REPORTS

METRIC DATA IN ARCHAEOLOGY:
A STUDY OF INTRA-ANALYST AND INTER-ANALYST VARIATION

R. Lee Lyman and Todd L. VanPool

Metric data are regularly presented, analyzed, and compared. Despite acknowledgment that metric data can vary both when collected by one observer and when collected by multiple observers, few studies of these sources of variation in archaeological metric data have been undertaken. Intra-observer and inter-observer measurement errors are examined across four dimensions of 23 modern bighorn sheep (Ovis canadensis) astragalus and five dimensions of 30 specimens of stone projectile point representing 4 culture-historical (temporal) types. Statistical and graphical analyses indicate that measuring dimensions of the same specimens multiple times facilitates determination of dimensions that can be readily and reliably measured and serves to screen data for data recording errors and for dimensions that may be subject to high levels of intra-observer and inter-observer variation.

Con regularidad se presenta, se analiza y compara los datos métricos. A pesar del reconocimiento que los datos métricos pueden variar cuando un observador los colecta y cuando varios observadores los colectan. Han hecho pocos estudios de estas fuentes de variación en los datos métricos arqueológicos. Se examinan los errores en medidas de intra-observador e inter-observador a través de cuatro dimensiones de astrágalo moderno de 23 bueyes de cornudo (Ovis canadensis) y cinco dimensiones de 30 especímenes del punto de proyectil de piedra que representan 4 tipos cultural-históricos (temporales). Análisis estadístico y gráfico indica que al medir las dimensiones de las mismas especímenes varias veces se facilita la determinación de las dimensiones que pueden ser fácilmente y correctamente medidas y sirve para examinar los datos para errores de la grabación de los datos y para las dimensiones que pueden ser sujeto a niveles altos de variación entre intra-observador y inter-observador.

Data from series which are taken at different times, under different conditions, and especially from series which are collected and measured by different persons, cannot be regarded as comparable with one another [F. S. Summer 1927].

Archaeology is a comparative science. What this means of course is that our analyses are based on measurement, defined as a procedure for assigning a descriptive label derived from an agreed upon set of units to an item according to a set of rules. Archaeologists measure attributes of bones, stones, pots, houses, mounds, strata, seeds, walls, temper, paint, iconography, and enumerable other things. Learning what to measure and how to do so is a primary topic of archaeological introductory texts, and is the career focus of many archaeologists who have developed and refined particular measurement methods (e.g., radiocarbon dating, bone chemistry, tree ring dating, obsidian sourcing). We are being only slightly facetious when we observe that archaeological fame (if not necessarily fortune) will be the reward of anyone, who identifies a measurement that archaeologists should record but have not, or who develops a better way to measure a variable that archaeologists consider important.

Because of the importance of measurements in archaeological analysis, the process of taking measurements has on occasion been a significant source of contention. This is especially true for those who focus on obtaining and interpreting specific types of measurements. For example, the proper mea-
measurements of lithic usewear has produced extremely detailed and at times heated exchanges (e.g., Bamforth 1988; Hurcombe 1988; Newcomer et al. 1986; 1988). Chemical sourcing is another area where the empirical and theoretical structure of measurement has been a topic of interest (Hughes 1998; Shackley 1998).

Archaeologists generally indicate why they take certain measurements, but seldom discuss how they take measurements (e.g., Rice 1996; Shott 1994). As a result, the link between measurements and archaeological analyses is underdeveloped and analysts do not consider the process of measuring except in superficial ways (e.g., which tool is best for measuring sherd thicknesses, where best can concern anyone or more of several properties), if at all. Yet the process of generating measurement data directly impacts all aspects of archaeological work and thus demands critical scrutiny.

Archaeologists regularly measure metric attributes of artifacts to distinguish types (e.g., Benfer 1967; Thomas 1981), detect functional differences (e.g., Shott 1997; Thomas 1978), track trends in shifting central tendencies (e.g., Braun 1987; Eerkens and Lipo 2005), detect tool maintenance (e.g., Hoffman 1989) or craft specialization (e.g., Crown 1995), and monitor social interaction (e.g., Kay 1975). Similarly, zooarchaeologists measure metric attributes of bones and teeth to identify the taxon represented by a specimen (e.g., Payne 1969), to estimate animal size, meat weight, or biomass (e.g., Emerson 1978; Noodle 1973; Reitz and Ruff 1994), and to detect chronoclinal variation (e.g., Ducos and Horwitz 1977; Klein 1991; Purdue 1989), domestication (e.g., Grigson 1989; Zeder and Hesse 2000), or both (e.g., Davis 1981). Indices have been designed to allow comparison of sets of metric data (reviewed by Meadow 1999), and researchers regularly utilize metric data generated by others in new analyses (e.g., Eerkens and Lipo 2005) or compare metric data they have generated with those generated by others for different analytical purposes (e.g., Drews 1990). Problems in the measuring process will undermine all of these endeavors, yet few synoptic reviews of measurement issues have been published (see Albarella 2002 and Boessneck and Driesch 1978 for notable but terse exceptions). Given archaeology’s long-proclaimed desire to be scientific, and the centrality of measurement to attaining that desire, we find it curious that the critical literature on comparing metric data is limited. We are, for example, aware of only three previous critical studies of archaeological metric data (Fish 1978, 1979; Gnaden and Holdaway 2000; Wilmshurst and Roberts 1978).

In this paper we discuss errors in the generation of ratio-scale size or metric data (e.g., length, width, thickness) and their impact on subsequent data comparisons. The issues we raise (as well as our conclusions) are applicable to data measured at other scales too. Qualitative or nominal-scale data are often recorded by archaeologists, but we do not consider this kind of data here for two reasons. First, qualitative data are categorical, or typological, to use a term familiar to archaeologists. There are already several studies of intra- and inter-observer variation in categorizing or typing artifacts (e.g., Beck and Jones 1989; Boyd 1987; Whittaker et al. 1998); we cannot improve on them. Second, metric data have a mystery about them: Because they are numerical, they are scientific and objective; that is, they are not influenced by observer fallibility. Thus, metric data have seldom been subjected to rigorous evaluation. We rectify that lacunae here.

To explore and illustrate the severity of measurement problems for comparative studies, our analyses focus on repeated ("replicate") measurement of the same specimens taken by the same and by different observers. This is not the typical archaeological case, which instead focuses on the measurement of different assemblages by one or more researchers who then either compare the assemblages with each other or collapse them into a larger data set. Still, our study reflects the likely sources of measurement error and their analytic significance, as well as illustrates rarely used analytical techniques, and exemplifies ways that measurement data generated for different assemblages by different analysts might be evaluated for comparability.

We begin by introducing the issue of mensuration (the act of measuring), focusing especially on definitions that clarify the process and potential difficulties. We then present analyses of measurements of the same dimensions of a sample of bones and a sample of stone projectile points to illustrate how inter-observer and intra-observer measurement error can be evaluated. We conclude by outlining procedures to detect significant mensuration problems and to potentially mitigate them.
Mensuration and Archaelogical Data

In what is still one of the most thorough statements on mensuration, fifty years ago Simpson et al. (1960:21–22) argued that measurements should be related to a research problem in a logical manner, they should be adequately reported, and they should be well defined, comparable, and standardized. Measurements are tools for one’s analysis, thus one’s theoretical perspective identifies which measurements are important, and ultimately the form those measurements must take (Kelley and Hanner 1988:362–368). Here we are not concerned with the theoretical validity or utility of the measurements and dimensions we describe but rather focus on assessing several kinds of measurement error.

Regardless of one’s favored archaeological theory, several factors are consistently important when evaluating the appropriateness of measurements as research tools. To begin with, the term measurement itself has two meanings; it denotes both the act of measuring and a documented record of a dimension. Problems with mensuration can be present in both; the technique used to measure specimens can be inaccurate (e.g., produces incorrect results; see Dibble and Bernard 1980 for an exemplary study) and a single measurement can be inaccurately documented (e.g., coding error when recording a particular measurement). The concepts of accuracy and precision are central to understanding and controlling for such difficulties.

