CHAPTER 32 The Pervasiveness of Daubert

Stephen D. Ousley and R. Eric Hollinger

Introduction

There have been many recent publications (e.g., Christensen and Crowder 2009; Foster and Huber 1999) about the meaning and impact of the Daubert decision (*Daubert v. Merrell Dow Pharmaceuticals* 1993) and subsequent rulings on forensic science. The role of science in legal proceedings has become preeminent and the "CSI effect," in which many juries expect to see cutting-edge scientific evidence, is undeniable: one jury recently refused to convict, despite enough circumstantial evidence, because a DNA test was not run on a half-eaten hamburger (Stockwell 2005). At the same time, in response to the Daubert ruling, previously used forensic methods (such as bite-mark analysis) have been largely discontinued and even fingerprint and DNA analyses are being reevaluated. The Daubert decision also played a role in the recent National Academy of Sciences findings (Holden 2009) calling for an independent governmental body to review scientific methods in the forensic sciences. We maintain that Daubert was necessary, forensic anthropology has improved since Daubert, and all of the forensic sciences are becoming more scientific because of it.

While technically applying only to scientific testimony, the main thrust of the Daubert decision is that scientists doing forensic work must do good science. Most importantly, scientific results do not stand alone; there are no results, *prima facie*, that prove any disputed fact: scientific conclusions must be qualified, or, ideally, quantified, as to the probability that the conclusions are correct, through estimated error rates. Daubert diminished the authority of the expert witness based solely on experience and reputation, and instead emphasized the quality of the expert's methods. The Frye standard for scientific evidence of "general acceptance" was in part superseded by emphasizing the more neutral concepts of reliability and validity as criteria to evaluate

A Companion to Forensic Anthropology, First Edition. Edited by Dennis C. Dirkmaat.

the correctness of scientific conclusions. Daubert thus allows more innovation in scientific methods. The Daubert decision resulted in significant changes in the Federal Rules of Evidence rule 702, which is in force for federal civil law and incorporated into many state criminal law proceedings (Foster and Huber 1999). Additionally, a recent supreme court case (*Melendez-Diaz v. Massachusetts* 2009) established that scientists and technicians who evaluate evidence in any way must be prepared to testify about any methods and techniques used to establish a fact, even the methods for determining the weight of drugs recovered from a crime scene.

In this chapter we will define the terms "reliability" and "validity" and explain why they are important in any scientific investigation, especially after Daubert, and how they can be estimated, especially in forensic anthropology. Following Carmines and Zeller (1979), we will emphasize that validity and reliability are not absolute, but are matters of degree. We will explore an additional factor in evaluating scientific work, which we refer to as "stupidity" that incorporates practitioner error. Stupidity is important because it is unavoidable. In contrast to validity and reliability, stupidity is inherently unquantifiable; it can however be intuitively qualified. We also will clarify the terms "error," "precision," and "accuracy" in scientific investigations, especially statistical analyses. Understanding all of these concepts is necessary to help choose the best methods in scientific investigations, to understand why incorrect conclusions will be reached at times even using the best methods, and to judge which conclusions are more likely to be true when conclusions derived from different methods disagree.

Error

The word "error" is ambiguous because it has been applied to a mistake, to an incorrect conclusion, or, in statistics, to normal variability, or to "noise." In this chapter we will only use "error" to mean a mistake, and as such we will use it in measuring reliability and qualifying stupidity. We also advocate using error in this restricted sense in discussing scientific concerns, as the following explanations illustrate.

Mistakes

If observations (measurements or trait states) are recorded incorrectly, then the use of those observations will be worthless or misleading. W.W. Howells (1973) delineated seven sources of mistakes when recording osteological measurements: (i) application of technique: how things are measured, defined, or observed; (ii) interobserver error: inconsistencies among individuals measuring; (iii) intraobserver error: inconsistencies by an individual measurer; (iv) instrument errors: defective or uncalibrated instruments; (v) instrument reading errors: misreading numbers; (vi) recording error: writing down an observation incorrectly; and (vii) computer data entry error: final stage error. All of these kinds of mistakes influence reliability in that they will tend to produce inconsistent recorded measurements from the same bone. The magnitude of intra- and interobserver errors (or, more neutrally, differences) can be calculated using various statistics. For instance, differences of 1–2 mm are not uncommon when measuring a large measurement such as glabello-occipital length, with a mean of about 175 mm, and is often at most 1.5% of the mean measurement. However, 1–2 mm

