessing age in Forensic Anthropology. As other fields of Forensic Sciences attempt to increase their precision and admissibility in court by carefully defining "best practice" standards for the acceptance and interpretation of their particular brand of forensic evidence, Anthropology may find this task difficult. So, how can we standardize an approach that includes so much intuitive and subjective assessment? How can we decide which method or combination of methods is best, when none of them is able to account for more than 50% of the variability in skeletal indicators? Perhaps, the best we can do is to make sure that our individual methods are statistically valid and educate ourselves regarding their systematic biases.

References

1. Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 US 579 (1993).
2. Haug M. Minimizing uncertainty in scientific evidence. In: Cwik CH, Witt HE, editors. Scientific Evidence Review. Monograph No. 7, January 2006.
3. Haug M, Baird E. Finding the error in *Daubert*. Hastings Law J 2011;62:737–56.
4. Gatowski S, Dobbin SA, Richardson JT, Ginsburg GP, Merlino ML, Dahir V. Asking the gatekeepers: a national survey of judges on judging expert evidence in a post-*Daubert* world. Law Hum Behav 2001;25(5):433–58.
5. Kumho Tire Company Ltd. v. Carmichael, 526 US 137 (1999).
6. Christensen AM, Crowder CM. Evidentiary standards for forensic anthropology. J Forensic Sci 2009;54(6):1221–6.
7. Chaudhary MA, Stearns SC. Estimating confidence intervals for cost-effectiveness ratios: an example from a randomized trial. Stat Med 1996;15:1447–58.

8. Davison AC. Statistical models. Cambridge, UK: Cambridge University Press, 2003.
9. Weerahandi S. Exact statistical methods for data analysis. New York, NY: Springer-Verlag, 2003.
10. Hirji KF. Exact analysis of discrete data. New York, NY: CRC Press/Chapman and Hall, 2006.
11. Baker SB. ''Gatekeeping'' in Texas: the practical impact of full implementation of the Texas rules of civil evidence regarding experts. 27 St. Mary's L J 1996;237:256–7.
12. Merrell Dow Pharmaceuticals Inc. v. Havner, 953 S.W.2d 706 Tex. 1997.

Stefano De Luca,[1] M.Sc.; Josefina Bautista,[2] Ph.D.; Inmaculada Alemán,[1] Ph.D.; and Roberto Cameriere,[3] Ph.D.
[1]Laboratory of Anthropology, Faculty of Medicine, University of Granada, Granada, Spain.
[2]Department of Physical Anthropology, INAH, Mexico City, Mexico.
[3]AgEstimation Project, Institute of Legal Medicine, University of Macerata, Macerata, Italy.
E-mail: stefanotlatoani@hotmail.com