Accuracy concerns how close a measurement is to the true value. In many analyses, perfect accuracy is unnecessary if not unattainable given limitations of the tools we use to measure an attribute (e.g., the accuracy of GPS units typically is not perfect at the scale of the nearest centimeter). What is important to worry about is whether inaccuracy is sufficient to create a significant yet false difference (or false similarity). Precision is the closeness of repeated measurements of the same specimen. Precise results do not differ greatly, even if they are inaccurate. Accuracy and precision do not necessarily correspond with one another. A scale that consistently reports a value that is 5 gm heavier than the true weight of individual specimens has poor accuracy but great precision. Likewise, a broken scale that has poor precision can, on occasion, provide accurate results.

The accuracy and precision of measurements can be evaluated in a variety of ways. Accuracy can be determined by study of replicate measurements taken using different tools by different people. In contrast, consistent results, on the one hand, suggest that individual measurements are accurate, although this will not be true if there is a systematic bias that is uniform among all of the recorders and their measurement tools. Repeating measurements of a sample by the same researcher using the same tool, on the other hand, allows evaluation of precision; precise measurements will be consistent among measurement sessions. We emphasize that there is no way to evaluate precision without repeated measurement of the same object. Perfect precision is unlikely if not impossible in most cases, but the degree of imprecision can be quantified. Minor differences between measurements may not be significant. If significant differences are present, then the measurement procedure (measurement technique, data recording procedure, tool used, conditions under which measurements are taken) should be improved. The degree of acceptable precision and accuracy should be decided before measurement has begun and must be evaluated empirically throughout the measurement process.

A related issue is the resolution of measurement. A measurement taken to the nearest .01 mm provides greater resolution than measuring to the nearest .1 mm. In general, as resolution increases, precision decreases while accuracy increases, up to the limits of the measuring tool (Sumner 1927). Measurements that exceed the tool’s resolution limits will provide false resolution, implying greater accuracy than there actually is. An analytically useful resolution level depends on one’s research question, but a good rule of thumb is to create measurements in which the number of unit steps from the smallest to the largest measurement is between 30 and 300 (VanPool and I. Leonard 2009). When measuring an assemblage of projectile points that vary in length between 1.2 cm and 5.7 cm, measuring to the nearest centimeter (which produces 6 unit steps) likely will obscure variation that is analytically useful; measuring to the nearest .01 cm (which produces 450 unit steps) will be time consuming and produce data at a higher resolution than required for most analyses. But measuring to the nearest .1 cm (which produces 45 unit steps) likely
will balance the analytic requirements of the data. We said "likely will balance" because some analyses require less resolution (e.g., using Thomas's [1978] 2 cm demarcation to differentiate between atlatl dart points and arrow points based on their shoulder width) requires only two unit steps (<2 cm and >2 cm) and other analyses may require more (e.g., room corners of prehistoric Southwestern Pueblos will likely require more than 300 unit steps between the smallest and largest structures).

Returning to Simpson et al.'s (1960:21-22) criteria, measurements should be comparable and standardized. Comparability concerns the mechanics of measurement and relates to accuracy, precision, and resolution. In order for a series of measurements of a dimension to be comparable, each measurement must be taken in the same way as every other one. Changing measurement tools, especially if they differ in their potential resolution and the way that they are used, will generally decrease precision and create problems with accuracy (Dibble and Bernard 1980). Calipers must, for example, be held in the same way and applied to or oriented relative to each specimen in the same manner to produce comparable data. Changing measurement instruments (calipers to dividers to rulers to tape measure) will likely result in low comparability, low precision and shifting levels of accuracy and resolution.

Standardization concerns the precise definition of the dimension measured. Simpson et al. (1960:23) note that the dimension "length" of a mammal tooth has been applied at least six different ways, indicating this dimension is not standardized. Differing definitions of dimensions may be useful in different analyses, but comparisons within and between data sets require dimensions that are standardized relative to each other (Simpson et al. 1960:22). Toward that end, archaeologists have devoted some effort to standardizing the dimensions that might be measured on lithic projectile points (e.g., Boaz 1984; Thomas 1981), ceramics (Rice 1996), and other artifacts, and zooarchaeologists have done the same with respect to the dimensions of bones and teeth (Driese 1976). Standardization should make clear what investigator A means when he/she labels a dimension "axial length" or "distal breadth," but standardized dimensions alone do not ensure consistency within and between data sets because of the other mensuration issues (e.g., comparability, accuracy, resolution).

A final important variable is the measurability of a dimension. Some dimensions are quite measurable, meaning they are precisely defined using easily recognized criteria and can be easily taken in instance after instance, whereas other dimensions are difficult to measure (Driese 1976:6, 11). Measurability directly impacts the standardization, comparability, accuracy, and precision, given that easily measured dimensions can be more precisely and accurately measured. With a linear dimension, measurability depends on the definition of the points or landmarks defining the dimension and between which a measurement is taken. To standardize the measurement of a dimension across multiple specimens, the two landmarks must be explicitly specified and easily relocated. Landmarks that cannot be consistently located on specimens after specimen, perhaps because of vague definition, create error (Gavan 1950:147). The distance between Seattle, Washington, and Albuquerque, New Mexico, is easily measured, but results may be inconsistent because the landmarks making up the twin ends of the linear distance are ambiguous. Are they the city limits, the geographic centers of the cities (defined how?), the city courthouses? A common way to determine whether two landmarks are easily relocated is to measure a dimension of a specimen or small set of specimens multiple times and compare the results. If the repeated measurements differ considerably, then a possible cause for the lack of precision is poorly defined landmarks that cannot be reliably located on specimen after specimen.

Study of inter-observer variation is important given that problems of mensuration may be acute when using multiple data sets generated by different researchers (Gnaden and Holdaway 2000; Wilmansen and Roberts 1978). Yet the use of multiple data sets generated by different observers is becoming increasingly common as the amount of archaeological data skyrockets and electronic means of communicating data make it easier for researchers to access and analyze previously existing data sets. Analysts may assume that inter-observer variation is of insignificant magnitude, but previous research indicates that would be unwise. Sumner (1927) found as much as a 6 percent difference between means of measurements of mice carcasses taken by different observers. Fish (1978, 1979) indicated a difference between lithic flakes. With much as a 10 percent measurements observers, but were statistically different. The measurement errors taken by one observer variate, repeated measurements use data from a sample of lithic flakes and electronic data noting that a ran may have experienced but may have relatively low error. The variation may be spurious, and mental problems in that some aspec- sequence breaks A blunder can happen (e.g., I am reading one's centimeter reading
ues (e.g., comparability) is the measurability of observations are quite meticulously defined using d can be easily taken on regardless of other dimensions (e.g., 1976:6, 11). Mea-

standardization, conception, given that easily more precisely and near dimension, mea-

sion is important definition of the points tension and between 1. To standardize the across multiple spec-

ies be explicitly spec-
mard that cannot remain after specimen, inition, create error ice between Seattle, ie, New Mexico, is may be inconsistent up the twin ends of us. Are they the city of the cities (defined A common way to arks are easily rela-

tion is important tion may be acute generated by difference Holdaway 2000; at the use of mul-
ter observers is as the amount of ts and electronic make it easier for e previously exist-

is that would be much as a 6 per-

of measurements at observers. Fish

(1978, 1979) found as much as a 3–4 percent difference between observers measuring attributes of lithic flakes. Wilmsen and Roberts (1978) found as much as a 10 percent difference between means of measurements of lithic flakes taken by multiple observers, but argued that overall the differences were statistically insignificant.

What factors influence variation between measurement episodes, whether measurements were taken by one observer or by multiple observers? Analyst error is likely a more significant problem at the end of a hot day excavating when compared to the same individual working in an air conditioned laboratory, some measurement tools may be prone to error when compared to other tools, and inexperienced and poorly-trained people will likely produce less accurate and less precise measurements relative to others, to name a few. These factors underscore that there are potential problems even when using metric data collected by one observer.

