Short Contribution

# Evaluation of Archaeological Sourcing Techniques: Reconsidering and Re-Deriving Hughes' Four-Fold Assessment Scheme

Ellery Frahm

Department of Earth Sciences, University of Minnesota

**Correspondence**
Corresponding author;
E-mail: frah0010@umn.edu

**Received**
26 July 2011
**Accepted**
8 December 2011

Scientific editing by Steven Shackley

Published online in Wiley Online Library (wileyonlinelibrary.com).

doi 10.1002/gea.21399

As new analytical techniques are brought to sourcing studies and researchers compile data into multi-laboratory databases, systematic evaluation is essential. The importance of precision and accuracy is clear, but Shackley (2005) also calls for "archaeological accuracy." Hughes (1998) offered a framework to consider precision and accuracy alongside the concepts of reliability and validity. These four concepts can serve as a foundation to evaluate archaeological sourcing data and procedures, but adoption of Hughes' framework has been nearly nonexistent. Unfortunately, Hughes' formulations of reliability and validity are somewhat at odds with their conventional definitions, hindering his framework. Furthermore, the concept of precision has become outdated in analytical circles, and superfluous terms (e.g., replicability) have emerged in the archaeological literature. Here I consider the basis of Hughes' framework and how its four components, when applied consistently by the sourcing community, are best applied to evaluate analytical data and techniques for sourcing.


## INTRODUCTION

With the continuing application of new analytical techniques in archaeological sourcing research (e.g., portable X-ray fluorescence [pXRF] in Craig et al., 2007; Jia et al., 2010; Liritzis, 2008; particle-induced X-ray emission [PIXE] in Lugliè et al., 2007; Rivero-Torres et al., 2008; Torrence et al., 2009; and laser-induced breakdown spectroscopy [LIBS] in Harmon et al., 2009), systematic evaluation is as important as ever. Furthermore, assessment standards are critical as researchers increasingly propose compiling data from numerous laboratories, using a variety of techniques and procedures, into a core repository for sourcing purposes (e.g., the "obsidatabase" advocated by Varoutsikos & Chataigner, 2011). Much like the proverbial bad apple spoiling the whole bunch, the inclusion of low-quality data, if unrecognized, could lead to source-assignment errors, invalidating the entire endeavor. Therefore, a systematic evaluation framework should be adopted before there are serious efforts to compile such sourcing databases.

The need for precision and accuracy of analytical data is evident, but there is also a sense that researchers must be concerned with *something* beyond only precise and accurate data. For example, Shackley (2005) claims that researchers have often made "the mistake of focussing on precision versus the *archaeological accuracy* we seek" in sourcing studies (7; emphasis added). In fact, Hughes (1998) proposes a four-fold framework to consider "archaeological accuracy." Initially suggested for obsidian sourcing, his framework is sound for nearly all archaeological sourcing studies. He contends that, besides precision and accuracy, any assessment should also incorporate both "the concepts of *reliability* and *validity*" (p. 108).

Reliability and validity are hardly new concepts in anthropological research (e.g., Landy, 1965; den Hollander, 1967; Euler, 1967; Brim & Spain, 1974; Pelto & Pelto, 1978; Baker, 1988). Nance (1987, p. 247), for example, declares their importance in archaeology in his chapter "Reliability, Validity, and Quantitative Methods in Archaeology," arguing it is

> . . . difficult to overemphasize the importance of these two concepts in archaeology or any science. They relate to our ability to make meaningful observations about the phenomena we study. It is a universal truth of science that if we cannot measure a phenomenon

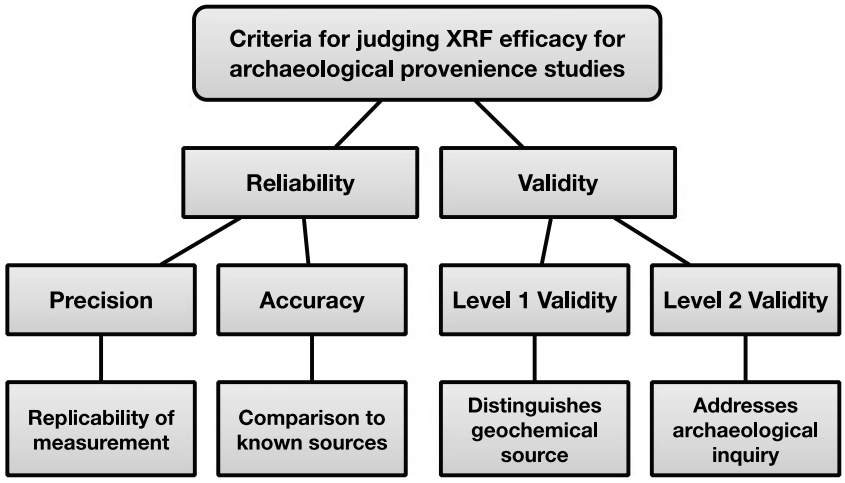


**Figure 1** Hughes' (1998) four-fold sourcing evaluation scheme represented diagrammatically as interpreted by (and redrawn from) Nazaroff et al. (2010). Both the terminology and structure are faithful to the original published figure.

properly, we can never truly understand that phenomenon. Reliability and validity lie at the very heart of the science of prehistory.

