TECHNIQUES FOR ASSESSING STANDARDIZATION IN ARTIFACT ASSEMBLAGES: CAN WE SCALE MATERIAL VARIABILITY?

Jelmer W. Eerkens and Robert L. Bettinger

The study of artifact standardization is an important line of archaeological inquiry that continues to be plagued by the lack of an independent scale that would indicate what a highly variable or highly standardized assemblage should look like. Related to this problem is the absence of a robust statistical technique for comparing variation between different kinds of assemblages. This paper addresses these issues. The Weber fraction for line-length estimation describes the minimum difference that humans can perceive through unaided visual inspection. This value is used to derive a constant for the coefficient of variation (CV = 1.7 percent) that represents the highest degree of standardization attainable through manual human production of artifacts. Random data are used to define a second constant for the coefficient of variation that represents variation expected when production is random (CV = 57.7 percent). These two constants can be used to assess the degree of standardization in artifact assemblages regardless of kind. Our analysis further demonstrates that CV is an excellent measure of standardization and provides a robust statistical technique for comparing standardization in samples of artifacts.

El estudio de estandarización y variación ha sido una importante y valiosa linea de interés en los análisis arqueológicos. Sin embargo, aún persisten dos problemas que son el enfoque de este estudio. En primer lugar, faltan medidas independientes para evaluar problemas de estandarización y variación. En otros términos, no hay nada que indica cómo se debe hacer una muestra arqueológica bien estandardizada o bien variable. En segundo lugar, no existe una técnica estadística segura para hacer comparaciones cuantitativas. El 'Weber fraction,' utilizado para la estimación de una línea amplia describe la diferencia mínima que seres humanos pueden percibir con sólo una inspección ocular. Este valor es utilizado para derivar una constante (CV = 1.7 percent) que representa la variación mínima obtenida a través de la producción manual de artefactos por seres humanos. Datos aleatorios son utilizados para determinar una segunda constante que representa la variación esperada bajo condiciones aleatorias (CV = 57.7 percent). De este modo, estas dos constantes pueden estar utilizadas para determinar el grado de estandarización en las colecciones de artefactos. También, este estudio proporciona una técnica estadística segura para comparar la estandarización en muestras de artefactos.

The study and interpretation of artifact variation is essential for understanding and explaining the archaeological record. The most visible contribution of this research is taxonomic, the creation of schemes that divide material culture into meaningful functional, temporal, and geographical categories. In recent years, however, inquiry has increasingly shifted from developing taxonomies to interpreting the variation that makes them work. The study of variation and standardization has become commonplace across a broad range of subject matters relevant to anthropological theory and culture history (see Rice 1991 for a review). Ceramics feature prominently in these studies (e.g., Arnold 1991; Blackman et al. 1993; Costin and Hagstrum 1995; Crown 1999; Longacre 1999; Longacre et al. 1988; Rice 1991; Rottlander 1966), but lithics (Bettinger and Eerkens 1997, 1999; Chase 1991; Eerkens 1997, 1998; Hayden and Gargett 1988; Torrence 1986), bone and antler (Dobres 1995), and textiles (Rowe 1978) are also well represented.

Variation is useful for understanding such a broad range of phenomena because it reflects the degree of tolerance for deviation from a standard size, shape, form, or method of construction. Higher tolerance increases variability, while lower tolerance decreases variability leading to standardization. Standardization, then, is a relative measure of the degree to which artifacts are made to be the same. Standardization is in turn related to the life cycle of the artifact type or

Jelmer W. Eerkens ■ Department of Anthropology, University of California, Santa Barbara, CA 93106
Robert L. Bettinger ■ Department of Anthropology, University of California, Davis, CA 95616

American Antiquity, 66(3), 2001, pp. 493–504

class in question, reflecting such things as production costs, consumer preferences, replication and learning behaviors, number of producers, concern with quality, producer skill, and access to resources. Unfortunately, the statistics of variation have not kept pace with this growing interest in variation. Although many approaches have been used, none is universally applicable, and, when the analysis proceeds to interpretation, the emphasis is always on qualitative rather than quantitative characterizations. Studies of variation have employed a sophisticated range of measures (e.g., standard deviation, coefficient of variation, skewness, etc.), but nothing in the theoretical or experimental literature provides an independent standard for interpreting these measures. Nor is it possible, given the present situation, to compare the amount of variation observed between two artifact classes, for example, between ground stone and chipped stone artifacts. In sum, the anthropological study of variation lacks a robust statistical approach.

This paper addresses these issues on two counts. First, it seeks to place observed artifact variation within a universal context by exploring theoretically derived guidelines or baseline values that can assist interpretation. The upper baseline (highest degree of standardization) describes the minimum amount of metrical variation humans can generate without such external aids as rulers. The lower baseline describes the amount of variation that will occur when there is no attempt at standardization at all, i.e., when production is random relative to a standard. We borrow from psychology and statistics to derive these boundaries. Second, we present a statistical method for comparing variation between assemblages that is applicable to cases where assemblages differ with respect to artifact class or attribute size. We argue that under most circumstances coefficient of variation *(CV)* is a stable and reliable measure of variation.

Human Error and Weber Fractions

Humans commit all kinds of errors when hand-crafting such objects as ceramic pots and stone projectile points. The kind of error we are interested in here is that which would result were one to show a skilled stone knapper a model projectile point, request 10 identical copies (to the best of his/her ability), and allow the knapper to discard any specimens s/he might regard as deviant. Observed variation in shape or size of the 10 points would then represent knapper error in replicating the model projectile point. Multiple factors would contribute to this error (Rice 1991:273 lists several), but a key source is what can be termed scalar error, stemming from errors in estimating object size and translating mental images into properly scaled physical objects. This error is neither random nor absolute. It is limited by human visual perception and motor skill and increases linearly with the magnitude or size of the intended end product (e.g., Coren et al. 1994). This makes it possible to define a quantitative boundary for the least amount of variation that can be expected under the most rigorous kind of production.