Here we explore inter-observer variation, intra-

observer variation, and measurement error across repeated measurements of the same specimens. We use data from a sample of bones and data from a sample of lithic projectile points. We explore variation in replicate measurements, and based on that variation we suggest how measurement error might be detected and perhaps minimized in other, more typical archaeological analyses.

Evaluating Inter- and Intra-Observer Variation in Linear Measurements

Three general "categories" of errors are considered: blunders, bias, and random errors. Daniels (1972) summarized some of the problems with using metric data from a conceptual perspective, noting that a random error should not create patterns, gross errors or blunders will likely be rare but may have large effects, and (systematic) bias will create spurious patterns. Blunders are fundamentally problems of accuracy and comparability in that some aspect of a standardized measurement sequence breaks down causing data to be incorrect. A blunder can be as simple as transposing numbers (e.g., 1 is recorded as 10), misplacing decimal places (e.g., 1.00 is reported as 10.0), or misreading one's tool (e.g., confusing the inch and centimeter readings on a tape measure). Blunders are not the result of mechanical problems with one's tool, or lack of standardized measurements, although a poorly standardized measurement without clear demarcations may be characterized by more blunders than other measurements. Instead, blunders reflect problems caused by the imperfections of the analyst. Today, blunders resulting from coding or transcription errors can be avoided by using calipers linked directly to a computer. Also, some software programs reduce errors of various sorts (Gnudt and Holdaway 2000:743).

Bias, defined as systematic differences in measurements among researchers, can be produced by various sources. Some tools may be biased relative to one another. This is typically a problem with accuracy, in that some (or all) of the tools produce measurements that are consistently too high or too low relative to an accurate measurement. Although extremely inaccurate measurements may be obviously incorrect, if the measurement tool is precise, problems with accuracy may not be evident until comparable data generated with other tools are available (e.g., the bias in uncalibrated radiocarbon dates was poorly understood until data derived from dendrochronology and other methods were available for comparison). Bias can also be produced by problems with comparability and standardization, in that researchers may use different tools in different ways to measure different attributes designated by the same name. Flake length for example could be considered the longest dimension OR the distance from the platform to the most distant detachment point parallel to the direction of force application. Although these measurements might produce the same measurement for some flakes, they will be different for other flakes. Importantly, the differences will be consistent based on the morphology of the flakes, thereby creating bias and incomparable data for some flakes.

Random variation is typically the least significant source of variation within and between data sets and is a reflection of the precision limits of a tool and the researcher, the standardization of a measurement, and the resolution of measurement. There are always limits to precision, simply because of the nature of measurement tools and the imperfections of humans. For example, a person using a set of manual calipers with a measurement dial must make decisions as to which integer to assign when the dial's indicator is midway between
values. There will be variation in this decision making process, especially given that subtle differences in the way the calipers are held and in the way the measurement is taken can lead to slightly different measurements of the same attribute of a specimen. Further, attributes that are poorly standardized may be inconsistently measured. The result of these problems is that there is variation in measurements that is not systematically larger or smaller than the true value.

Non-systematic, random errors typically have less influence on comparability than biased results because they tend not to change the average values significantly. However, comparisons of variation using quantitative methods such as the coefficient of variation can be influenced by differences in precision. Further, problems with bias and random variation can become more or less significant as the resolution of measurement changes. Random errors in measurements taken to the nearest .01 mm may not be significant when the resolution is shifted to the nearest cm. By the same token, the significance of biased variation may be reduced as the resolution is decreased.

Materials and Methods

Lyman (2009) recently measured two dimensions (greatest lateral length; distal breadth) of one astragalus from each of 60+ individual modern bighorn sheep (Ovis canadensis) astragali (see Driesch [1976] for definitions of dimensions); these are hereafter referred to as Lyman1. He then acquired a copy of an unpublished thesis that contained measurements of these dimensions from 23 of the same specimens (Lawler 1992). Both sets of measurements were taken without either individual knowing of the other’s work. Lyman subsequently (a month after the first set of measurements was taken) again measured the two dimensions (hereafter referred to as Lyman2) to evaluate intra-observer variation, and two other dimensions measured by Lawler (medial length, lateral depth) to evaluate inter-observer variation. Lawler (1992) recorded measurements to the nearest tenth of a millimeter; Lyman recorded measurements to the nearest two hundredths of a millimeter and rounded to the nearest tenth of millimeter for comparison with Lawler’s data.

The second series of data are from thirty complete projectile points recovered from eastern Washington State that represent four point types as defined by archaeologists working in the area (Desert Side Notch [n = 7], Rosegate [n = 10], Gypsum 4 [n = 5], Martis [n = 8]; type names after Jennings [1986]). Five linear dimensions were measured on each: maximum length, axial length (less than maximum length on specimens with a concave base), neck width, shoulder width, and thickness. Lyman and VanPool measured all of the points twice, with at least a month passing between the measurement episodes (Lyman3 and Lyman4; VanPool1 and VanPool2). Other than agreeing on the specimens to be measured and on the measurements to be taken, no other discussion of measuring protocol took place between Lyman and VanPool. Both of Lyman’s data sets and VanPool’s were taken using the same pair of spreading dial calipers. Lyman recorded measurements to the nearest .02 mm whereas VanPool recorded measurements to the nearest .001 cm. VanPool2 was taken using a pair of Fisher Scientific digital spreading calipers with measurements taken to the nearest .01 mm.

Physical anthropologists, particularly those interested in anthropometrics, and assorted human-health professionals have explored measurement error in great detail and developed a set of graphing techniques and descriptive statistics that reveal aspects of precision, reliability, intra-observer variation, inter-observer difference, and the like (e.g., Adams and Byrd 2002; Altman and Bland 1983; Bland and Altman 1986; Heathcote 1981; Himes 1989; Jamison and Ward 1993; Jamison and Zegura 1974; Mueller and Martorell 1988; Pederson and Gore 1996; Ulijaszek and Kerr 1999; Ulijaszek and Lourie 1994; Uttermohle and Zegura 1982; Ward and Jamison 1993; Weinberg et al. 2005). Here we adopt and use some of these methods. Because some of them will likely be unfamiliar to archaeologists, we briefly describe all methods we use.

Basics

Replicate measurement sets are compared in terms of several of their properties; that is, several kinds of similarity of measurement sets are evaluated. Comparison of central tendencies involves comparison of the means or averages of sets of replicate measurements; if the means are not significantly different, then it can be argued that the measurements of (replicate measurement error was compared in terms or spread of the coefficients of variation between measurement (T are described) behavior are being measured similar unless the variation in one replicate measurement (and should be) is the reliability of using a measure.

Graphs

A graph of data is a visual way to bivariate plot the relationship between dimensions again specimens should the diagonal (the value of replicates is measured at 0.5 the equality 1983; Bland and near the diagonally determine the value of replicates if the pair, a technique calculated by Blad and Ho refer to as significant correlation and data agreement between measurement highlights outliers between replicates suggests system between replicate and should be cl (Altman and Bla}
covered from eastern four point types as working in the area, osagate \( n = 10 \).GYp-type names after Jenkins. Dimensions were in length, axial length on specimens with a shoulder width, and of measured all of the 10th from passing between Lyman3 and Lyman4; her than agreeing on ed and on the means of discussion of mean between Lyman and ta sets and VanPool 1 air of spreading dial measurements to the Pool recorded meal cm. VanPool2 was antidigital spreadments taken to the near, particularly those that and assorted human-shoulder widths from the same population (replicate measurements do!) and that measurement error was random. Replicate sets are also compared in terms of measures of the dispersion or spread of the values constituting each set using coefficients of variation (CVs) and in terms of the between sets using the technical error of measurement (TEM) and the related \%TEM (both are described below). Because the same specimens are being measured, two compared CVs should be similar unless there is a large amount of measurement error in one or both sets. The TEM and \%TEM estimate how much variation is a result of differences between specimens in a sample and how much variation is a result of difference between replicate measurements; typically the former is (and should be) much greater than the latter. Finally, the reliability of repeated measurements is assessed using a measure of replicability.