These four concepts—precision, accuracy, reliability, and validity—can, if consistently applied, serve as a foundation for evaluating sourcing data and procedures. It should be noted that, while Hughes' sourcing framework was new, the concepts were not. Although they have analogues as far back as Aristotelian thought (Neugebauer, 1969), these concepts and their conceptual pairings (accuracy and precision, reliability and validity) were codified with their modern meanings in the late 19th century by measurement and evaluation theorists.

Since Hughes called for these four concepts to be included at the core of archaeological sourcing studies, use of this framework has been nearly nonexistent. Aside from a few one-off uses of the word "reliability" without defining it (e.g., Bavay et al., 2000; Constantinescu et al., 2002), very few sourcing studies (e.g., Frahm, 2010, in press; Nazaroff et al., 2010) have used this four-concept framework as an evaluation basis. Unfortunately, Hughes' formulations of reliability and validity are somewhat at odds with their conventional definitions in the literature. Consequently, Nazaroff et al. (2010), who use his framework to assess pXRF for Mesoamerican obsidian sourcing, also use his atypical formulations (Figure 1). Furthermore, the concept of precision has become outdated in analytical circles, and superfluous terms (e.g., replicability) have emerged in the recent literature. Additionally, idiosyncratic statements (e.g., "This result indicates that each method is *internally accurate*," Jia et al., 2010, p. 1672, emphasis added) hint at fundamental misconceptions regarding the key concepts.

Here I consider the basis of Hughes' framework and how the four components, when used and reported consistently by the sourcing community, are best applied to evaluate their analytical data and techniques (see Figure 2). Adoption of consistent assessment reporting standards would enable greater confidence in the quality of analytical data and permit the community to focus instead on interpreting sourcing data in terms of human behavior. It should be acknowledged that these issues are relevant to (and have been considered in) other fields, such as geological sourcing of sediments in archaeological (e.g., Woodward et al., 2001) and non-archaeological (e.g., Collins et al., 1997) contexts. Ultimately, this framework is apropos in a wide variety of scientific contexts, but my particular focus here is sourcing archaeological artifacts.

## PRECISION

Precision is the degree to which a series of measurements with the same conditions give identical results (Taylor, 1996). Sufficient precision is essential because archaeological sourcing studies involve measuring geological specimens and artifacts and seeking tight clusters within the data to differentiate raw-material sources and identify any characteristic signals in artifacts. Analytical techniques have inherent limitations in precision, but researchers' choices can affect precision as much as the instruments and specimens (e.g., Long, 1995).

Precision in archaeological sourcing studies is actually a complex issue. Given the large number of specimens needed for sourcing, analytical sessions scheduled over many months or even years are often necessary, meaning that conditions inevitably vary somewhat, even if all procedures remain the same. Most instruments must be newly recalibrated for each analytical session, and the

  

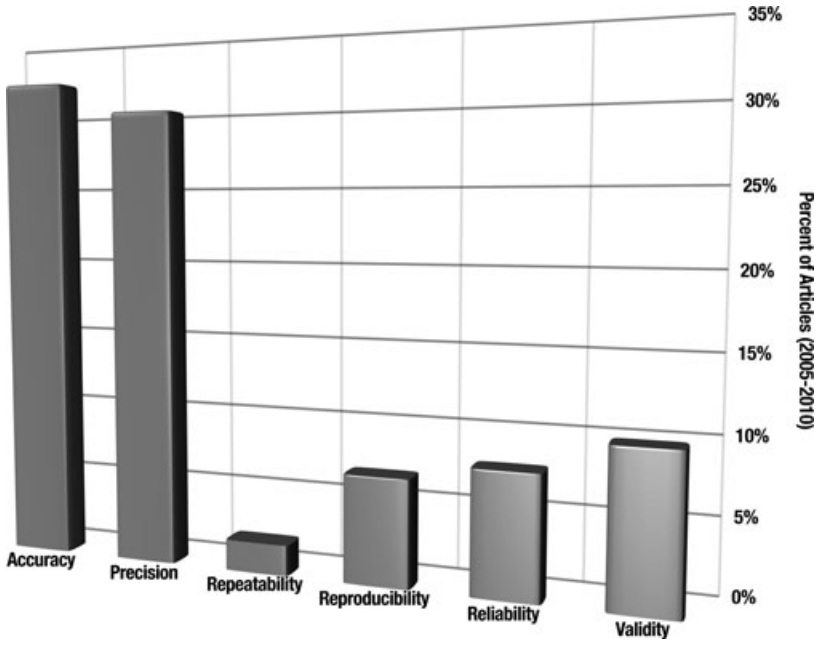


**Figure 2** The results from a survey of 60 articles on archaeological sourcing published from 2005 to 2010. These articles, which used compositional data to source artifacts, were examined for references, even cursory, to the evaluation concepts discussed here. Fewer than one in three articles mentioned accuracy, and the same portion mentioned precision. Fewer than one in ten mentioned validity, and even fewer mentioned reliability or separated precision into the concepts of reproducibility and repeatability.