When humans attempt to estimate the size or magnitude of an object visually, without reference to an independent scale (i.e., without a ruler), they make mistakes that grow larger in absolute size as the size of object increases; the larger the object, the larger the absolute error in estimated size (Coren et al. 1994:39–43; Kerst and Howard 1978; Teghtsoonian 1971). Similarly, when people attempt to make an object from a mental image or model, they make mistakes that increase in absolute size as template size increases. If a person makes 10 objects independently from the same mental image, both the range and standard deviation of size of those 10 finished objects will increase as the template size increases. In short, error and size are correlated; people make larger absolute errors when making larger objects. More importantly, the rate at which error and intended size are correlated is linear. Such scaling error is frequently discussed in the psychophysics literature (e.g., Algom 1992; Coren et al. 1994; Gescheider 1997; Miller 1956; Stevens 1975) and has also been observed in archaeological materials (e.g., Bettinger and Eerkens 1997; Eerkens 1998; Shott 1997) and replication experiments (Eerkins 2000).

This phenomenon is a product of how the human brain interprets, measures, and compares visual and other sensory information. In the mid-1800s E. H. Weber observed that the ability of individuals to discriminate between objects of different weight depended on the mean weight of the objects involved (Coren et al. 1994:39–43; Weber 1834). In lifting experiments Weber discovered that to be perceived as differing in weight, heavy objects had to differ by a greater absolute amount than lighter objects. Weber also determined, however, that the relative difference needed to make such distinctions remained relatively constant. Specifically, he found that two

objects had to differ by more than about 2 percent (1/50) for a difference in weight to be detected, meaning that two large objects had to differ more in absolute size than two small objects. Thus, unlike mechanical scales that determine weight within an invariant unit of error (e.g., ± .1g), human appreciation of heaviness is scaled relative to object weight (see Jones 1986; Ross 1981, 1995; Stevens 1979 for more recent work with weight). This value (2 percent) has come to be called the Weber's fraction for heaviness (see also Norwich 1987; Ross 1997; Ross and Gregory 1964; Teghtsoonian 1971).

Human perception of length and area are similarly scaled. The Weber fraction for perception of the length of a line is similar to that for heaviness, about 3 percent (Teghtsoonian 1971). This number varies slightly from person to person, but does not vary significantly by gender, age, or within an individual over the course of time (Verrillo 1981, 1982, 1983) although remembered length seems to vary more with increasing time (Kerst and Howard 1978, 1981, 1984) and context (Hotopf et al. 1983; Pagano and Donahue 1999). In this respect the Weber fraction for length perception is surprisingly constant over an extremely wide range of sizes (Coren et al. 1994; Laming 1997; Poulton 1989; see also Ross 1997). Recent work with other aspects of vision, such as color and contrast recognition, stereopsis, blur discrimination, and depth perception, show similar magnitude and error-scaling properties, though the Weber Fraction value and the structure of the relationship can change (Howard and Rogers 1995; Mather 1997; Schwartz 1999; Smallman et al. 1996).

Thus, the ability of humans to perceive a difference in the size of two objects, or between a mental image of an object and the object itself, is limited by our sensory system. This difference must be at least 3 percent. This does not apply when a physical standard, such as a ruler, is used as the method of measurement. In that context, the ability of a subject to measure size or length is independent of absolute object size, turning instead on subject ability to differentiate between marks on the ruler. With a ruler, the error in measuring 10-cm objects and 1000-cm objects is the same.

Scaling and Artifact Variation

That scalar error and object size are linearly and positively correlated in human perception of weight, length, and area has several implications for understanding artifact variability. Foremost, it implies that scalar error divided by size will be constant in sets of handmade artifacts that are manufactured without rulers. This is convenient because archaeologists frequently express artifact variation in precisely this manner using the Coefficient of Variation *(CV)*, defined as the sample standard deviation divided by the sample mean, which is often multiplied by 100 and expressed as a percentage. Thus, the Weber fraction and *CV* both express variation scaled to magnitude. Further, it is easy to convert the Weber fraction into *CV* form by using the notion of a uniform distribution.

A uniform distribution defines a range within which all values are equally frequent or probable. This might be the case if one were randomly picking numbers between 0 and X out of a hat, each number being represented once and having the same chance of being drawn. Such a population is uniformly distributed between 0 and X, with a mean of $X/2$. The width of the range, then, is twice, or 200 percent of, the mean, running from $X/2 - X/2$ (= 0) to $X/2 + X/2$ (= X). Regardless of the size of X, all such distributions have a *CV* of 57.7 percent (= $1/\sqrt{3}$).[1] In comparison, Weber error should generate distributions that are uniform but much more narrowly limited around the mean. As we have seen, acting alone, Weber error will cause humans to produce collections of objects whose range in size is 6 percent of the mean, i.e., from 97 percent of the mean to 103 percent of the mean. Such a relationship might be expected if subjects were asked to draw a line equal in length to a reference line and they did so without any additional error due to motor-skill inaccuracy. When the line is drawn within +3 percent or –3 percent of the reference, subjects perceive the two as equal and stop drawing, though in reality the lines would differ by some finite amount. Since human perception is unable to discern smaller differences, values will be uniformly distributed within these extremes (i.e., all values are equally likely).[2] Since the *CV* for uniform distributions whose range is 200 percent of the mean is $1/\sqrt{3}$ (= 57.7 percent), it follows that the *CV* for the narrower uniform distribution defined by the Weber fraction (range = 6 percent of the mean) will be (6 percent / 200 percent) x $1/\sqrt{3} \cong$ 1.7 percent. This is essentially identical to the values obtained in psychological experiments where subject estimates of line-segment length display a *CV* of 1.6 percent (Ogle 1950:231). We can produce the same result empiri-