**Graphs**

A graph of data that are being compared provides a visual way to evaluate measurement error. A bivariate plot of one set of measurements of a dimension against another set taken from the same specimens should have all plotted points falling on the diagonal (signifying equal values for each pair of replicate measurements) if there is minimal difference between the two sets \( (X = Y) \). Calculation of a correlation coefficient would indicate the strength of the relation between the two sets, but not the equality of the values (Altman and Bland 1983; Bland and Altman 1986). If all the points fall near the diagonal line it may be difficult to visually determine differences in values between the two sets of measurements. Altman and Bland (1983) suggest plotting the difference between a pair of replicate measurements against the mean of the pair, a technique introduced to archaeology by Gnaden and Holdaway (2000) that we hereafter refer to as difference against mean plots. A difference against mean plot reveals the magnitude of disagreement between the replicates and also highlights outliers that may represent blunders. A significant correlation between the difference between replicates and the mean of the replicates suggests systematic bias. The mean difference between replicate measurements is the relative bias and should be close to zero in unbiased data sets (Altman and Bland 1983). We present both bivariate scatterplots and difference against mean plots in all analyses to highlight their differences.

**Descriptive Statistics**

If the means of two sets of replicate measurements are significantly different as indicated by Student's \( t \), then either there are small systematic differences between the two data sets or there are large non-systematic differences (Ward and Jamison 1991). To begin to sort out which is which in a given case, we compare coefficients of variation (CV) of replicate measurement sets. In conjunction with the two plots of data, Student's \( t \) and CV statistics suggest whether or not there has been significant measurement error.

The technical error of measurement (TEM) reflects measurement precision—it is a measure of imprecision—and is expressed in the original units of measurement (Jamison and Ward 1993; Mueller and Martorell 1988; Pederson and Gore 1996; Uljaszek and Laurie 1994; Weinberg et al. 2005). It is a descriptive statistic that is calculated as the square root of the withinspecimen variance obtained from replicate measurements. If a measurement is taken twice, TEM is calculated as:

\[
TEM = \sqrt{\frac{\sum D^2}{2N}}
\]

where \( D \) is the difference between two replicate measurements and \( N \) is the number of specimens measured. The main source of imprecision or TEM is random imperfections in the measurement instrument, measurement technique, or recording technique (Mueller and Martorell 1988). There is no way to evaluate the statistical significance of TEM; values must be assessed subjectively by considering the question: How much error are you willing to accept?

The relative technical error of measurement (rTEM or \%TEM) is a kind of coefficient of variation; the magnitude of error is reported as a percentage of the average size of the dimension under study (Jamison and Ward 1993; Pederson and Gore 1996; Weinberg et al. 2005). It is calculated as:

\[
\%TEM = \frac{TEMP}{\text{grand mean}} \times 100
\]

where the grand mean is the mean of all measurements, including replicates, of a dimension. As with TEM, the \%TEM is a descriptive statistic that must
be subjectively evaluated. Like other measures of spread, it provides a comparative measure of magnitudes of dispersion of multiple measurements, whether replicate or not. Two sets of values can be similar in terms of their mean, but different in terms of their dispersion; or, they may have different means but similar dispersions. %TEM may be a small absolute value (e.g., 2–3 percent), but depending on the resolution desired (to the nearest mm, the nearest tenth of a mm, or the nearest one hundredth of a mm), a small %TEM may have a large influence on analytical results.

Finally, because we are particularly interested in reliability (repeatability), we calculated a coefficient of reliability (R) which ranges from 0 to 1, with values closer to 1 indicating greater reliability (Himes 1989; Jamison and Ward 1993; Uljaszek and Lourie 1994). Recall that precision concerns obtaining consistent results. R estimates the degree of precision or consistency when intraobserver variation is under scrutiny; it estimates measurement error when interobserver variation is examined. A simple way to calculate R is:

\[ R = \frac{\sigma_2}{\sigma_1^2 + \sigma_2^2} \]

where \( \sigma_2^2 \) is the variance of the grand mean (mean of all measurements, including replicates) and \( \sigma_1^2 \) is the variance of the mean difference between a set of replicate measurements (\( \Sigma D/N \), where D is the difference between two replicate measurements and N is the number of specimens). R reveals the proportion of between-subject variance that is free from measurement error. The rationale for R is that variability has two components—that existing between specimens that are measured (\( \sigma_1^2 \)), and that resulting from random errors in measurement (\( \sigma_2^2 \)). Reliability (R) increases as \( \sigma_2^2 \) decreases (Uljaszek and Kerr 1999; Ward and Jamison 1991). If R = .95, then 95 percent of the variance in measurements is due to factors such as between-specimen variability rather than measurement error.

Results

The bone data and lithic data differ in several important ways. First, bone is a bit softer than stone, and this might create greater differences between replicate measurements of bone than of stone. Second, some bone data were not recorded with this study in mind (Lawler and Lyman) whereas all of the lithic data were specifically recorded with this study in mind. This too could influence results when replicate data sets are compared. For each material type, we first present exemplary analyses of intraobserver variation, followed by exemplary analyses of interobserver variation. Because we were primarily interested in intraobserver and interobserver variation that would be typical of most analyses, we did screen our data for obvious blunders, just as most analysts would (e.g., Adams and Byrd 2002). We did discover a handful of blunders when comparing the handwritten data to the data after it was entered into an electronic database. Other blunders were detected only after some analyses had been performed, requiring that the analyses be repeated after correcting them. However, a few significant errors escaped our screening, and were only discovered as a result of the methods we employ here to detect them. We retain these in our analyses for illustrative purposes.

Bones

The means for lateral length of Lyman1 and Lyman2 are not significantly different (Student’s t = 1.164, \( p > .25 \)); they differ by \(< .02 \) mm (Table 1). The means of distal breadth, on the other hand, are significantly different (Student’s t = 4.12, \( p = .0005 \); they differ by an average of .17 mm. Dissimilarity is likely the result of ambiguous landmarks defining the distal breadth dimension, a problem with measurability as defined above. The lateral and medial surfaces of the distal trochlea diverge from the po antero-dorsal surfar rather than points. 1 observer error rest measurability illus a significant issue. spurious, given that reflect the same ass of such differences error could be signi

The effects of in are not consistently of the assemblages of such errors may sorts of compars length (\( t = .129, p > . \)) are for example, suggest precision from the f second for d impact of measure vary is in one : measurement error entire measureme there are no differen­ ments might mi­clude that accuracy turn would bolster tence in means of dis

The lack of differ­lengths indicates si­ilar accuracy, and the ation as reflected in \( t \) and lateral length is one set of measure­ments of the other. However distal breadth sug­gests that error created sences between the re the fact that all m people experienced analyst i

Can the problems detected using the va...
n1) whereas all of the
recorded exemplary analyses of
fluence results when
ferred by exemplary intra-observer error resulting from the problems with measurability illustrates that measurement error is a significant issue. The differences are obviously spurious, given that the distal breadth measurements reflect the same assemblage, but the analytic impact of such differences when caused by measurement error could be significant in a given analysis.

The effects of intraobserver measurement error are not consistently manifested in all comparisons of the assemblages, suggesting that the influence of such errors may differentially impact different sorts of comparisons. The two CVs for lateral length ($t = 4.12, p = .001$) and the two for distal breadth ($t = .083, p > .5$) are statistically indistinguishable, for example, suggesting minimal difference in the precision from the first measurement session to the second for each dimension. While the lack of impact of measurement error on comparisons of variation is in one sense good, and indicates that measurement error did not necessarily taint the entire measurement process, the conclusion that there are no differences in the precision of the measurements might misguide an archaeologist to conclude that accuracy is equally consistent, which in turn would bolster the conclusion that the difference in means of distal breadth is meaningful.

The lack of difference between mean lateral lengths indicates similar precision and likely similar accuracy, and the absence of differences in variation as reflected in the CVs for both distal breadth and lateral length indicates that the precision of one set of measurements is similar to the precision of the other. However, differences in the means of distal breadth suggest that intra-observer measurement error created statistically meaningful differences between the replicate measurements despite the fact that all measurements were taken by an experienced analyst in ideal laboratory conditions.