calibrations inevitably "drift" over time due to environmental, mechanical, and electronic variations (i.e., an instrument and its environment are never identical between two analytical sessions scheduled a month apart). Accordingly, the National Institute of Standards and Technology (NIST), part of the United States Department of Commerce, has not recognized the concept of precision for analytical data for nearly two decades, having divided the concept in two: repeatability and reproducibility (Taylor & Kuyatt, 1994). The former is the agreement between sequential measurements under identical conditions, and the latter is agreement when observers, conditions, or instruments change or after time has passed. Repeatability is generally what most researchers consider to be precision, and the interval that quantifies repeatability of a measurement is also known as its uncertainty (e.g., Pitblado et al., 2008).

It is widely accepted in sourcing studies that precision is determined by taking repeated measurements on a specimen, not as a calibration standard, but as an unknown. Sometimes the specimen is an artifact (e.g., Doelman et al., 2008), geological specimen (e.g., Craig et al., 2007), or, most preferably, international reference standard (e.g., Shackley, 2009). Commonly precision is quantitatively reported as the relative standard deviation for the repeated measurements. It is recommended that NIST's terminology be adopted and that "repeatability" replaces "precision" to describe this concept, so it is clear that the measure applies to a discrete study, that is, a particular instrument with the same observers and conditions.

## ACCURACY

Any particular archaeological sourcing study does not actually need accurate analyses if its data are repeatable and thus internally consistent. Accuracy is essential, though, if one wishes to use the data from one study or laboratory with the data from another. Analytical techniques have no set accuracy, and the literature (e.g., Goldstein et al., 1981) recognizes that researchers' choices, informed by their theoretical and practical "know how" (i.e., *connaissances* and *savoir-faire* in the terminology of anthropologist Pierre Lemonnier), can readily affect accuracy.

Accuracy, as defined by NIST, is the "closeness of the agreement between the result of a measurement" and the actual (i.e., accepted) value of the quantity being measured (Taylor & Kuyatt, 1994). Since accuracy compares measured to actual values, it is nonsensical to have an "internally accurate" analytical technique (Jia et al., 2010:1672). Researchers will often address accuracy simply by displaying their data alongside other published values from another study (e.g., Seelenfreund et al., 2010); however, accuracy can be expressed quantitatively as the percent relative error. More importantly, as explained by Mark and Workman (2003, p. 214), assessing accuracy is challenging "because the usual statement of 'accuracy' compares the result obtained with 'truth' ... 'truth' is usually unknown, making this comparison difficult."

Hughes (1998) offers two principal approaches to accuracy assessment. First, accuracy should be checked against specimens of known composition, preferably at least one reference standard from an internationally

 

recognized organization (e.g., Smithsonian VG-568, United States Geological Service RGM-1, or NIST SRM 278 for obsidian). Second, specimens should be analyzed using two or more analytical techniques for direct comparisons (e.g., Frahm, 2010, in press; LeBourdonnec et al., 2010, Pitblado et al., 2008). Hancock and Carter (2010, p. 245) contend that, "although analytical chemistry is not a democratic process, the agreement of specific elemental concentration data between (among) independent analytical techniques adds credibility to the relative accuracy of their numbers." Both of Hughes' approaches to accuracy assessment are strongly encouraged, especially in conjunction.

Mark and Workman (2003) also advocate "round robin" interlaboratory comparisons to assess accuracy: pieces of a specimen (or specimens) are sent to various laboratories to analyze, and the results are compiled and shared (e.g., Glascock, 1999). It is assumed that, with multiple participating laboratories using different techniques, the resulting mean values represent a good approximation of "true" values. Individual laboratories may, in turn, use these values to assess their own accuracy. A weakness of round robins, though, is that commonly only one or two specimens are sent out and analyzed, so when any inaccuracies are observed, it is difficult to generalize because a systematic error cannot be revealed with just one or two references. Thus, comparative data for multiple specimens are recommended, if possible, for assessing accuracy via such interlaboratory comparisons.

## RELIABILITY

Regarding reliability, Hughes cites a classic definition from Carmines and Zeller (1979): "*reliability* concerns the extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials" (p. 11). From this, Hughes (1998) concludes that reliability in sourcing "involves ... both precision and accuracy" (p. 108). Consider, though, the following few examples from the anthropological literature:

- "*Reliability* is the name given to the degree of reproducibility of a measure. If we were able to make repeated, independent, determinations of a measure, we should hope to obtain values that were close together" (Bartholomew, 1996, p. 24).
- "*Reliability* refers to whether or not you get the same answer by using an instrument to measure something more than once" (Bernard, 1994, p. 38).
- "A *reliable* measure is one that, if administered in the same situation, will provide the same result" (Kuznar, 1997, p. 37).
- "*Reliability* is the degree to which observations of a study are repeatable. A measuring instrument is reliable if it generates consistent observations at two points in time" (Madrigal, 1998, p. 4).