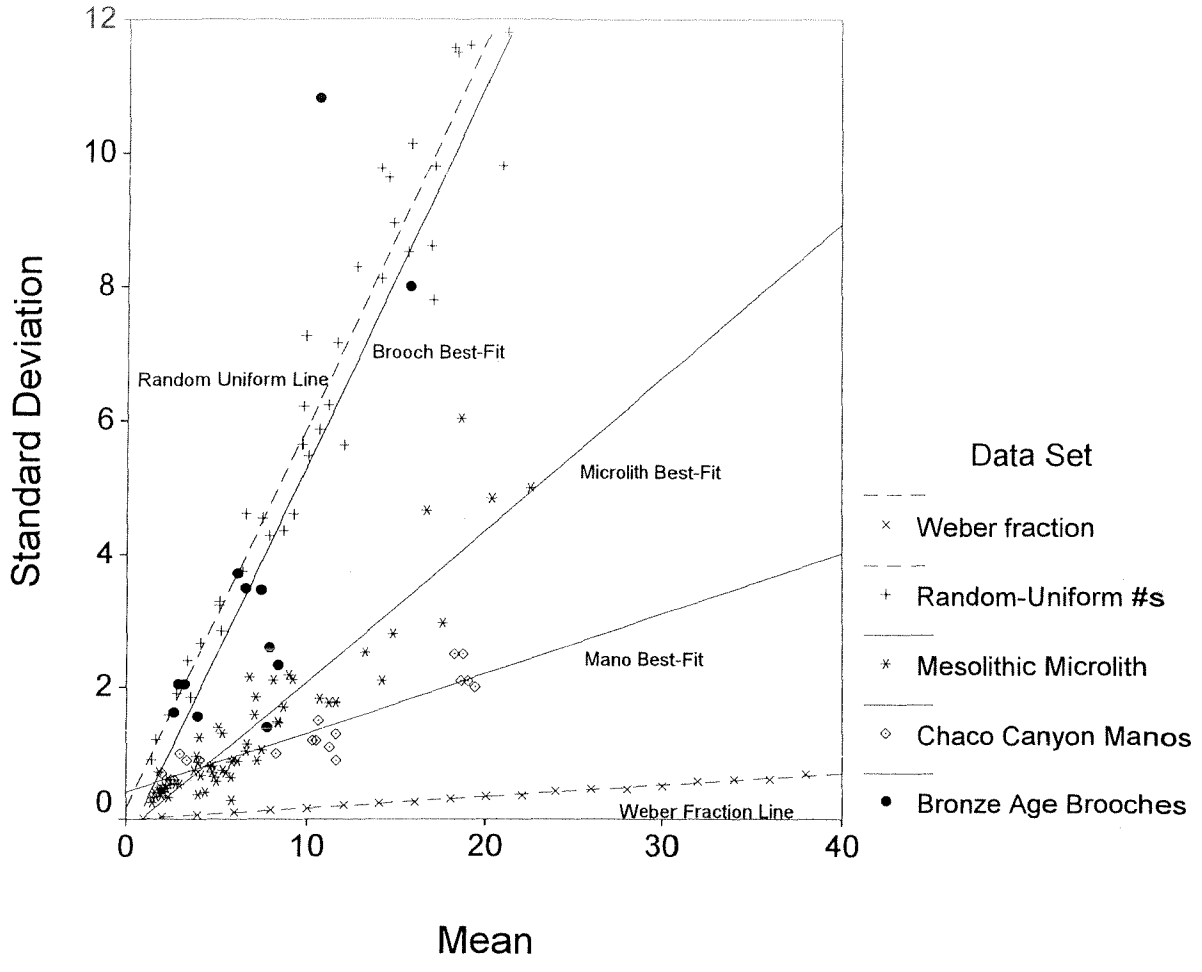

Figure 1. Mean-standard deviation relationships for three archaeological data sets and relationship to random data and Weber fraction. Solid lines represent best-fit regression lines through relevant data points. Dashed lines represent best-fit regression lines through data generated by Weber fraction and random-uniform data.

cally by repeatedly drawing random numbers from uniform distributions whose means are different but whose ranges are always from 97 percent of the mean to 103 percent of the mean. The lower line in Figure 1 presents such a simulation. Each of the 20 cases shown represents a sample of 200 numbers randomly drawn from a uniform distribution, producing a corresponding mean, standard deviation, and *CV*. As shown, the expected *CV* for each case is 1.7 percent. The simulation obeys this expectation with *CV*s falling very near 1.7 percent for each of the 20 cases.

The *CV* of 1.7 percent derived for the Weber fraction should represent the *minimum* amount of variability attainable by humans for length measurements. Variation below this threshold is not possible given the visual perception capabilities of most humans. Sets of artifacts that display *CV*s less than 1.7 percent imply automation or use of an independent standard. Of course, small errors in motor skills and memory will introduce additional variability in the manual production of artifacts (Algom 1992; Kerst and Howard 1984; Moyer et al. 1978). Eerkens (2000) suggests that *CV*s in the range of 2.5–4.5 percent are more typical of the minimum error attainable by individuals in manual production without use of external rulers. Similarly, Longacre (1999) reports *CV* values in the range of 2–5 percent for aperture, circumference, and height for "standardized" handmade pots, constructed without the use of a ruler by highly skilled Philippine specialists. Quite clearly, these artifacts are highly standardized, approaching the Weber fraction *CV*, and are probably close to the minimum *CV* attainable in manual production.