Can the problems in variable measurability be detected using the various methods we introduced? If the measurements of Lyman1 and Lyman2 were identical, then a perfectly straight line should be evident when the measurements for each specimen are plotted on a bivariate graph. A diagonal line is approximated by replicate measurements of both dimensions (Figures 1a and 2a), and the difference against mean plots also suggest the replicate measurements are similar (Figures 1b and 2b). The data in the difference against mean plots are not correlated for either dimension ($r < .1, p > .15$ in both), indicating there are no systematic biases in these data. These graphic results are consistent with our earlier analysis of lateral length and distal breadth in that the variation in replicate measurements of both variables is a result of random variation as opposed to systematic bias and the replicates are similar in regards to the accuracy and precision of their measurement. Statistical comparisons of Lyman1 and Lyman2 for both dimensions also suggest that measurement error is low and replicability (and likely precision) is high for lateral length but less so for distal breadth (Table 2). The TEM, TEM, and R values all suggest intra-observer error.
ever, all four CVs for Lawler are greater than the error in their precision. Bivariate scatterplots of corresponding CVs for Lyman1 (Table 1). If there is no consistently greater variation in Lawler relative to distal breadth of 23 astragali. Solid line ill illustrate the difference between the replicates. A lateral breadth than lateral length, implying lower meas­urement of distal breadth than lateral length. This is of course reflected in the difference in mean distal breadth between Lyman1 and Lyman2.

Lawler’s mean values for all four dimensions are not significantly different from Lyman1’s mean values (Student’s t < 1.6 and p > .1 in all). However, all four CVs for Lawler are greater than the corresponding CVs for Lyman1 (Table 1). Although none of the differences in CVs are statistically significant (t < .75 and p > .3 in all), the consistently greater variation in Lawler relative to that evident in Lyman1 suggests the data sets differ in their precision. Bivariate scatterplots of variation in replicate measurements is greater in distal breadth than lateral length, implying lower measurability of distal breadth than lateral length. This is of course reflected in the difference in mean distal breadth between Lyman1 and Lyman2.

Table 2. Summary Results of Comparing Replicate Measurements of Four Dimensions of Bighorn Sheep Astragali.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>TEM</th>
<th>%TEM</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lateral length, Lyman1 to Lyman2</td>
<td>.051</td>
<td>.130</td>
<td>997</td>
</tr>
<tr>
<td>Distal breadth, Lyman1 to Lyman2</td>
<td>.179</td>
<td>.140</td>
<td>979</td>
</tr>
<tr>
<td>Medial length, Lawler to Lyman1</td>
<td>.356</td>
<td>.998</td>
<td>872</td>
</tr>
<tr>
<td>Lateral length, Lawler to Lyman1</td>
<td>.428</td>
<td>.953</td>
<td>858</td>
</tr>
<tr>
<td>Lateral depth, Lawler to Lyman1</td>
<td>.390</td>
<td>1.820</td>
<td>779</td>
</tr>
<tr>
<td>Distal breadth, Lawler to Lyman1</td>
<td>.529</td>
<td>2.090</td>
<td>775</td>
</tr>
</tbody>
</table>

Note: n = 23 in all. TEM, technical error of measurement (in mm); %TEM, relative technical error of measurement; R, reliability. Lyman1, first set of measurements recorded by Lyman; Lyman2, second set of measurements recorded by Lyman; Lawler, measurements recorded by Lawler (1992).

Lawler against Lyman1 weakly reveal this difference and suggest that inter­observer variation could be a problem, whether the measurable lateral length dimension (e.g., Figures 3a) or the less measurable distal breadth dimension is considered (Figure 4a). The difference against mean plots clearly and strongly reveal that compiling and using data generated by these two observers may produce results that have been significantly influenced by inter­observer variation, despite the lack of statistical differences between the means for all four attributes (Figures 3b and 4b). Still, none of the pairs of difference and mean values for the four measured dimensions are correlated (r < .34, p > .1), suggesting there is no systematic difference between observers. The mean difference between replicate measurements is not large across all four dimensions (mean difference < .15 mm), suggesting random error of small magnitude.

Statistical comparison of Lawler with Lyman1 shows TEM > .35, %TEM > .8, and R < .9 for each dimension (Table 2). The R statistic indicates that 90 percent of the variation in replicate measurements is due to between specimen variability, implying that as much as 10 percent of the variation may be the result of measurement error. Inter­observer variation is considerably greater than intra-observer variation (Table 2). Lawver and Lyman both measured the same standardized dimensions (Driesch 1976). Given that our statistical analyses suggest that differences in the means are the result of random error, as opposed to systematic bias, it is unlikely that the measurement error is caused by variation in measurement technique, which would be the result of random error.

Figures 2, 3, and 4 illustrate the difference between the replicate measurements recorded by Lyman; Lyman2, second set of measurements recorded by Lyman; Lawler, measurements recorded by Lawler (1992).
Comparing Replicate Isorations of Bighorn Sheep

Means of the replicate measurements of the five dimensions represented in Lyman 3 and Lyman 4 are statistically indistinguishable (Student's *t* < 2.0 and *p* > 0.05 in all; Table 3). We note, however, that Lyman 3 and Lyman 4 data for neck widths appear to differ more than other dimensions (difference = .08 mm in Figure 5) because the neck width of some specimens was not as measurable as other dimensions.

Figure 3. Bivariate (a) and difference against mean (b) plots of Lyman 1 and Lyman 2 measurements of lateral length of 23 astragali. Solid line in (a) is diagonal (X = Y). Points should fall on the diagonal if there is no difference between the replicate measurements. Projectile points would have likely resulted in statistically significant differences in the means. The lack of statistically significant differences in the means suggests a lack of intersample variation could have produced statistically significant differences in the means.

Figure 4. Bivariate (a) and difference against mean (b) plots of Lyman 1 and Lyman 2 measurements of lateral length of 23 astragali. Solid line in (a) is diagonal (X = Y). Points should fall on the diagonal if there is no difference between the replicate measurements. Projectile points would have likely resulted in statistically significant differences in the means.
Table 3. Summary Statistics for Five Dimensions of 30 Projectile Points Measured Twice by Each of Two Investigators.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Mean</th>
<th>SD</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum length, Lyman3</td>
<td>26.09</td>
<td>8.81</td>
<td>34.04</td>
</tr>
<tr>
<td>Maximum length, Lyman4</td>
<td>26.06</td>
<td>8.81</td>
<td>34.08</td>
</tr>
<tr>
<td>Maximum length, VanPool1</td>
<td>26.10</td>
<td>8.79</td>
<td>33.68</td>
</tr>
<tr>
<td>Maximum length, VanPool2</td>
<td>26.02</td>
<td>8.83</td>
<td>33.93</td>
</tr>
<tr>
<td>Axial length, Lyman3</td>
<td>25.22</td>
<td>9.07</td>
<td>36.25</td>
</tr>
<tr>
<td>Axial length, Lyman4</td>
<td>25.25</td>
<td>9.08</td>
<td>35.96</td>
</tr>
<tr>
<td>Axial length, VanPool1</td>
<td>25.26</td>
<td>9.06</td>
<td>35.87</td>
</tr>
<tr>
<td>Axial length, VanPool2</td>
<td>25.22</td>
<td>9.06</td>
<td>35.92</td>
</tr>
<tr>
<td>Neck width, Lyman3</td>
<td>7.58</td>
<td>2.52</td>
<td>33.25</td>
</tr>
<tr>
<td>Neck width, Lyman4</td>
<td>7.50</td>
<td>2.54</td>
<td>33.66</td>
</tr>
<tr>
<td>Neck width, VanPool1</td>
<td>7.56</td>
<td>2.63</td>
<td>37.88</td>
</tr>
<tr>
<td>Neck width, VanPool2</td>
<td>7.58</td>
<td>2.57</td>
<td>33.96</td>
</tr>
<tr>
<td>Shoulder width, Lyman3</td>
<td>13.48</td>
<td>3.07</td>
<td>22.77</td>
</tr>
<tr>
<td>Shoulder width, Lyman4</td>
<td>13.50</td>
<td>3.04</td>
<td>22.52</td>
</tr>
<tr>
<td>Shoulder width, VanPool1</td>
<td>13.47</td>
<td>3.09</td>
<td>23.80</td>
</tr>
<tr>
<td>Shoulder width, VanPool2</td>
<td>13.48</td>
<td>3.08</td>
<td>22.85</td>
</tr>
<tr>
<td>Thickness, Lyman3</td>
<td>3.96</td>
<td>1.58</td>
<td>39.90</td>
</tr>
<tr>
<td>Thickness, Lyman4</td>
<td>3.96</td>
<td>1.58</td>
<td>39.90</td>
</tr>
<tr>
<td>Thickness, VanPool1</td>
<td>4.03</td>
<td>1.56</td>
<td>38.71</td>
</tr>
<tr>
<td>Thickness, VanPool2</td>
<td>3.97</td>
<td>1.58</td>
<td>39.81</td>
</tr>
</tbody>
</table>

Note: All measurements are in millimeters. Lyman3, first set of measurements recorded by Lyman; Lyman4, second set of measurements recorded by Lyman; VanPool1, first set of measurements recorded by VanPool; VanPool2, second set of measurements recorded by VanPool.