Reliability is synonymous with reproducibility, repeatability, and consistency. Hence reliability involves precision but not accuracy, as Hughes contends. Indeed, Bartholomew (1996) asserts, "Reliability is then equivalent to precision" (p. 24). Others also consider reliability and accuracy distinct concepts (Bernard, 1994, p. 39; Kaplan, 1964, pp. 202–203).

Reliability is actually most analogous to reproducibility, as defined by NIST. Consider the following descriptions of reliability from the literature on evaluation theory:

- "Accordingly, reliability is often interpreted as a kind of intersubjectivity: the agreement of *different observers* on the measures to be assigned in particular cases. But *changes in the circumstance of measurement* other than the identity of the person making the measurement are also involved in reliability" (Kaplan, 1964, p. 200; emphasis added).
- "Reliability concerns the extent to which measurements are repeatable – by the same individual using *different measures* of the same attribute or by *different persons* using the same measure of an attribute" (Nunnally, 1967, p. 172; emphasis added).
- "If *several doctors* use the same thermometer to measure the temperature of the same individual but obtain strikingly dissimilar results, the thermometer is unreliable" (Zeller & Carmines, 1980, p. 6; emphasis added).
- "A measuring instrument is said to be reliable according to the degree to which it generates consistent observations *at two points in time*. Or a measure is reliable to the degree that *two different researchers* using the same instrument on the same sample would generate the same observations" (Bohrnstedt & Knoke, 1988, p. 14; emphasis added).

Reliability requires agreement when observers and conditions change or time has passed. Hence reliability is directly analogous to reproducibility and, thus, is definable quantitatively. It can be quantified using either the relative standard deviation or a test to identify a statistical difference among populations (e.g., Student's *t*-test). Again it is recommended that NIST's terminology be adopted so that "reproducibility" replaces "reliability."

An important implication is that it is difficult for any individual study to determine the reproducibility of an analytical technique. For example, one study is incapable of establishing the reproducibility of portable Raman spectroscopy for obsidian sourcing in a particular region (e.g.,

  

Kelloway et al., 2010). The technique must be tested with other observers, instruments, and conditions over time, and only then may its reproducibility be claimed. See Hancock and Carter (2010) for a recent exploration of this issue with respect to obsidian sourcing.

## VALIDITY

Regarding validity in archaeological sourcing studies, Hughes (1998) contends there are two levels. The first level "concerns the extent to which measurement units are suited to goals of research (i.e., are the units *themselves* valid measures for identifying distinct geochemical varieties of obsidian and for matching artifacts to them?)" (p. 109). One expects that, when he refers to valid "units" for the research, he may mean suitable conceptions of "source," site types, or some other spatiotemporal analysis unit, perhaps even the elements chosen for distinguishing raw-material sources. Instead, Hughes only discusses the units in which the chemical analyses are reported (i.e., percent or parts-per-million rather than instrument-specific units like X-ray counts). His concern is legitimate and has been echoed elsewhere (e.g., Shackley, 2005), but it does not entail validity as traditionally conceived in the evaluation literature.

Hughes' (1998, p. 109) second level "concerns the degree to which geochemical data serve archaeological ends." Sourcing studies assume that raw-material distribution patterns can be interpreted in terms of human behavior, such as exchange and mobility. He criticizes the assumption that obsidian from distant sources constitutes evidence of trade: "archaeologists do not study trade; they study artifacts ... sourcing studies are conducted to inform on such nongeochemical topics, yet they do not speak directly to these issues" (Hughes, 1998, pp. 109–111). Ultimately his core criticism is that the "trade" label is too readily applied to the spatial displacement of raw materials and that exchange, direct procurement, and mobility may look similar archaeologically. His concerns are again legitimate, but this is a broad theory-based criticism regarding archaeological interpretation that has little to do with evaluating an analytical technique for sourcing.

There are, in fact, two types of validity found in the evaluation literature. One equals accuracy while the second type is related to whether one actually is measuring the phenomenon one wishes to measure. Consider the following examples from the literature:

- "For example, if the shots from a well-anchored rifle hit exactly the some location but not the proper target, the targeting rifle is consistent (and hence reliable) but it did not hit the location that it was supposed to hit (and hence it is not valid)" (Zeller & Carmines, 1980, p. 77).
- "If the perforations on a target made by successive shots from a rifle ... are all clustered in the bull's-eye, the rifle is also performing validly" (Ebel, 1965, p. 310).
- "For example, let us assume that a particular yardstick does not equal 36 inches; instead, the yardstick is 40 inches long. Thus, every time this yardstick is used to determine the height of a person (or object), it systematically underestimates height by 4 inches for every 36 inches... This particular yardstick, in short, provides an invalid indication of height" (Carmines & Zeller, 1979, pp. 13–14).