It follows from all of the above that variation in artifacts produced manually without the use of an

independent ruler should be scaled positively and linearly to the mean. Attributes that fail to show such scaling imply an alternative mode of production.

Variation in random distributions is also scaled to the mean. As we have just shown, any uniform distribution of positive numbers, with a lower limit of 0, an upper limit of X, and a mean of $X/2$, will have a CV of 57.7 percent. A set of artifact attributes distributed in such a manner would imply that variation within 100 percent on either side of the mean was within production standards; this as opposed to 3 percent on either side of the mean defined by the Weber fraction. Production where anything from 0 percent of the mean to 200 percent of the mean is tolerated would, indeed, be extreme, and clearly, humans do not produce artifacts in uniformly distributed ways (again, see note 2). However, even a normally distributed variable with a CV of 57.7 percent displays nearly as many variates that fall more than half of the mean from the mean as a uniform distribution (39 percent against 50 percent); and the normally distributed variable displays *more* variates that fall further than the mean from the mean than the uniform distribution (8 percent against 0 percent). In short, whether populations are normally or uniformly distributed, CVs greater than or equal to 57.7 percent are derived from *extremely* variable populations in which approximately 40–50 percent of the variates fall more than half the mean from the mean.

As just noted, artisans producing material goods are unlikely to be working with an arbitrary size interval ranging from an unspecified value of X all the way down to 0. In the real world, therefore, unstandardized assemblages should display CVs less than 57.7 percent. Observed minimum and maximum values can be used to obtain a more conservative, empirical standard for random production in specific cases = $((A\text{-}B)/(A+B))$ x .577, where A is the maximum observed value and B is the minimum observed value. This avoids the implication that objects of zero length or zero size are acceptable when production is random (which is obviously not so), but risks the possibility that the observed maximum and minimum values underestimate the true limits of production tolerance. Because of the latter, we prefer to use the theoretically derived value (CV = 57.7 percent) as the baseline standard for random production, noting that under the proper conditions important insights may be gained through the use of an empirically determined standard.

Independent Standards and Use of the CV

The CVs derived from the Weber fraction and the uniform-random distribution provide two baseline measures against which variability in artifact assemblages can be compared. The uniform-random CV value (57.7 percent) does not involve the kind of psychological limitation that gives rise to the Weber fraction CV (1.7 percent). Nevertheless, it provides a useful measurement to examine variability encountered in archaeological situations. Variation below 1.7 percent suggests use of a scale or external template to measure and manufacture artifacts and should be typical of settings where items are mechanically produced (i.e., perhaps from a mold or by a machine). Variation above 57.7 percent suggests intentional inflation of variation and may indicate situations where individual manufacturers are actively trying to differentiate their products from those of others, thereby increasing variation. An intentional increase is necessarily implied because variation is greater than would occur when production is completely random. Alternatively, such cases might also describe situations where an archaeologist has unknowingly lumped two or more discrete classes of artifacts into a single category, thereby artificially increasing variation.

Between these extremes a wide range of possibilities exist. Figure 1 also shows mean-standard deviation relationships for several archaeological collections including microliths from Mesolithic sites in Northern England, manos or milling handstones from Chaco Canyon, New Mexico, and Bronze Age safety pin brooches from Switzerland.

As Figure 1 shows, archaeological data often show linear correlation between mean and standard deviation (see Bettinger and Eerkens 1997 for a similar discussion for Great Basin projectile points). Lines running through the data indicate best-fit regressions. However, the nature of the regression, as measured by the slope, varies by collection. Steeper slopes denote collections characterized by less-standardized attributes (i.e., standard deviation increases relatively sharply relative to size). For example, Chaco Canyon manos show the least variation with increasing mean (see Cameron 1997), though they show more variation than the Weber fraction for length (indicated by a dashed line near the bottom of Figure 1). This suggests that of the three collections, the Chaco Canyon manos are the most

standardized or most consistently made. Microliths (see Eerkens 1997, 1998) show greater variation on average than Chaco Canyon manos, but less than Bronze Age brooches (see Doran and Hodson 1975), which equal the variation expected under random conditions. We are reluctant to characterize the production of brooches as random, since each was carefully made in a certain way. However, the high *CV*s suggest manufacturers were relatively unconcerned with conformance to a specific size. In this respect, the brooches represent a very unstandardized set of artifacts—at least with respect to size (see also Torrence 1986:158–159 for examples of highly variable lithic data sets where *CV*s exceed 57.7 percent).

Another way to think of this is in terms of the intensity of constraints or forces acting to reduce variation within a data set (perhaps how intensely the data set has been winnowed or selected). The strength of the regression as measured, for example, by r^2, describes the consistency in standardization within the data set from sample to sample. Collections with high goodness-of-fit values suggest that the intensity of selection is roughly equal on all samples, while lower values imply that some attributes or samples are more standardized than others.

Importantly, the figure demonstrates that standard deviation is inappropriate as a statistic to compare standardization between samples, because it fails to scale variation properly. Samples with smaller means will have smaller standard deviations simply because their means are small, hence will appear more standardized. For example, consider two samples of random numbers drawn from uniform distributions, the first with a mean of 10.03 and standard deviation 5.47 ($n = 50$) and a second with mean .96 and standard deviation .56 ($n = 50$). These two samples represent two distinct points in the left-hand line in Figure 1. Since both samples contain completely random numbers, neither is more standardized than the other. However, any statistical test used to compare standard deviation between these two samples, including the *F*-Test and Brown-Forsyth test (see below), would find statistically significant differences between the two. Such a test would wrongly conclude that the second sample is more standardized than the first. A test comparing *CV*s, on the other hand, would find no difference, which is the desired result (see below). *CV* is an inappropriate comparative measure, however, when the relationship between mean and standard deviation is either negative or non-linear.