Figure 5. Bivariate (a) and difference against mean (b) plots of replicate measurements Lyman3 and Lyman4 measurements of neck width of 30 projectile points. Solid line in (a) is diagonal (X = Y); points should fall on the diagonal if there is no difference between the replicate measurements. Note scale of the vertical axis in (b) and compare with Figure 6.

Figure 6. Bivariate (a) plots of replicate measurements of maximum length, Lyman3 and Lyman4. Solid line in (a) is diagonal if there is no difference. Lymph width of Lyman3 and Lyman4 are not significantly different (p > .3). This again emphasizes the influence of the measurability of a dimension (reliably locating landmarks) on measurement error.

For all five dimensions, CVs for Lyman3 and Lyman4 are not significantly different (t < .4, p > .05), indicating similar precision across the two episodes of measuring. Points in bivariate scatterplots of replicate measurements for all five dimensions fall close to the diagonal suggesting high precision. Difference against mean plots for all dimensions suggest high precision and replicability, as exemplified by data for maximum length (Figure 6). There is no indication of systematic bias across any pair of replicate measurements (r < .2, p > .3 in all comparisons of differences with means). The mean difference for all five dimensions is small (mean ≤ .1 mm for all). Measurement error is low across all five dimensions (TEM < .19; %TEM < .2.4) and repeatability 2.5 percent of the measurement error is due to between worst measurement other statistics) is 2.305, likely because of the neck width.

Results of comparing the mean values of all dimensions and replicate sets are significant and differences > .85 for all dimensions. The minimal dispersion of difference against mean...
2.4) and repeatability is high (R > .99). Less than 2.5 percent of the variation can be attributed to measurement error and > 99 percent of the variation is due to between-specimen variability. The worst measurement error (as expected based on other statistics) is with neck width (%TEM = 2.305), likely because of relatively low measurability of the neck width of stemmed points.

Results of comparing VanPool1 with VanPool2 are similar to those for Lyman3 and Lyman4. None of the mean values of replicate measurements are significantly different for any dimension (t < .2, p > .85 for all dimensions). None of the CVs for replicate sets are significantly different (t < .6, p > .5 for all dimensions). Bivariate scatterplots exhibit minimal dispersion of points from the diagonal and difference against mean plots suggest high precision and no systematic bias, although occasional outliers indicate possible blunders (Figure 7; r < .2, p > .4 for all dimensions). TEM < 0.3, %TEM is < 4.0; and R > .97 for all five dimensions (Table 4), indicating that most of the variation is the result of differences between specimens and < 4 percent of the variation is the result of measurement error. Interestingly, both the VanPool1 and VanPool2 replicates and the Lyman3 and Lyman4 replicates display high measurement error in neck width (about 2.5 to 4.0 percent), again suggesting a systematic lower degree of measurability for this dimension than for the others.

Mean values of each dimension for Lyman3 and VanPool1 are not significantly different (t ≤ 1.64, p > .1 for all five); mean differences between measurement sets are < .07 mm. CVs for the measure-
Table 4. Summary Results of Comparing Replicate Measurements of Five Dimensions of 30 Projectile Points.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>TEM</th>
<th>%TEM</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum length, Lyman3 to Lyman4</td>
<td>0.029</td>
<td>0.112</td>
<td>0.999</td>
</tr>
<tr>
<td>Maximum length, VanPool1 to VanPool2</td>
<td>0.178</td>
<td>0.716</td>
<td>0.999</td>
</tr>
<tr>
<td>Maximum length, VanPool1 to Lyman3</td>
<td>0.184</td>
<td>0.705</td>
<td>0.999</td>
</tr>
<tr>
<td>Axial length, Lyman3 to Lyman4</td>
<td>0.139</td>
<td>0.551</td>
<td>0.999</td>
</tr>
<tr>
<td>Axial length, VanPool1 to VanPool2</td>
<td>0.170</td>
<td>0.673</td>
<td>0.999</td>
</tr>
<tr>
<td>Axial length, VanPool1 to Lyman3</td>
<td>0.168</td>
<td>0.667</td>
<td>0.999</td>
</tr>
<tr>
<td>Neck width, Lyman3 to Lyman4</td>
<td>0.174</td>
<td>2.305</td>
<td>0.991</td>
</tr>
<tr>
<td>Neck width, VanPool1 to VanPool2</td>
<td>0.297</td>
<td>3.940</td>
<td>0.973</td>
</tr>
<tr>
<td>Neck width, VanPool1 to Lyman3</td>
<td>0.389</td>
<td>5.139</td>
<td>0.954</td>
</tr>
<tr>
<td>Shoulder width, Lyman3 to Lyman4</td>
<td>0.139</td>
<td>1.031</td>
<td>0.996</td>
</tr>
<tr>
<td>Shoulder width, VanPool1 to VanPool2</td>
<td>0.089</td>
<td>0.664</td>
<td>0.998</td>
</tr>
<tr>
<td>Shoulder width, VanPool1 to Lyman3</td>
<td>0.090</td>
<td>0.671</td>
<td>0.967</td>
</tr>
<tr>
<td>Thickness, Lyman3 to Lyman4</td>
<td>0.045</td>
<td>1.134</td>
<td>0.998</td>
</tr>
<tr>
<td>Thickness, VanPool1 to VanPool2</td>
<td>0.159</td>
<td>3.983</td>
<td>0.981</td>
</tr>
<tr>
<td>Thickness, VanPool1 to Lyman3</td>
<td>0.161</td>
<td>4.040</td>
<td>0.980</td>
</tr>
</tbody>
</table>

Note: TEM, technical error of measurement (in mm); %TEM, relative technical error of measurement; R, reliability. Lyman3, first set of measurements recorded by Lyman; Lyman4, second set of measurements recorded by Lyman; VanPool1, first set of measurements recorded by VanPool; VanPool2, second set of measurements recorded by VanPool.

or less is the result of measurement error or, in this case, inter-observer difference. The same lack of statistically significant differences coupled with high replicability is reflected in comparisons between Lyman3 and VanPool1, Lyman4 and VanPool1, and Lyman4 and VanPool2, which we do not present because of their redundancy with the previously described results.

Discussion

As mentioned above, we screened the replicate data sets in various ways prior to performing the analyses. The importance of "proofing" one's data cannot be overstated, but is sometimes overlooked by archaeologists rushing against deadlines and dwindling budgets. For example, Lyman failed to note VanPool1's inversion of all measurements for projectile point specimens 2 and 3 on the handwritten data sheets, an error that VanPool discovered when comparing the typed data with the handwritten original. However, some blunders are not so easily discovered. For example, when comparing VanPool1 with VanPool2, we noted that the shoulder width for projectile point 17 in VanPool1 was considerably different than the value recorded in VanPool2. We compared both with Lyman3 and Lyman4, and noted that VanPool and Lyman4 but VanPool2, a re-measuring of that value for the measurement, was able to correct it. The remaining undetected situation of comparing replicate measurements that more than two and against mean plots replicability and blunders such as the evident statistical significance coupled with the high replicability is reflected in comparisons between Lyman3 and VanPool1, Lyman4 and VanPool1, and Lyman4 and VanPool2, which we do not present because of their redundancy with the previously described results.