The examples involve a systematic error to which a measurable correction can be applied (i.e., adjustment of the rifle scope using trigonometry, subtracting 4 inches per yard or cutting the yardstick) to become valid. This "quantitative" validity is therefore identical to accuracy.

Contrast this to the type of validity described in the following examples:

- "The number of words in a poem ... would be readily accepted as a valid measure of the length of the poem. It would not, however, be accepted by most poets or literary critics as a valid measure of the literary merit" (Ebel, 1965, p. 310).
- "'Validity' refers to the degree to which scientific observations actually measure or record what they purport to measure ... we all understand the validity of the temperature as measurement by thermometers and the measures of distance we can gauge with yardsticks and rulers" (Pelto & Pelto, 1978, p. 33).
- "Since a kilogram is equivalent to 2.2 pounds, the scale *is* valid since it is in fact measuring the concept it is intended to measure – weight. If the second scale had been measuring percent body fat, then it would have been invalid as a measure of weight" (Bohrnstedt & Knoke, 1988, p. 13).

This is a different type of validity: a conceptual one. It is concerned with whether or not the variables examined actually correspond to the concept one wishes to study. One cannot apply some sort of correction and make an invalid measure valid. Note that, in the last example above, validity is unit-independent: mass in kilograms can be converted into weight in pounds (provided one uses the scale on the surface of the Earth), so both are valid units for a scale (i.e., bathroom scales have a switch to report pounds or kilograms, laboratory scales can be set to report grams or ounces). This highlights that Hughes' concerns about measurement units, while a legitimate

 

issue regarding data reporting, do not involve validity. Just as mass can be converted to weight on the surface of a particular planet, so too can the X-ray counts, for example, from an element be validly converted to percent or parts-per-million for a particular instrument.

A form of conceptual validity separate from accuracy is also suggested by Bernard (1994, pp. 39–40) in his book *Research Methods in Anthropology*:

> What if the spring were not calibrated correctly ... and the scale were off? ... Suppose it turned out that your scale [readings] were always incorrectly lower by 5 pounds ..., then a simple correction formula would be all you'd need in order to feel confident that the data from the instrument were pretty close to the truth ... The data from this instrument are valid (it has already been determined that the scale is measuring weight – exactly what you think it's measuring); the data are reliable (you get the same answer every time you step on it) ... But they are not *accurate* (emphasis in original).

Here validity, reliability, and accuracy are individual concepts, leaving only the conceptual-type validity intact since quantitative validity equals accuracy.

One cannot evaluate the conceptual validity of techniques or data alone. Validity is assessed in light of a particular phenomenon:

> ... it is quite possible for a measuring instrument to be relatively valid for measuring one kind of phenomenon but entirely invalid for assessing other phenomena. Thus, one validates not the measuring instrument itself but the measuring instrument in relation to the purpose for which it is being used (Carmines & Zeller, 1979, p. 16).

Meter sticks and bathroom scales are not inherently valid instruments. Instead, their use to measure particular phenomena must be considered: meter sticks are valid for measuring length, not mass, and bathroom scales are valid for measuring weight, not body fat. Thus accuracy, as conceived here, is a qualitative assessment and a nominal variable.

Hughes' second-level validity evaluates the technique or data directly in light of human behavior, skipping over assessing middle-range theories. Fortunately, Neff (1998, p. 116), in the same volume, mentions an interpretation of validity for ceramic sourcing: "The instrument is a *valid* indicator to the extent that composition really does measure 'source' as a location in geographic space and not some other concept." This formulation of validity is more reasonable for technique assessment: does an analytical technique, when combined with data analysis, discern raw-material sources (i.e., geochemical groups) and assign artifacts to them?

An implication is that a researcher cannot determine the validity of his or her analytical technique and data-analysis procedure independently. The two must interact to produce a source identification for archaeological materials. For example, the validity of XRF data with obsidian and principal components analysis does necessarily not mean, without additional tests, that XRF data with basalts and cluster analysis would be valid. The geographic region (i.e., the numbers and geochemical types of raw-material sources) is another variable in the validity assessment (e.g., Nazaroff et al., 2010), as is proper element selection. It is also assumed that there are no methodological deficiencies (e.g., inadequate sampling of raw-material sources). An analytical technique cannot be considered valid for sourcing without a specific context.