represents a relatively larger error when measuring something like interorbital breadth, with a mean measurement of 23 mm. Miscalibrated instruments, most often digital calipers that were not zeroed before measurement, or a spreading caliper with bent arms, will likewise lead to inconsistent measurements compared to properly calibrated instruments. Naturally, in order to help minimize interobserver differences, practitioners must study measurement definitions and must practice measuring. Practice is even more important in scoring nonmetric traits because the consequences of interobserver differences are much greater: for a three-state trait (small, medium, large), a difference of one grade in scoring results in 50% disagreement between observers; for a two-state trait (presence/absence), a difference of one grade in scoring results in 100% disagreement between observers. Additionally, qualifying a nasal breadth as small, medium, or large will likely be influenced by the last 100 nasal apertures seen by the subjective observer. In contrast, measuring a nasal breadth will produce an objective quantity that contains more information. Also, measured nasal breadths can be converted later to small, medium, and large if desired, but not vice versa. Therefore, explicit trait definitions and standardization are essential for replicating the observation methods used in scientific publications and for high reliability (consistency or repeatability), and measures of repeatability, such as intra- and interobserver differences, are important and currently are published more often. Observational methods with higher repeatability are inherently better than those with lower repeatability. If there is inconsistency in recording observations, the observations will show low reliability and will be of little practical use in analyses.

Incorrect conclusions

Daubert emphasizes a method's known or potential "error rate." It is advised to use techniques that have the lowest error rates, but what is meant really is a lower probability of drawing an incorrect conclusion; in other words, we should use methods with the highest validity. Because the *potential* error rate of *any* method is 100%, we should choose the method that most likely provides correct answers, assuming no mistakes in observations, analysis, interpretations, and conclusions using the data and method. Following Daubert, the expert witness should be able to estimate the probability of being correct objectively, using the specified data and methods, rather than subjectively. There are no techniques that are 100% correct because of a persistent nonzero probability of stupidity, but certain techniques are more likely to produce correct answers than other techniques. Partial fingerprints are not nearly as informative for individual identification as a complete 10-fingerprint set. A single bone is not as informative as a complete skeleton. A sex classification function that classifies male and female reference samples 95% correctly has, in the long run and under the best conditions, a 5% chance of incorrectly classifying all future remains analyzed, which is sometimes termed a 5% classification error rate. When the expert has little evidence to work with, the expert must simply accept that only weak conclusions can be made with confidence. Depending on the condition and completeness of the remains present, estimating sex may be no better than guessing. Incomplete remains having a 50% probability of being male only involves "error" if one draws a firm conclusion about a particular sex, because there is a 50% probability that the conclusion is incorrect. A higher probability of correctness means that the probability of making an

incorrect conclusion is lower. A conclusion can end up being correct for the right reasons, or for the wrong reasons, such as estimating sex by flipping a coin. We maintain that a conclusion with a higher probability of being correct that later is shown to be incorrect is more defensible than a conclusion with a lower probability of being correct that later is shown to be correct. We maintain that coming to a conclusion that has greater support is more defensible than coming to a conclusion with less support, no matter which one eventually is shown to be correct. It would indeed be an error, a mistake, to draw a conclusion with less support rather than drawing a different conclusion with greater support.

Of course, the value of reaching a correct conclusion depends on the background, or prior, probabilities. Predicting heads or tails correctly 50% of the time when flipping a coin is not very impressive. As mentioned in Chapter 16 in this volume, on stature estimation, our conclusions should be more accurate and specific, when possible, than those that can be made using no information from the present case and merely based on the prior probabilities. Importantly, although randomness is often emphasized, it is most often thought of as equal probabilities, a 50/50 chance when there are two choices, as in the case of coin flipping. But that is not always the case. In the classroom one of us (SDO) demonstrates an amazing computer program that predicts handedness of an individual based on mathematical manipulations of his or her date of birth. In every undergraduate class tested so far, it has consistently been at least 85% correct, and in many classes it is 100% correct. The students are impressed by the demonstration, in which every prediction is that the student is right-handed. Once informed that in the general population, 90% are right handed, the students are not so impressed. The high prior probability of being right handed is the reason that methods for determining handedness from the skeleton are difficult to justify, because any proposed method would need to be much better than 90% correct.