In sum, where the mean-standard deviation relationship is linear and positive, especially when the regression line passes near the origin, *CV* will be the more reliable measure of variation because it scales standard deviation to the mean. When these conditions hold, *CV* facilitates comparison of variation across different-sized attributes (i.e., large vs. small), as well as across attributes measured by different scales (i.e., centimeters vs. grams). Most metric attributes typically measured in artifact assemblages (e.g., length, width, thickness, diameter, and weight) meet these conditions. Provided the range of values is not excessive (i.e., not greater than 180 degrees), angular data should also meet these criteria. For these reasons, *CV* should be the standard statistic in studies of variation.

Quantifying and Measuring Standardization

The *CV* is commonly used in other natural sciences such as medicine, biology, and psychology. Although some archaeological studies have made qualitative comparisons of *CV*s (e.g., Arnold 1991; Benco 1988; Longacre et al. 1988; Torrence 1986), quantitative analyses with this statistic are notably absent. It has even been argued that it is not possible to test the statistical significance of *CV* (e.g., Arnold and Nieves 1992; Blackman et al. 1993). This is not so. Statistical research provides several techniques for creating confidence intervals and testing equality of *CV* (Bennett 1976; Doornbos and Dijkstra 1983; Gupta and Ma 1996; Vangel 1996), some of which are robust to departures from normality (Feltz and Miller 1996).

Many archaeological studies rely on the *F*-ratio test to compare variation. However, as Kvamme et al. (1996) have pointed out, this test requires normality in the underlying sample populations, an assumption that does not hold in many archaeological situations. Instead, they recommend use of alternative homogeneity of variance (HOV) tests, such as the Brown-Forsyth test (Brown and Forsythe 1974), that are robust to departures from normality (see Conover et al. 1983 for a comparison and discussion of over 50 HOV tests). Unfortunately, use of HOV tests, even those that are robust to non-normality, are of little use in studying variation unless the analyst is certain that the means of the samples being compared are approximately equal. This is

Table 1. Average and Range of CV (Percentage) Values for Various Material Artifact Data Sets.

Data Set	Avg. CV (%)	Range of CV (%)	Source
Machine-Produced Items	.1	.1 – .2	Eerkens 2000
Weber Fraction	1.6	1.6 – 1.7	Ogle 1950
Pots by specialized potters	4	2 – 6	Longacre 1999
Cut–outs from mental image	5	2.5 – 8	Eerkens n.d.
Duna Are Kou	10	8 – 11	White and Thomas 1972
Chaco Canyon Manos	17	8 – 35	Cameron 1997
English Mesolithic Microliths	19	5 – 39	Eerkens 1997, 1998
Great Basin projectile points	22	6 – 55	Bettinger and Eerkens 1997
Owens Valley Handstones	22	10 –32	This article
Random Uniform Data	58	50 – 65	This article
Stylistic elements on SW pots	66	46 – 84	Kantner 1999
Safety pin brooch attributes	74	25 – 113	Doran and Hodson 1975:224

because, as shown above, variance is often scaled to the mean. A standard deviation of five indicates something quite different in a sample with a mean of 10 ($CV = 50$ percent) than in a sample with a mean of 100 ($CV = 5$ percent). Tests for HOV are not sensitive to this, and only compare absolute measures of variance. Unless variation is scale-independent or sample means are approximately equal, HOV tests should not be used in studies of artifact variation.

Tests comparing *CV*, on the other hand, are sensitive to differences in magnitude or mean. Moreover, *CV* is a reliable statistic even at small sample sizes (Simpson 1947; Simpson et al. 1960). For this reason, the *CV* is appropriate for archaeological studies comparing sample variation. Unfortunately the techniques have not yet been incorporated into popular statistical packages (Reh and Scheffler 1996). Presented below is the formula for one such test developed by Feltz and Miller (1996) that is reasonably robust to departures from normality and allows simultaneous comparison of CVs from k sample populations with unequal sample sizes. This statistic is recommended for use in standardization and variation studies.

$$D'AD = \frac{\sum_{j=1}^{k} m_j \left(\frac{s_j}{\bar{x}_j} - D \right)^2}{D^2(0.5 + D^2)}, \text{ where } D = \frac{\left(\sum_{j=1}^{k} m_j \bullet \frac{s_j}{\bar{x}_j} \right)}{\sum_{j=1}^{k} m_j}$$

In D'AD, k is the number of samples, j is an index referring to the sample number, n_j is the sample size of the jth population, $m_j = (n_j - 1)$, s_j is the standard deviation of the jth population, and $\bar{x}_j$ is the mean of the jth population. $D'AD$ is distributed as a χ^2 random variable with $k - 1$ degrees of freedom, and basically describes how far sample *CV*s lie from the estimate of the overall population *CV*. Unlike the Brown-Forsyth statistic, $D'AD$ is simple to determine and can be computed from summary statistics only (mean, standard deviation, and number of samples).

Examples

How do archaeological samples stack up against the *CV* boundary values of 1.7 percent and 57.7 percent presented earlier? Table 1 lists the average and the range of *CV* values for various attributes on material artifacts from a variety of studies. This sample is nonrandom and obviously incomplete, but represents a range of artifact and attribute types typically encountered by archaeologists. Obviously, there is much variation in *CV*s across these data sets. Items made by a few people, such as Philippine pots (Longacre 1999) and Duna Are Kou stone tools (White and Thomas 1972), are much more standardized than generalized assemblages of microliths from England (Eerkens 1997, 1998) and projectile points from the Great Basin (Eerkens and Bettinger 1997), which were likely made by hundreds if not thousands of different flintknappers. Similarly, artifacts typically considered functional, such as projectile points from the Great Basin and manos from Chaco Canyon (Cameron 1997) have much lower *CV* values than attributes typically considered stylistic, such as line elements painted on Southwestern pots (Kantner 1999) or Swiss Bronze Age safety pin brooches (Doran and Hodson 1975). *CV* values on the latter often exceed 57.7 percent, suggesting individuals were resisting conformity to a central or ideal template.