To establish validity, researchers must actually test the success of their procedures by applying them to artifacts from known raw-material sources and showing that the artifacts are attributed to their correct sources. This permits a qualitative, perhaps even subjective, validity assessment to become quantitative. The technique's success rate (e.g., the fraction of artifacts correctly attributed to sources) provides a measure of validity that could otherwise simply be a nominal variable. When a researcher's analytical technique and data analysis assign artifacts to known sources with a high success rate, there may be confidence in their validity (e.g., Craig et al., 2009, Figure 3; Ericson & Glascock, 2004, Figures 7–10; Parish, 2011, Table 2). Statistical approaches to quantifying the validity of artifacts' source assignments have also been

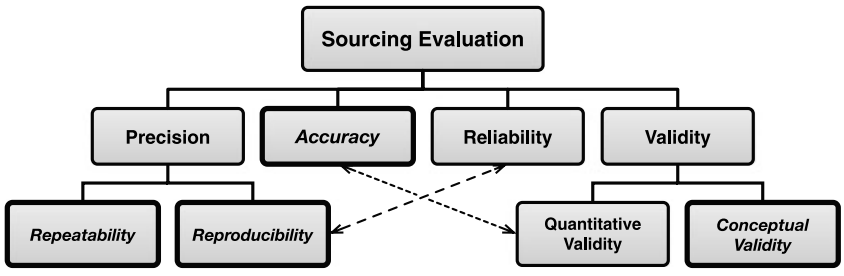


**Figure 3** Hughes' (1998) evaluation scheme re-derived from first principles and hierarchically represented. The equivalencies (i.e., reproducibility and reliability, accuracy and quantitative validity) are shown by dashed lines, and the four core concepts are highlighted in bold.

  

proposed (e.g., Mulholland & Pulford, 2007; Rapp et al., 2000).

## CONCLUSIONS

These four concepts—accuracy, repeatability (often termed precision), reproducibility (equal to reliability), and (conceptual) validity—are discrete, and each is useful for evaluating sourcing data and procedures (Figure 3). This framework is independent of analytical technique and may be applied equally to data from pXRF, PIXE, and other forms of spectrometry. Archaeological sourcing studies should remain consistent in the usage of these concepts and avoid extraneous terms without analytical definitions (e.g., replicability). The result will enable superior inter-study comparisons of new analytical techniques or procedures applied to sourcing research and existing techniques applied to new materials or geographic regions. Consequently, the sourcing community can either establish confidence in new techniques or highlight problems that require additional work. Evaluation standards can lead to greater confidence in data quality and allow us to focus on other issues regarding raw-material use and acquisition.

The journal editors as well as two anonymous reviewers are thanked for their comments and suggestions that led to a stronger version of the final manuscript.

### REFERENCES

Baker, T.L. (1988). Doing social research. New York: McGraw-Hill.

Bartholomew, D.J. (1996). The statistical approach to social measurement. San Diego: Academic Press.

Bavay, L., De Putter, T., Adams, B., Navez, J., & Andre, L. (2000). The origin of obsidian in predynastic and early dynastic upper Egypt. Mitteilungen des Deutschen Archäologischen Instituts, Abteilung Kairo, 56, 5–20.

Bernard, H.R. (1994). Research methods in anthropology: Qualitative and quantitative approaches, 2nd ed. Lanham, MD: AltaMira Press.

Bohrnstedt, G.W., & Knoke, D. (1988). Statistics for social data analysis, 2nd ed. Itasca, IL: F.E. Peacock Publishers.

Brim, J.A., & Spain, D.H. (1974). Research design in anthropology: Paradigms and pragmatics in the testing of hypotheses. New York: Holt, Rinehart, and Winston.

Carmines, E.G., & Zeller, R.A. (1979). Reliability and validity assessment. Quantitative applications in the social sciences. London: Sage Publications.

Collins, A.L., Walling, D.E., & Leeks, G.J.L. (1997). Source type ascription for fluvial suspended sediment based on a quantitative composite fingerprinting technique. Catena, 29, 1–27.

Constantinescu, B., Bugoi, R., & Sziki, G. (2002). Obsidian provenance studies of transylvania's Neolithic otols using PIXE, Micro-PIXE and XRF. Nuclear Instruments and Methods in Physics Research B, 189, 373–377.

Craig, N., Speakman, R., Popelka-Filcoff, R., Glascock, M., Robertson, J., Shackley, M.S., & Aldenderfer, M. (2007). Comparison of XRF and PXRF for Analysis of archaeological obsidian from Southern Perú. Journal of Archaeological Science, 34(12), 2012–2024.

Craig, N., Speakman, R.J., Popelka-Filcoff, R.S., Aldenderfer, M., Blanco, L.F., Vega, M.B., Glascock, M. & Stanish, C. (2009). Macusani obsidian from southern Peru: A characterization of its elemental composition with a demonstration of its ancient use. Journal of Archaeological Science, 37(3), 569–576.

den Hollander, A.N.J. (1967). Social description: The problems of reliability and validity. In D.G. Jongmans, & P.C.W. Gutkind (Eds.), Anthropologists in the field (pp. 1–34). Assen: Van Gorcum.