Noise or variability

"Error," the "error term," and the "standard error" are all related to statistical analysis but can have very different underlying meanings depending on the analysis. Early statistical methods were employed to estimate constants such as the circumference of the Earth. Any deviations from the constants were, by definition, errors in estimation or calculation, artificial, and therefore statistical "noise." When biological measures first were studied, such as human statures, people from different countries showed different mean measurements, and a normal curve was often observed within groups. Unfortunately, "error" was also used to describe these normal deviations from the mean, normal variations, which are seen in nearly all linear measurements in virtually all animals (Stigler 1999). Likewise, in linear regression, the "standard error" represents deviations in actual values from predicted values represented by the regression line. In forensic stature estimation, there is not a perfect relationship between bone length and stature, because a bone length is one small component, along with many other components, that make up stature. Accounting for normal variation is exceedingly important in stature estimation, with an important trade-off between precision and correctness: the more precise the stature estimation is, the less certain we can be that the actual stature is contained within the prediction interval. Stature estimates are often given in prediction intervals, which define how often we should be correct in the

long run; this is an estimate of validity. We will be incorrect less often when we use a 99% prediction interval for stature given a bone length than when we use a 90% prediction interval. The 99% prediction interval is rather wide and may not help narrow down possible identifications, but we must accept the variability in stature given specific bone lengths. As long as the stature prediction is accurate (unbiased), we can be correct nearly 100% of the time if we use a plus/minus 15 cm prediction interval, but we would sacrifice precision for correctness.

Reliability

"Reliability" in a scientific and legal sense is frequently misunderstood due to its ordinary language meaning, similar to dependability, which is actually closer in meaning to validity. Reliability is necessarily a component of validity but, in a forensic sense, using reliable methods – that is, measuring something correctly and consistently – may have little or nothing to do with reaching the correct conclusion, which reflects validity. We follow Hand (2004), in that reliability is best thought of as consistency in measurement, or how closely different observers measure or score the same phenomenon. Highly reliable measurements show very low or no interobserver errors and high repeatability. A clock that shows the correct time consistently is a reliable and valid clock; a clock that is consistently 1 hour ahead is reliable but invalid; in each case the clock presents a consistent representation of time. Thus, reliability ideally should be used only for observations and measurement, the beginnings of any analysis that will lead to meaningful conclusions. Inter- as well as intraobserver differences are naturally of paramount concern when measuring reliability. Collecting reliable bone measurements involves using an instrument, and for small bones a digital sliding caliper accurate to 0.01 mm should be more reliable than using an osteometric board, which should be more reliable than holding a bone next to a meter stick and "eyeballing" a measurement. When measuring a long bone such as a femur, using an osteometric board is likely more reliable than using multiple measurements from a 200 mm digital sliding caliper. Of course, the digital caliper and osteometric board should be calibrated. If digital calipers are not properly zeroed, they can nonetheless produce repeatable and reliable results for many observers, at least until they are zeroed. In this limited case it could be said that the measurements written down are reliable but invalid (in the strict sense of "face validity") representations of the bone length (Carmines and Zeller 1979). In other words, the measurements are consistent, but consistently wrong. Also, some measurements are inherently more reliable than others, no matter what instrument is used. Measurements involving maximum lengths of long bones show higher reliability than a bone measurement such as the condylomalleolar length of the tibia, which requires the correct orientation of the tibia and a specialized osteometric board (Moore-Jansen et al. 1994). On the cranium, the measurement of frontal chord (nasion–bregma) shows lower interobserver differences than measuring orbital breadth, and especially mastoid height, because the frontal chord involves measuring the distance between two well-defined points. Importantly, sources of observational error occur with DNA sequencing, and extracting an incorrect DNA sequence due to DNA contamination may be more likely than due to computer or instrument error. It should be clear that reliability is

not absolute, in that no data-extraction method can be said to be perfectly reliable. Rather, the degree of reliability represents consistency and repeatability of measurements or observations using specific instruments and procedures.