Discussion

As mentioned above, we screened the replicate data sets in various ways prior to performing the analyses. The importance of "proofing" one's data cannot be overstated, but is sometimes overlooked by archaeologists rushing against deadlines and dwindling budgets. For example, Lyman failed to note VanPool1's inversion of all measurements for projectile point specimens 2 and 3 on the handwritten data sheets, an error that VanPool discovered when comparing the typed data with the handwritten original. However, some blunders are not so easily discovered. For example, when comparing VanPool1 with VanPool2, we noted that the shoulder width for projectile point 17 in VanPool1 was considerably different than the value recorded in VanPool2. We compared both with Lyman3 and Lyman4, and noted that VanPool and Lyman4 but VanPool2, a re-measuring of that value for the measurement, was able to correct it. The remaining undetected situation of comparing replicate measurements that more than two and against mean plots replicability and blunders such as the evident statistical significance coupled with the high replicability is reflected in comparisons between Lyman3 and VanPool1, Lyman4 and VanPool1, and Lyman4 and VanPool2, which we do not present because of their redundancy with the previously described results.

Discussion

As mentioned above, we screened the replicate data sets in various ways prior to performing the analyses. The importance of "proofing" one's data cannot be overstated, but is sometimes overlooked by archaeologists rushing against deadlines and dwindling budgets. For example, Lyman failed to note VanPool1's inversion of all measurements for projectile point specimens 2 and 3 on the handwritten data sheets, an error that VanPool discovered when comparing the typed data with the handwritten original. However, some blunders are not so easily discovered. For example, when comparing VanPool1 with VanPool2, we noted that the shoulder width for projectile point 17 in VanPool1 was considerably different than the value recorded in VanPool2. We compared both with Lyman3 and Lyman4, and noted that VanPool and Lyman4 but VanPool2, a re-measuring of that value for the measurement, was able to correct it. The remaining undetected situation of comparing replicate measurements that more than two and against mean plots replicability and blunders such as the evident statistical significance coupled with the high replicability is reflected in comparisons between Lyman3 and VanPool1, Lyman4 and VanPool1, and Lyman4 and VanPool2, which we do not present because of their redundancy with the previously described results.

Discussion

As mentioned above, we screened the replicate data sets in various ways prior to performing the analyses. The importance of "proofing" one's data cannot be overstated, but is sometimes overlooked by archaeologists rushing against deadlines and dwindling budgets. For example, Lyman failed to note VanPool1's inversion of all measurements for projectile point specimens 2 and 3 on the handwritten data sheets, an error that VanPool discovered when comparing the typed data with the handwritten original. However, some blunders are not so easily discovered. For example, when comparing VanPool1 with VanPool2, we noted that the shoulder width for projectile point 17 in VanPool1 was considerably different than the value recorded in VanPool2. We compared both with Lyman3 and Lyman4, and noted that VanPool and Lyman4 but VanPool2, a re-measuring of that value for the measurement, was able to correct it. The remaining undetected situation of comparing replicate measurements that more than two and against mean plots replicability and blunders such as the evident statistical significance coupled with the high replicability is reflected in comparisons between Lyman3 and VanPool1, Lyman4 and VanPool1, and Lyman4 and VanPool2, which we do not present because of their redundancy with the previously described results.

Discussion

As mentioned above, we screened the replicate data sets in various ways prior to performing the analyses. The importance of "proofing" one's data cannot be overstated, but is sometimes overlooked by archaeologists rushing against deadlines and dwindling budgets. For example, Lyman failed to note VanPool1's inversion of all measurements for projectile point specimens 2 and 3 on the handwritten data sheets, an error that VanPool discovered when comparing the typed data with the handwritten original. However, some blunders are not so easily discovered. For example, when comparing VanPool1 with VanPool2, we noted that the shoulder width for projectile point 17 in VanPool1 was considerably different than the value recorded in VanPool2. We compared both with Lyman3 and Lyman4, and noted that VanPool and Lyman4 but VanPool2, a re-measuring of that value for the measurement, was able to correct it. The remaining undetected situation of comparing replicate measurements that more than two and against mean plots replicability and blunders such as the evident statistical significance coupled with the high replicability is reflected in comparisons between Lyman3 and VanPool1, Lyman4 and VanPool1, and Lyman4 and VanPool2, which we do not present because of their redundancy with the previously described results.
Comparing Replicate Os of 30 Projectile Points.

Table 5. Matrix of the Number of Variates Larger and Smaller than the Grand Median for Distal Breadth in Lyman I and Lawler. Lawler has only 22 instead of 23 variates because the median value was excluded from the matrix.

<table>
<thead>
<tr>
<th></th>
<th>Lyman I</th>
<th>Lawler</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number less than grand median</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Number greater than grand median</td>
<td>12</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 6. Matrix Reflecting Systematic Bias of +20mm in Lyman I (+2mm) Lawler

<table>
<thead>
<tr>
<th></th>
<th>Lyman I (+2mm) Lawler</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number less than grand median</td>
<td>7</td>
</tr>
<tr>
<td>Number greater than grand median</td>
<td>16</td>
</tr>
</tbody>
</table>

noted that VanPool 2 was rather close to Lyman 3 and Lyman 4 but VanPool 1 was not. This prompted a re-measuring of specimen 17, which determined that the value for VanPool 1 was a blunder during the measurement process. We were consequently able to correct this blunder, but it may have remained undetected in the typical archaeological situation of comparing different assemblages without replicate measurement.

When analyzing replicate data sets, substantial, or statistically (and hence analytically) significant blunders such as those just described are likely to be easily identifiable. Blunders of large magnitude will be manifest as a difference between replicate measurements that exceeds the mean difference by more than two standard deviations. Difference against mean plots reveal more about precision, replicability and blunders than do bivariate plots, and more consistently. Consider, for example, Figure 7. No obvious outliers are apparent in the bivariate scatterplot (Figure 7a), but the arrow identifies the specimen that is the extreme outlier in the difference against mean plot (Figure 7b). Examination of the data suggests that the value recorded for this specimen in VanPool 1 was a blunder. Bivariate scatterplots seem to reveal blunders only if their scales are of sufficiently high resolution, as in Figure 8a; this blunder is apparent in both graph styles (Figure 8). Difference against mean plots may also suggest a low degree of measurability, such as with the imprecise measurability of neck widths evident on stemmed projectile points (Figure 5b).

There were significant differences in variation between Lyman’s and Lawler’s data sets, and in one case, these differences produced differences in the mean. The variation is likely random, and the result of differences in precision and, in the case of distal breadth, poor measurability. Given that random variation is less likely to create an analytic problem than bias, it may be important to differentiate between cases where measurement error is random, and when it reflects bias. There is a statistically means of doing so using the median and a chi-square test. The median perfectly splits a distribution such that half of the values are larger and half are smaller. In a distribution with an even number of variates, the median is the average of the two central variates. To evaluate whether there is bias in two or more assemblages, determine the grand median for all of the assemblages put together. If there is no bias, roughly half of the variates of each sample will be larger than the grand median and half will be smaller. In a biased data set, this will not be the case, given that one or more of the data sets will systematically be larger than the other data set(s). This will result in the data sets differing significantly in the number of variates on either side of the median.

In the case of the bone data, the median distal breadth for Lyman I and Lawler is 25.1 mm. Determining the frequency of values above and below the median produces the matrix in Table 5. The frequencies for each data set greater than and less than the median are equal, exactly what is expected if the variation is random, as opposed to biased. Imagine that Lyman used a caliper that produced precise but inaccurate results that were consistently 2 mm larger than they should be. Modifying the data in this manner produces the matrix in Table 6. The preponderance of variates below the median in one sample and those above the median in the other suggests biased measurements. A chi-square test indicates the differences are significant ($\chi^2 = 7.04, p < .01$), suggesting the (contrived) data are systematically biased.