The amount of variation in most of these cases is

far above the minimum humans are able to detect (at 1.7 percent) and produce (at 2–4 percent), often by a factor of 10 or more. This may stem from several factors. First, people may accept visually detectable variation (more than 1.7 percent) because within some margin an artifact may be close enough to the ideal shape that spending more time modifying it is not worth the extra effort (i.e., to possibly obtain a small increase in performance). In other words, beyond some point, imperfect artifacts may still be good enough. This concept has been referred to elsewhere as design constraint or design tolerance (Aldenderfer 1990; Bleed 1986, 1997). Items needed for exact or specialized work are likely to have high design constraints (low tolerance for deviation from the optimal shape), and should display lower *CV*s than less-specialized tools.

Second, as we have seen, the number of people responsible for a set of artifacts may be important. Different people are likely to have slightly different ideas and definitions of what constitutes an "ideal" shape for a particular item. As such, samples of artifacts that archaeologists typically compare when studying variation may differ simply as a result of the number of manufacturers contributing to samples. For example, Eerkens (1997, 1998) has compared Later Mesolithic microliths from generalized site contexts, likely representing numerous individuals, with those from specialized "hoard" or "group" find-spots representing the work of a single individual. Not surprisingly, *CV*s from the latter are much smaller than the former. Routinization is likely to play a role here as well. Large numbers of artifacts made over a short amount of time with a similar and well-remembered mental image will have lower *CV* values than those made one at a time over a longer period of time.

Third, archaeologists may unknowingly group artifacts that were considered distinct by their makers, thereby artificially increasing *CV* values. In other words, elevated *CV*s may be a product of the etic categories archaeologists define, as opposed to the emic categories and restricted *CV*s manufacturers were originally working with. Longacre et al. (1988) has recognized this problem in an ethnoarchaeological study of Kalinga pots, where inadvertent lumping of multiple size classes of pots by archaeologists led to artificially inflated values of variation.

Finally, different raw materials, such as clay and stone, exhibit different forming properties. Some, such as clay, are easier to control, form, and modify, while others, such as flaked stone, are less predictable and controllable, and can only be modified through further reduction of the artifact. Media that are more difficult to control are likely to have inflated *CV* values. Of course, standardization and design tolerances are relative to different media, technologies, and intended artifact functions (Aldenderfer 1990; Bleed 1997:100). Thus, *CV*s that might be considered standardized within a flaked-stone technology producing projectile points may not be in a clay technology producing pots, clay being easier to shape. The study of each technology will need to empirically derive *CV* values that represent what is called "standardized."

Our point here is that there is a limit to how standardized things can get, based on the human ability to differentiate size. In the example above, people are likely to see that, in an absolute sense, there is more variability among projectile points than pots. However, the effort that it would take to make the projectile points as standardized as the pots through additional careful flaking may not be worth whatever benefits might accrue. In this sense, we *can* compare standardization and variation between different technologies. However, the results might tell us more about the inherent difficulties in controlling different media than whether one technology is more standardized, and that people were more careful or concerned about it, than another.

Table 2 uses seven samples drawn from Table 1 to illustrate the superiority of the *D'AD* test over the *F*-ratio test. The *F*-ratio test (see Runyon and Haber 1988:324), which is occasionally used in archaeological studies (e.g., Arnold 1991; Arnold and Nieves 1992; Longacre et al. 1988), examines the ratio of squared sample standard deviations (sample variances) to test equality of variance. Unlike *D'AD*, then, *F*-ratio does not incorporate sample mean. The samples compared in Table 2 include attributes that are typically considered "functional" (microlith length and thickness, projectile-point length), attributes typically considered "stylistic" (Swiss safety-pin bow width, painted line width on black-on-white ceramics), as well as two sets of random data drawn from uniform distributions with means of 10 and 1. The *D'AD* tests clearly demonstrate distinct differences between the "functional" and "stylistic" attributes. The *CV*s of all functional samples are statistically distinct from all stylistic samples (i.e., p

Table 2: Statistical Comparison of Variation *CV* in Various Assemblages; *F*-Ratio (Using Standard Deviation) in Upper Right, and *D'AD* (Using *CV*) in Lower Left.

F-Ratio→ D'AD	Microlith length	Microlith thick.	DSN length	SW pot line wid.	Brooch bow wid.	Random Data 1	Random Data 2
Prestatyn microlith length ($\bar{x}$ = 20.32; s = 4.83; n = 16)		180.0 p = .000	1.862 p = .08	43.78 p = .000	1.9 p = .07	1.28 p = .25	74.1 p = .000
Prestatyn microlith thick. ($\bar{x}$ = 1.97; s = .36; n = 25)	1.24 p = .27		96.94 p = .000	4.11 p = .000	95.52 p = .000	230.87 p = .000	2.43 p = .01
Owens Valley DSN length ($\bar{x}$ = 22.05; s = 3.54; n = 28)	3.03 p = .08	0.4 p = .52		23.52 p = .000	1.0 p = .47	2.39 p = .004	39.82 p = .000
SW pots line width ($\bar{x}$ = 1.12; s = .73; n = 190)	6.96 p = .01	15.21 p = .000	19.28 p = .000		22.99 p = .000	56.15 p = .000	1.69 p = .02
Brooch bow width ($\bar{x}$ = 6.57; s = 3.5; n = 30)	6.71 p = .01	14.42 p = .000	24.84 p = .000	0.97 p = .32		2.44 p = .003	38.9 p = .000
Random Data 1 ($\bar{x}$ = 10.03; s = 5.47; n = 50)	6.71 p = .01	17.11 p = .000	23.14 p = .000	1.24 p = .27	.012 p = .91		95.1 p = .000
Random Data 2 ($\bar{x}$ = .93; s = .561; n = 50)	8.62 p = .003	18.4 p = .000	24.5 p = .000	.244 p = .62	.327 p = .57	.3 p = .58	

Notes: wid. = width, thick. = thickness, DSN = desert side-notched projectile point.