Doelman, T., Torrence, R., Popov, V., Ionescu, M., Kluyev, N., Sleptsov, I., Pantyukhina, I., White, P., & Clements, M. (2008). Source selectivity: An assessment of volcanic glass sources in the Southern Primorye Region, Far East Russia. Geoarchaeology, 23(2), 243–273.

Ericson, J.E., & Glascock, M.D. (2004). Subsource characterization: Obsidian Utilization of subsources of the Coso Volcanic Field, Coso Junction, California, USA. Geoarchaeology, 19(8), 779–805.

Ebel, R.L. (1965). Measuring educational achievement. Englewood Cliffs, NJ: Prentice-Hall.

Euler, R.C. (1967). Ethnographic methodology: A tri-chronic study in culture change. Informant reliability and validity from the Southern Paiute. In C.L. Riley & W.W. Taylor (Eds.), American historical anthropology: Essays in honor of Leslie Spier (pp. 61–67). Carbondale: Southern Illinois University Press.

Frahm, E. (2010). The Bronze-age obsidian industry at Tell Mozan (Ancient Urkesh), Syria: Redeveloping electron microprobe analysis for 21st-century sourcing research and the implications for obsidian use and exchange in Northern Mesopotamia after the Neolithic. Ph.D. dissertation, Department of Anthropology, University of Minnesota. Available online at University of Minnesota's Digital Conservancy: http://purl.umn.edu/99753. Accessed 1 July 2011.

Frahm, E. (in press). Non-destructive sourcing of Bronze-Age near Eastern obsidian artifacts: Redeveloping and reassessing electron microprobe analysis for obsidian sourcing. Archaeometry, in press. doi: 10.1111/j.1475-4754.2011.00648.x

Glascock, M.D. (1999). An inter-laboratory comparison of element compositions for two obsidian sources. International Association for Obsidian Studies Bulletin, 23, 13–25.

Goldstein, J.I., Newberry, D.E., Echlin, P., Joy, D.C., Fiori, C., & Lifshin, E. (1981). Scanning electron microscopy and X-ray microanalysis. New York: Plenum Press.

 

Hancock, R.G.V., & Carter, T. (2010). How reliable are our published archaeometric analyses? Effects of analytical techniques through time on the elemental analysis of obsidians. Journal of Archaeological Science, 37(2), 243–250.

Harmon, R.S., Gottfried, J.L., Remus, J., Baron, D., Draucker, A., & Yohe, R. (2009). Provenance determination of California obsidian sources by laser-induced breakdown spectroscopy. Geological Society of America Annual Meeting, Portland, OR. Abstract in Geological Society of America Abstracts with Programs, 41(7), 554. Retrieved fromhttp://gsa.confex.com/gsa/2009AM/finalprogram/abstract˙160229.htm. Accessed 1 July 2011.

Hughes, R.E. (1998). On Reliability, validity, and scale in obsidian sourcing research. In A.F. Ramenofsky & A. Steffen (Eds.), Unit issues in archaeology: Measuring time, space, and material (pp. 103–114). Salt Lake City: University of Utah Press.

Jia, P.W., Doelman, T., Chen, C., Zhao, H., Lin, S., Torrence, R., & Glascock, M.D. (2010). Moving sources: A preliminary study of volcanic glass artifact distributions in northeast China using PXRF. Journal of Archaeological Science, 37, 1670–1677.

Kaplan, A. (1964). The conduct of inquiry: Methodology for behavioral science. San Francisco: Chandler Publishing Company.

Kelloway, S.J., Kononenko, N., Torrence, R., & Carter, E.A. (2010). Assessing the viability of portable Raman spectroscopy for determining the geological source of obsidian. Vibrational Spectroscopy, 53, 88–96.

Kuznar, L.A. (1997). Reclaiming a scientific anthropology. Lanham, MD: AltaMira Press.

Landy, D. (1965). Tropical childhood: Cultural transmission and learning in a rural Puerto Rican village. New York: Harper & Row.

LeBourdonnec, F.-X., Bontempi, J.-M., Marini, M., Mazet, S., Neuville, P.F., Poupeau, G., & Sicurani, J. (2010). SEM-EDS characterization of western Mediterranean obsidians and the Neolithic site of A Fuata (Corsica). Journal of Archaeological Science, 37(1), 92–106.

Liritzis, I. (2008). Assessment of Aegean obsidian sources by a portable ED-XRF analyser: Grouping, provenance and accuracy. In Y. Facorellis, N. Zacharias, & K. Polikreti (Eds.), Proceedings of the 4th Symposium of the Hellenic Society for Archaeometry (pp. 399–406). BAR International Series 1746. Oxford: Tempus Reparatum.

Long, J.V.P. (1995). Microanalysis from 1950 to the 1990s. In P.J. Potts, J.F. Bowles, S.J.B. Reed & R. Cave (Eds.), Microprobe techniques in the earth sciences (pp. 1–48). London: Chapman & Hall.