Obviously, bias, blunders, and random errors are most easily detected (and perhaps only detectable) when the same objects are re-measured and the results compared as we have done here. Any differences between the data samples must be attributed to measurement error. Determining the presence of error in cases where different specimens are being compared will be considerably trickier, given that differences in medians, means,
CVs, and any other statistical measure could reflect real differences in the data sets, not differences in measurement protocol, the use of unstandardized dimensions, problems with measurability, or the presence of mechanical errors.

Unfortunately, such "tricky" cases include nearly all comparisons archaeologists wish to make. Given that archaeology is a comparative science, we are in an awkward position. Archaeologists intuitively know, and we clearly demonstrated above, that measurement error is a significant confounding issue for the comparisons we make every day. Yet it is impossible to differentiate from analytically meaningful differences in most contexts. How can one determine the significance, if any, of measurement error?

Our suggestion is that, when possible, archaeologists should re-measure a sample of specimens from each collection to test for measurement error, bias, and blunders. In our experience, such remeasurement is not typical in archaeology, although other sciences, especially those focused on experimental replication do it more frequently. In fact, we personally know of no examples (including our own work) where it has been used by archaeologists, but even the reanalysis of a small sample from each of the assemblages one wishes to compare will be adequate to detect the presence of bias and problems with measurability, precision, and accuracy (Gnaden and Holdaway 2000). Such measurement error is best, and often only, detected when comparing measurements of the same specimens. We are of course proposing additional work for archaeologists who labor under time and budgetary constraints. After all, who really wants to reanalyze artifacts that have already been measured once? However, we ask which is worse—spending a small amount of time reanalyzing a sample of artifacts (a sample of 20 from each collection will likely be enough to detect significant measurement error), or spending weeks, months, or years creating reports, professional papers, and journal articles attempting to explain spurious differences (or similarities) in behavioral terms that reflect nothing more than measurement errors? Archaeological analyses will be improved, and our time and budgets better utilized, if archaeologists pay serious attention to and control for measurement error.

Conclusions

In a non-random sample of 10 publications on projectile point morphometrics by ten authors published between 1972 and 1999, 5 of the authors illustrate the dimensions they measured, 4 authors simply name the dimensions (e.g., length, width), and one indicates that he measured dimensions defined and illustrated by Thomas (1981). Only 3 of 10 authors specify the tool they used to take measurements. All 10 authors indicate the resolution of the measurements they took, but 4 of those authors do not mention resolution in the text and instead resolution information must be gleaned from the data tables. Resolution ranges from .1 cm to .001 mm; the resolution mode is .1 mm (n = 7). None of the ten authors says anything about measurement error, replicability, precision, blunders, measurability, bias, or observer variation. Given the implications of measurement error we found under ideal circumstances, comparisons based on measurement data irreconcilably will produce spurious trends, and we These measurements any single individuals, but are instances such as archi Our general and equally measured in a uniform accurate measure testing and increasing frame creating. However problems those! problems are every measure everyone knows when using the sa or five sentences ment technique, tion would be we. However, ourtions that some de'surements may be The magnitude o some, depending. For example, the fidelity of cult'several reasons (e. cate and mimic p come within 3 per have made a perfect 2001; Eerkens and within assembl no other reason that more variation is i transmission. This if documented var inter-observer (or consider the following)
The projectile p the Great Basin is to 600 $^{14}$C yr B.P.) arrow point (Bettin and Bettinger 2001) efforts to rigorously point styles based ologists working in
measurement data run significant dangers of being irreconcilably tainted by measurement error that will produce spurious patterns, obscure meaningful trends, and waste archaeological time and effort. These measurement problems are not the fault of any single individual (or even all of the individuals), but are instead inherent in a comparative science such as archaeology.

Our general lack of consistently standardized and equally measurable attributes that are measured in a uniform way that guarantees precise and accurate measurements reflects the incredibly interesting and increasingly robust analytic and methodological frameworks that archaeologists are creating. However, we must be cognizant of the problems those frameworks create. Some of the problems are easily dealt with by simply clarifying the measurement procedure to ensure that everyone knows that they mean the same thing when using the same term. We suggest that the four or five sentences necessary to describe measurement technique, tool used, and dimension definition would be well worth the page space.

However, our results parallel earlier observations that some degree of difference between measurements may be a function of different observers. The magnitude of this difference may be worrisome, depending on the analyses one is performing. For example, recent research has revealed that the fidelity of cultural transmission is imperfect for several reasons (e.g., a novice who lacks replication and mimicry) and that it may come within 3 percent of doing so and think they have made a perfect replica; Eerkens and Bettinger 2001; Eerkens and Lipo 2005). Variation between and within assemblages is therefore expected for no other reason than copying error, but significantly more variation is interpreted as a lack of cultural transmission. This conclusion may be unwarranted if documented variation is in truth the product of inter-observer (or even intra-observer) error. Consider the following instance.

The projectile point type known as Rosegate in the Great Basin is a relatively late type (ca. 2000 to 600 $^{14}$C yr B.P.) that likely represents a style of arrow point (Bettinger and Eerkens 1999; Eerkens and Bettinger 2001). Ever since Thomas’s (1981) efforts to rigorously define Great Basin projectile point styles based on point morphometry, archaeologists working in the area have measured thousands of projectile points and often they have published the metric data. Thomas and Bierworth (1983), for example, published the metric data for eight dimensions of nearly four dozen Rosegate points recovered from Gatecliff Shelter, Nevada. Similarly, Drews (1990) published the metric data for the same eight dimensions of 33 Rosegate points recovered from James Creek Shelter, Nevada. Summary data for each set of points and each dimension are presented in Table 7. Mean values of three of the eight dimensions are significantly different (Student’s $t \geq 2.0$ and $p < .05$ for the three).

Perhaps the differences in means at Gatecliff Shelter and James Creek Shelter are the result of local variation in the production of Rosegate points, or perhaps the differences are the result of copying error, a particular kind of horizontal transmission between the two sites or the geographic separation of the two (which would influence transmission). The two sites are about 250 km apart, with James Creek Shelter in northeastern Nevada and Gatecliff Shelter in central Nevada. Whatever the case, our analyses indicate that measurement error, whether of the intra-observer or inter-observer sort, should be assessed before cultural processes are inferred on the basis of the measurement data in Table 7.

Our analysis has focused on the similarity of replicate measurements of the same specimens, whether taken by one observer or two. Most archaeological analyses, however, involve comparing sets of measurements or collapsing sets into one large set, all taken from different specimens or assemblages, typically by more than one observer. Our results suggest the usual archaeological process is reasonable insofar as dimensions are standardized and (importantly) measurable, and similar measurement techniques and tools are used. Still, we suggest that a sample of the specimens being compared or collapsed should be measured by the same person at least twice. Analyses like those described here can then be used to determine if there are differences in precision and/or accuracy created by low measurability that might complicate any planned analyses. High levels of replicability mean high measurability, precision, and (probably) accuracy, providing an empirical warrant for comparisons and collapsing. Low levels of replicability would suggest caution during comparisons, deletion of unmeasurable attributes, or a reason to not collapse data sets. Outliers probably represent blun-
Acknowledgments. We thank M. Lawler for a copy of his thesis, Denny Walker for the loan of his collection of bighorn sheep astraguli, and two anonymous reviewers for encouragement.

References Cited


Klein, Richard G. 1991 Size Variation (situs) and Late Quarter Between Cape Pro Research 36:243-255.


Lyman, R. Lee 2006 Identifying Bil Bones: How Symm Final of Archaeological 2009 The Holocene canadiensis) in East USA. The Holocene


Utermohle, Charles J., and Stephen L. Zegura

VanPool, Todd L., and Robert D. Leonard

Ward, R. E., and P. L. Jankowski

Weinberg, Seth M., Nicole M. Scott, Katherine Neiswanger, and Mary L. Marazita

Whittaker, John C., Douglas Caulkins, and Kathryn A. Kemp


Zeder, Melinda A., and Brian Hesse

Submitted January 12, 2009; Accepted March 24, 2009.