< .05). This is not true with the *F*-ratio tests, which in several instances failed to distinguish ($p > .05$) a functional attribute from a stylistic one (e.g., microlith and projectile-point length are statistically indistinguishable from both brooch bow width and random data). Moreover, variation in the two random data sets is statistically equivalent by the *D'AD* test but statistically different according to the *F*-ratio. The results demonstrate that the *F*-ratio test is inadequate for the task of comparing variation and evaluating degree of standardization.

Discussion and Conclusions

Most archaeological studies of technology recognize only the role of the physical and social environment in shaping material culture (Bleed 1997:98) by focusing on how a raw material is modified using various tools, and how different social and physical processes influence the final product (Schiffer and Skibo 1997). As Bleed (1997:98) has discussed, the human body has seldom been seen as part of this process. As we hope to have made clear, the human body, with all of its attendant sensory systems and limitations, is a medium through which technology operates. Our abilities to see, feel, and modify material items are limited and affected not only by culture, but by the physics and psychophysics of the human body as well.

Understanding these limitations can help archaeologists to ask new questions from the material record. As we have shown above, the psychophysical limitations of size discrimination quantified by the Weber Fraction can help in recognizing different modes of artifact production and degrees of standardization. Weber fractions also have implications in symbolic archaeology because humans are limited in their ability to view, interpret, and discriminate artifacts in the same way they are limited in their ability to produce them in standardized form. Thus, two potters using color and size of painted design elements to differentiate their products must make them different enough that they exceed the just-noticeable-difference (derived from the Weber fraction) for color contrast and size. Even if we, as archaeologists, can discriminate finer differences using Munsell color charts, rulers, calipers, or Scanning Electron Microscopes (SEM), prehistoric people may not have been able to.

Finally, we feel the research is of relevance to studies of artifact change and the evolution of technologies through time. The study demonstrates that people are unable to differentiate subtle differences in the size of objects beyond a certain point. In the transmission of cultural information these limits are just as applicable, affecting how accurately people can copy from and learn from others, and how precisely artifact traits will be transmitted between people. Although beyond the scope of this paper, it should be possible to use the Weber fraction to make some predictions about the degree of drift expected in artifact populations through time, if people are attempting to faithfully copy traits and are randomly making small errors due to the limits of visual perception. These predictions could be tested against the

archaeological record to see if changes in artifacts follow those expected under drift. If variation is less than this value, other variation-minimizing forces may be at work. Alternatively, if variation exceeds this level, various variation-inflating forces may be responsible.

In the last analysis, size-correlated error tolerance is probably telling us something important about the evolutionary setting in which humans evolved, specifically about the penalties suffered in matching tool size to intended task or duplicating tools made by others. The evidence would suggest that errors became, or were perceived as becoming, more costly as tool size decreased. In such a context small tools are specialized tools by definition. Alternatively, the Weber fraction for estimating size may have evolved in an altogether different context, perhaps foraging where, as prey size decreases, absolute error in estimating prey size increases return-rate variability, hence risk of resource shortfall. If so, one would expect to find evidence of size-correlated error tolerance in a wide range of species other than humans. We are unaware of any animal studies of this phenomenon, though these would clearly be worth pursuing as would studies comparing size-correlated error between different hominid forms.

In sum, we have presented evidence showing that *CV* should be the standard statistic in studies of variation and have offered two baseline measures for placing observed *CV*s of length measurements along a continuum of variation from 1.7 percent, the limit of human ability to perceive a difference in size, to 57.7 percent, the variability expected when production is random or near-random and uniform (i.e., completely unstandardized). Using the *CV*, future studies can use these baseline values to evaluate the degree of variation or standardization in independent sets of artifacts. The *D'AD* test facilitates statistical comparison of *CV*s from samples of artifacts of differing size or magnitude to help evaluate degree of standardization. We hope that further exploration of the psychophysical literature will lead to a deeper understanding of how technology and the human body interact to create the material record, and variation therein, that we study.

Acknowledgments: Thanks to Mark Aldenderfer, Peter Bleed, Mike Jochim, John Kantner, Dwight Read, Kevin Vaughn, and an anonymous reviewer for reading earlier drafts of this paper. Their comments greatly helped to improve the final product. We also kindly thank Tim Kohler for critique, suggestions, and editorial comments, and Elizabeth Klarich and Christina Torres for their Spanish skills in translating the abstract.

References Cited

Aldenderfer, M. A.
1990 Defining Lithics-Using Craft Specialists in Lowland Maya Society through Microwear Analysis: Conceptual Problems and Issues. In *The Interpretative Possibilities of Microwear Studies*, edited by B. Graslund et al., pp. 53–70. Societas Archaeologica Upsaliensis, Uppsala, Sweden.

Algom, D.
1992 Memory Psychophysics: An Examination of Its Perceptual and Cognitive Prospects. In *Psychophysical Approaches to Cognition*, edited by D. Algom, pp. 441–512. Elsevier, New York.