Lugliè, C., Le Bourdonnec, F.-X., Poupeau, G., Atzeni, E., Dubernet, S. Moretto, P., & Serani, L. (2007). Early Neolithic obsidians in Sardinia (western Mediterranean): The Su Carroppu case. Journal of Archaeological Science, 34, 428–439.

Madrigal, L. (1998). Statistics for anthropology. Cambridge: Cambridge University Press.

Mark, H., & Workman, J. (2003). Statistics in spectroscopy, 2nd ed. San Diego: Elsevier Academic Press.

Mulholland, S.C., & Pulford, M.H. (2007). Trace-element analysis of native copper: The view from northern Minnesota, USA. Geoarchaeology, 22(1), 67–84.

Nance, J.D. (1987). Reliability, validity, and quantitative methods in archaeology. In M.S. Aldenderfer (Ed.), Quantitative research in archaeology: Progress and prospects (pp. 244–293). Newberry Park: Sage Publications.

Nazaroff, A.J., Prufer, K.M., & Drake, B.L. (2010). Assessing the applicability of portable X-ray fluorescence spectrometry for obsidian provenance research in the Maya Lowlands. Journal of Archaeological Science, 37, 885–895.

Neff, H. (1998). Units in chemistry-based ceramic provenance investigations. In A.F. Ramenofsky & A. Steffen (Eds.), Unit issues in archaeology: Measuring time, space, and material (pp. 115–127). Salt Lake City: University of Utah Press.

Neugebauer, O. (1969). The exact sciences in antiquity. New York: Courier Dover Publications.

Nunnally, J.C. (1967). Psychometric theory. New York: McGraw Hill.

Parish, R.M. (2011). The application of visible/near-infrared reflectance (VNIR) spectroscopy to chert: A case study from the Dover Quarry Sites, Tennessee. Geoarchaeology, 26(3), 420–439.

Pelto, P.J., & Pelto, G.H. (1978). Anthropological research: The structure of inquiry. 2nd ed. Cambridge: Cambridge University Press.

Pitblado, B.L., Dehler, C., Neff, H., & Nelson, S.T. (2008). Pilot study experiments sourcing quartzite, Gunnison Basin, Colorado. Geoarchaeology, 23(6), 742–778.

Rapp, G., Jr., Allert, J., Vitali, V., Jing, Z., & Henrickson, E. (2000). Determining geologic sources of artifact copper: Source characterization using trace element patterns. Lanham: University Press of America.

Rivero-Torres, S., Calligaro, T., Tenorio, D., & Jiménez-Reyes, M. (2008). Characterization of archaeological obsidians from Lagartero, Chiapas Mexico by PIXE. Journal of Archaeological Science, 35, 3168–3171.

Seelenfreund, A., Pino, M., Glascock, M.D., Sinclaire, C., Miranda, P., Pasten, D., Cancino, S., Dinator, M.I., & Morales, J.R. (2010). Morphological and geochemical analysis of the Laguna Blanca/Zapaleri obsidian source in the Atacama Puna. Geoarchaeology, 25(2), 245–263.

Shackley, M.S. (2005). Obsidian: Geology and archaeology in the North American southwest. Tucson: University of Arizona.

Shackley, M.S. (2009). The topaz basin archaeological obsidian source in the transition zone of central Arizona. Geoarchaeology, 24(3), 336–347.

Taylor, B.N., & Kuyatt, C.E. (1994). Guidelines for evaluating and expressing the uncertainty of NIST measurement

results. NIST Technical Note 1297. Retrieved from http://www.nist.gov/pml/pubs/tn1297/index.cfm. Accessed 1 July 2011.

Taylor, J.R. (1996). Introduction to error analysis: The study of uncertainties in physical measurements. 2nd ed. Sausalito: University Science Books.

Torrence, R., Swadling, P., Kononenko, N., Ambrose, W., Rath, P., & Glascock, M.D. (2009). Mid-holocene social interaction in Melanesia: New evidence from hammer-dressed obsidian stemmed tools. Asian Perspectives, 48(1), 119–148.

Varoutsikos, B., & Chataigner, C. (2011). Obsidatabase project: Collecting and organizing data on prehistoric Caucasian and near eastern obsidian. Society for American Archaeology 76th Annual Meeting, Sacramento, CA. Abstract available online: Abstracts of the SAA 76th Annual Meeting: http://www.saa.org/Portals/0/SAA/Meetings/2011%20Abstracts/Abstracts.pdf. Accessed 1 July 2011.

Woodward, J.C., Hamlin, R.H.B., Macklin, M.G., Karkanas, P., & Kotjabopoulou, E. (2001). Quantitative sourcing of slackwater deposits at Boila rockshelter: A record of lateglacial flooding and Palaeolithic settlement in the Pindus Mountains, northwest Greece. Geoarchaeology, 16(5), 501–536.

Zeller, R.A., & Carmines, E.G. (1980). Measurement in the social sciences: The link between theory and data. London: Cambridge University Press.

 