Validity

Validity is the strength of the connection between a hypothesis and the real-world application or conclusion (Carmines and Zeller 1979). Validity can be subdivided but as a whole it covers the entire process of data collection, data entry, data analysis, reaching general conclusions, drawing inferences, and then coming to specific conclusions about the case at hand (Foster and Huber 1999). For example, measurements are collected from the bones of an unidentified person; statistical analyses are employed; overall, the measurements are more similar to those in a reference sample of males as opposed to females; due to the greater similarity with males, we infer that the remains come from a male. Validity is estimated through the process of validation, in which the same techniques and methods are used under similar circumstances and any discrepancies in conclusions are noted. Reliability is a part of validity because reliability, or consistency in measurements or observations, is required during data collection. Unreliable data cannot lead to valid conclusions. Validity is more often applied to methods, rather than data, except when an instrument measures something of direct importance, such as a clock, which measures time. As mentioned, a clock that is consistently 1 hour ahead is reliable but invalid, because it never shows the correct time (although with the knowledge that it consistently runs ahead, making it reliable, we can calculate the correct time). A stopped clock is neither reliable nor correct because it is inconsistent and never (well, okay, twice a day) tells the correct time. In analyzing a reliable mitochondrial DNA sequence, for example, heteroplasmy can cause false exclusions or false matches, and is due to the imperfect validity of matching an unidentified mitochondrial DNA sample to a known sample.

Estimating validity, specifically what is known as predictive validity, is related to a method's known or potential error rate (i.e., how often it provides a correct conclusion) referenced in Daubert. Predictive validity can be estimated and quantified most easily using statistical methods such as discriminant function analysis (DFA) or other classification procedures, of course, assuming that the measurements are taken correctly. DFA provides classification accuracies for specified measurements and groups as part of the procedure using a validation sample that is very similar to the reference samples (Hastie et al. 2001). In DFA, bone measurements of the reference groups are statistically manipulated to separate them as much as possible. Then, individuals in the validation sample as well as the unknown individual simply are classified into the group to which they are most morphologically similar. Some of the classifications will be incorrect, but the classifications are tabulated and the classification percentage correct is calculated. As you may imagine, with more measurements, the accuracy increases; there is more information and greater intergroup differences when using 10 measurements than when using two. Higher classification accuracies represent higher validity for the method of estimating sex from multiple bone measurements. Naturally, the inference must be made that the individual comes from one of the groups, the groups are most relevant to the questions at hand, and the estimated percentages

correct will hold up in the long run. If improper samples are used, the conclusions will be erroneous, which involves another aspect of validity, termed external validity, because the inference that the group samples are appropriate may not be correct. For example, Giles and Elliot (1962) published a novel way of using DFA to classify skeletal remains into three groups using cranial measurements, but their samples were from nineteenth-century American blacks and whites and a limited sample of American Indians. As Ayres and Jantz (1990) illustrated, secular changes in cranial morphology cause more incorrect classifications of modern crania using the Giles and Elliot formulas. Likewise, İşcan and Cotton's (1990) DFA using postcranial measurements performed very poorly when tested against modern Americans (Ousley and Jantz 1997).

Stupidity

Stupidity is a term we use largely for practitioner mistakes, any errors in execution that can affect reliability and validity, or can be largely independent of them. This category may be necessary because Christensen and Crowder (2009: 5) remarked that "Practitioner mistakes, especially those that result in misidentification, challenge the view of *method reliability* regardless of the validity of the method" (emphasis added). As mentioned, we agree with other authors in that analytical methods are best understood in terms of validity rather than reliability. Otherwise, there is a virtually unlimited number of analytical methods depending on which specific tools and procedures are used at what time. For example, working on a case while fatigued will probably result in more errors, and if these errors are more probable because of the method, then this analysis, in the presence of fatigue, is technically a different method. However, practitioner mistakes, or stupidity, can affect the validation process, or can be independent of reliability and validity. Elements of stupidity would include incorrectly writing down a measurement even though the measurement was taken correctly (possibly affecting reliability and validity), inadvertently entering the wrong number into a computer or calculator when calculating something such as a discriminant function score (possibly affecting validity), or incorrectly writing down the result of a test that results in an incorrect conclusion (affecting validity). More specific examples of the role of stupidity involve analyzing craniometrics to estimate sex or ancestry. For instance, to analyze a case, I could take measurements using a tape measure, enter those measurements into a calculator, and use the Giles and Elliot (1962) formulas; or I could take measurements using a spreading caliper, enter those measurements into a calculator, and use the Giles and Elliot (1962) or other published formulas; or I could measure using a vernier caliper, calculate the DFA coefficients using Fordisc 3 (Jantz and Ousley 2005), and use a calculator to analyze the measurements; or I could measure using a digital caliper and analyze the measurements using any number of methods available in commercial statistical packages such as R (R Development Core Team 2009), SAS (SAS Institute 2001), or SYSTAT (Systat Software 2004); or I could measure using a digital caliper and enter the measurements directly into Fordisc 3; or I could use a three-dimensional digitizer and software to register landmarks, have software calculate the standard measurements, then import the measurements directly into Fordisc 3 for analysis. In each of these analyses, there are different reliability and validity concerns, and the method may change only slightly, but the

opportunity for stupidity to affect results diminishes with each successive approach. From these examples we can see that the same analytical method can be used in a variety of ways, and not only are some methods better than others, some procedures for employing the same method are more likely to produce errors than others. In general, automated methods are more easily replicable, but certain kinds of mistakes are harder to detect using automated procedures. Most importantly, unlike measures of reliability and validity, the measurement of stupidity cannot be quantified, but can only be qualified because we know that certain methods (such as manual data recording and entry) are more prone to mistakes than others.

Two Cautionary Tales

Estimating sex using the mandible

Loth and Henneberg (1996) proposed a method for estimating sex from the adult mandible by examining the flexure of the posterior border of the ramus, which was a trait they discovered and defined, and supposedly found only in males. They reported an overall accuracy in sex estimation of 94% for healthy mandibles with teeth present. Koski (1996) concluded that the method did not perform as well as claimed, but his validation study was limited to radiographs from females only, many of whom were under 10 years of age; it was a poor test of the method. Later, Donnelly et al. (1998) tested the method more appropriately using dry bones from adults. In their blind tests, they found that 63% of the mandibles were sexed correctly on average, with a male bias in estimating sex, the mandibular flexure trait was difficult to recognize and score as present or absent, and interobserver differences were high. Estimating sex using mandibular flexure, then, is an extreme case: it is unreliable due to interobserver differences and invalid because classifications using it are no better than by chance. There have been several other validation studies that echo the conclusions of Donnelly et al. (1998). The importance of validation studies through independent and blind tests of methods, which incorporate the concerns of reliability, validity, and stupidity, cannot be overemphasized.

The bones of Everett Ruess?

In May 2008, skeletal remains were found in a remote part of Utah, USA, after following up on a Navajo family's story of the murder of a white man in 1934. An initial DNA test from a molar indicated that the DNA came from a European, rather than a Native American. Dennis Van Gerven, at the University of Colorado at Boulder, excavated the burial site further, analyzed the remains, and decided that the remains came from a white male about 5 feet 8 inches (172.7 cm) tall, aged between 19 and 22 years old. Everett Ruess, a 5 foot 8 inch, 20-year-old writer and artist, exploring isolated parts of Utah, had disappeared in 1934, so a possible identification was obvious. Van Gerven and his assistant reassembled the cranium and then used skull–photo superimposition to compare the remains and photographs of Everett Ruess. Van Gerven was quite certain in his conclusions: "Everett had unique facial features, including a really large, jutting chin. This guy had the same features. And the bones match the photos in every last detail, even down to the spacing between the teeth.

The odds are astronomically small that this could be a coincidence…. I'd take it to court. This is Everett Ruess" (Roberts 2009). Later, a more precise DNA test was conducted comparing the DNA of the remains to DNA from saliva of four of Ruess' nephews and nieces. Dr Kenneth Krauter, a molecular biologist at the University of Colorado at Boulder, analyzed 600 000 DNA markers (much more than the usual 18 or so used in forensic comparisons) and found that the individual shared 25% of the markers with Ruess' nieces and nephews, the expected percentage in such relatives, and did not share nearly as many of the same markers with a random sample of other individuals. Krauter concluded "This is a textbook case…. The evidence is irrefutable that the bones are from a close relative of the four, … Combined with the facial reconstruction, that makes this an irrefutable case," and "The combination of the forensic and genetic analyses makes it an open and shut case…. I believe it would hold up in any court in the country" (National Geographic 2009; Roberts 2010). Van Gerven later added "If this were going before a court of law, you'd want to build a case…. That's what we've done here, with Navajo oral tradition, the forensic analysis and now the DNA test. We can be certain that this is Ruess" (National Geographic 2009).

Ruess' relatives received the remains and were planning to cremate them but concerns were raised by the Utah state archaeologist and a physical anthropologist (Jones and Kopp 2009): most importantly, the lack of peer review, with results only available in magazine articles and press releases; haphazard recovery techniques; inconsistencies in the Navajo family story about the murder; a much greater amount of tooth wear seen in photos of the mandible than found in contemporary European Americans and more typical for American Indians; a probable large untreated tooth cavity, which Ruess would have had treated; questions about the skull–photo superimposition methods; and finally, they suggested the DNA results should be tested by a laboratory with experience analyzing ancient DNA. The relatives of Ruess agreed to a second DNA test, done this time by the Armed Forces DNA Identification Laboratory, which has extensive experience in analyzing DNA from American soldiers and testing against possible relatives. In October 2009, the shocking results came back: not only were the remains not from Everett Ruess, they were from a Native American.

What went wrong? How could different analyses, especially of the same DNA, give different conclusions? Apparently, there was no contamination of the individual's DNA. Also, the DNA sequences were read correctly, which indicates that reliability is not the issue. The analysis of the DNA was at fault. Krauter used sophisticated software to compare the 600 000 DNA markers that is used in medical research in the living. When using the software on ancient DNA, which in many cases is degraded, for some reason the software treated the greater noise associated with degraded DNA as DNA marker matches with the relatives. Krauter paid far more attention to the results of the method than to the process of arriving at those results. He later said "We screwed up by relying on the technology too much," and "Fortunately, the error uncovered how the extreme sensitivity can be misleading if a researcher takes its output at face value" (Roberts 2010). Thus, while DNA evidence is often compelling, it can be misleading. The methods of analyzing DNA evidence are evolving still, and each new method must be evaluated as to its validity in providing answers to forensic questions.

What about Van Gerven's inexorable conclusions from the independent skull–photo superimposition? Van Gerven expressed disappointment that his conclusions were shown to be incorrect and offered no further comments on the "unique" skeletal and dental features that identified the remains as Ruess. Despite the impression of unique features, it may well simply be possible to manipulate antemortem photographs and images of a cranium to produce a minimum of discrepancies. The Ruess case illustrates why skull–photo superimposition may involve as much art as science, because we do not know the random probability of matching a number of features in the cranium to photos of different people. Uncanny resemblances and similarities, using skull–photo superimposition or other methods, may be mere coincidence. In other words, skull–photo superimposition has uncertain validity for uniquely identifying an individual.

We suspect that the Ruess case is an example of confirmation bias; namely, when you look for something, you will probably find it, and you rarely will rethink your own results when they meet your expectations. We are all susceptible to confirmation bias, which has been demonstrated in a blind test of fingerprint experts, who were told the opinion (which disagreed with the experts' previous assessment) of a fictitious FBI fingerprint expert who examined the same prints. When the experts reexamined the prints, the experts often unknowingly reversed their former conclusions (Dror et al. 2006). Van Gerven likely was influenced by the initial DNA finding indicating European ancestry and possible identification, and his results in turn likely influenced Krauter's DNA conclusions, though technically they all should be independent: results from one independent test should not be considered as supporting the results of another test.

Conclusions

The Daubert ruling clarified what good scientific methods involve: accurate, consistent, and repeatable measurement techniques (reflecting reliability), making assumptions and inferences explicit, and using the most accurate methods available for reaching conclusions (reflecting validity). Methods with greater reliability and validity are preferred over those with lower reliability and validity. At the same time, stupidity should be minimized, and precision maximized, as much as is practical. When different methods produce different conclusions we will choose the method that most likely produces the correct answer given the concerns of reliability, validity, and stupidity. Nevertheless, we will at times come to an incorrect conclusion even when we use the best methods; this is only a mistake if we made mistakes during data collection or analysis. Validity and reliability are probabilistic, and no method is 100% correct at producing results or conclusions.

REFERENCES

Ayres, H.G., Jantz, R.L., and Moore-Jansen, P.H. (1990). Giles and Elliott race discriminant functions revisited: a test using recent forensic cases. In Gill, G.W. and Rhine, J.S. (eds), *Skeletal Attribution of Race* (pp. 65–71). Anthropological paper no. 4. Maxwell Museum of Anthropology, Albuquerque, NM.

Carmines, E.G. and Zeller, R.A. (1979). *Reliability and Validity Assessment*. Sage Publications, Newbury Park, CA.

Christensen, A.M. and Crowder, C.M. (2009). Evidentiary standards for forensic anthropology. *Journal of Forensic Sciences* 54: 1211–1216.

Daubert v. Merrell Dow Pharmaceuticals, 509 US 579, 113 S.Ct. 2786, 125 L.Ed. 2d 469 (US June 28, 1993) (No. 92–102).

Donnelly, S.M., Hens, S.M., Rogers, N.L., and Schneider, K.L. (1998). Technical note: a blind test of mandibular ramus flexure as a morphologic indicator of sexual dimorphism in the human skeleton. *American Journal of Physical Anthropology* 107: 363–366.

Dror, I.E., Charlton, D., and Peron, A.E. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International* 156: 74–78.

Foster, K.R. and Huber, P.W. (1999). *Judging Science: Scientific Knowledge and the Federal Courts*. MIT Press, Cambridge, MA.

Giles, E. and Elliott, O. (1962). Race identification from cranial measurements. *Journal of Forensic Sciences* 7: 147–157.

Hand, D.J. (2004). *Measurement Theory and Practice: The World Through Quantification*. Arnold, London.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Holden, C. (2009). Forensic science needs a major overhaul, panel says. *Science* 323: 1155.

Howells, W.W. (1973). *Cranial Variation in Man: A Study by Multivariate Analysis of Patterns of Difference Among Recent Human Populations*. Papers of the Peabody Museum, vol. 67. Peabody Museum of Archeology and Ethnology, Harvard University, Cambridge, MA.

İşcan, M.Y. and Cotton, T.S. (1990). Osteometric assessment of racial affinity from multiple sites in the postcranial skeleton. In Gill, G.W. and Rhine, J.S. (eds), *Skeletal Attribution of Race* (pp. 83–90). Anthropological paper no. 4. Maxwell Museum of Anthropology, Albuquerque, NM.

Jones, K. and Kopp, D. (2009). *Everett Ruess – A Suggestion to Take Another Look*. http://history.utah.gov/archaeology/ruess.html.

Koski, K. (1996). Mandibular ramus flexure–indicator of sexual dimorphism? *American Journal of Physical Anthropology* 101: 545–546.

Jantz, R.L. and Ousley, S.D. (2005). *FORDISC 3: Computerized Forensic Discriminant Functions*. Version 3.1. University of Tennessee, Knoxville, TN.

Loth, S.R. and Henneberg, M. (1996). Mandibular ramus flexure: a new morphologic indicator of sexual dimorphism in the human skeleton. *American Journal of Physical Anthropology*, 99:473–85.

Melendez-Diaz v. Massachusetts, 129 S.Ct. 2527, 69 Mass. App. 1114, 870 N.E. 2d 676, reversed and remanded (2009) (No. 07–591).

Moore-Jansen, P.M., Ousley, S.D., and Jantz, R.L. (1994). *Data Collection Procedures for Forensic Skeletal Material*, 3rd edn. University of Tennessee, Knoxville, TN.

National Geographic. (2009). *Press release: after 75 years, National Geographic ADVENTURE solves mystery of lost explorer*. April 30. http://adventure.nationalgeographic.com/2009/04/everett-ruess/dna-test-text.

Ousley, S.D. and Jantz, R.L. (1997). The forensic data bank: documenting skeletal trends in the United States. In K. Reichs (ed.), *Forensic Osteology*, 2nd edn (pp. 297–315). Charles C. Thomas, Springfield, IL.

R Development Core Team. (2009). *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. www.R-project.org.

Roberts, D. (2009). *Finding Everett Ruess*. April/May. http://adventure.nationalgeographic.com/print/2009/04/everett-ruess/david-roberts-text.

Roberts, D. (2010). *Everett Ruess Update: How the DNA Test Went Wrong*. February 2. http://ngadventure.typepad.com/blog/2010/02/everett-ruess-how-the-dna-test-went-wrong.html.

SAS Institute (2001). *SAS/SHARE 9 User's Guide*. SAS Institute Inc., Cary, NC.

Stigler, S.M. (1999). *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, MA.

Stockwell, J. (2005). Defense, prosecution play to new 'CSI' savvy. *Washington Post* May 22. www.washingtonpost.com/wp-dyn/content/article/2005/05/21/AR2005052100831.html.

Systat Software (2004). *Systat Version 11*. Systat Software Inc., Point Richmond, CA.