Arnold, D. E., and A. L. Nieves
1992 Factors Affecting Ceramic Standardization. In *Ceramic Production and Distribution*, edited by G. J. Bey III and C. A. Pool, pp. 93–113. Westview, Boulder, Colorado.

Arnold, P. J.
1991 Dimensional Standardization and Production Scale in Mesoamerican Ceramics. *Latin American Antiquity* 2:363–370.

Benco, N.
1988 Morphological Standardization: An Approach to the Study of Craft Specialization. In *A Pot for All Reasons: Ceramic Ecology Revisited*, edited by C. Kolb and L. Lackey, pp. 57–72. Temple University, Philadelphia.

Bennett, B. M.
1976 On an Approximate Test for Homogeneity of Coefficients of Variation. *Contributions to Applied Statistics*, edited by W. J. Ziegler, pp. 169–171. Birhauser Verlag, Stuttgart, Germany.

Bettinger, R. L., and J. W. Eerkens
1997 Evolutionary Implications of Metrical Variation in Great Basin Projectile Points. In *Rediscovering Darwin: Evolutionary Theory and Archaeological Explanation*, edited by C. M Barton and G. A. Clark, pp. 177–191. Archaeological Papers of the American Anthropological Association, Arlington, Virginia.
1999 Point Typologies, Social Transmission and the Introduction of Bow and Arrow Technology in the Great Basin. *American Antiquity* 64:231–242.

Blackman, J. M., P. B. Vandiver, and G. J. Stein
1993 Standardization Hypothesis and Ceramic Mass Production: Technological, Compositional, and Metric Indexes of Craft Specialization at Tell Leilan, Syria. *American Antiquity* 58:60–80.

Bleed, P.
1986 The Optimal Design of Hunting Weapons: Maintainability or Reliability. *American Antiquity* 51:737–847.
1997 Content as Variability, Result as Selection: Toward a Behavioral Definition of Technology. In *Rediscovering Darwin: Evolutionary Theory and Archaeological Explanation*, edited by C. M. Barton and G. A. Clark, pp. 95–103. Archaeological Papers of the American Anthropological Association, Arlington, Virginia.

Brown, M. B., and A. B. Forsythe
1974 Robust Tests for the Equality of Variance. *Journal of the American Statistical Association* 69:364–367.

Cameron, C. M.
1997 An Analysis of Manos from Chaco Canyon, New Mexico. In *Ceramics, Lithics, and Ornaments of Chaco Canyon, Volume III*, edited by J. Mathien, pp. 997–1012. National Park Service, Santa Fe, New Mexico.

Chase, P. G.

1991 Symbols and Paleolithic Artifacts: Style, Standardization, and the Imposition of Arbitrary Form. *Journal of Anthropological Archaeology* 10:193–214.

Conover, W. J., M. E. Johnson, and M. M. Johnson
1981 A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data. *Technometrics*: 23, No. 4, 351–361.

Coren, S., L. M. Ward, and J. T. Enns
1994 *Sensation and Perception*. 4th ed. Harcourt Brace, Fort Worth, Texas.

Costin, C. L., and M. B. Hagstrum
1995 Standardization, Labor Investment, Skill, and the Organization of Ceramic Production in Late Prehispanic Highland Peru. *American Antiquity* 60:619–639.

Crown, P. L.
1999 Socialization in American Southwest Pottery Decoration. In *Pottery and People*, edited by J. M. Skibo and G. M. Feinman, pp. 25–43. University of Utah Press, Salt Lake City.

Dobres, M.
1995 Gender and Prehistoric Technology: On the Social Agency of Technical Strategies. *World Archaeology* 27:25–49.

Doornbos, R., and J. B. Dijkstra
1983 A Multi Sample Test for the Equality of Coefficients of Variation in Normal Populations. *Communications in Statistics: Simulation and Computation* 12:147–158.

Doran, J. E., and F. R. Hodson
1975 *Mathematics and Computers in Archaeology*. Harvard University, Cambridge, Massachusetts.

Eerkens, J.W.
1997 Variability in Later Mesolithic Microliths of Northern England. *Lithics* 17/18:51– 65.
1998 Reliable and Maintainable Technologies: Artifact Standardization and the Early to Later Mesolithic Transition in Northern England. *Lithic Technology* 23:42–53.
2000 Practice Makes Within 5% of Perfect: The Role of Visual Perception, Motor Skills, and Human Memory in Artifact Variation and Standardization. *Current Anthropology* 41:663–668.

Feltz, C. J., and G. E. Miller
1996 An Asymptotic Test for the Equality of Coefficients of Variation from K Populations. *Statistics in Medicine* 15:647–658.

Gescheider, G. A.
1997 *Psychophysics: The Fundamentals*. L. Erlbaum, Hillsdale, New Jersey.

Gupta, R. C., and S. Ma
1996 Testing the Equality of Coefficients of Variation in K Normal Populations. *Communications in Statistics: Theory and Method* 25:115–132.

Hayden, B., and R. Gargett
1988 Specialization in the Paleolithic. *Lithic Technology* 17:12–18.

Hotopf, W. H. N., M. C. Hibberd, and S. A. Brown
1983 Position in the Visual Field and Spatial Expansion. *Perception* 12:469–476.

Howard, I. P., and B. J. Rogers
1995 *Binocular Vision and Stereopsis*. Oxford University Press, Oxford.

Jones, L. A.
1986 Perception of Force and Weight: Theory and Research. *Psychological Bulletin* 100:29–42

Kantner, J.
1999 The Influence of Self-Interested Behavior on Sociopolitical Change: The Evolution of the Chaco Anasazi in the Prehistoric American Southwest. Unpublished Ph.D. dissertation. Department of Anthropology, University of California, Santa Barbara, CA.

Kerst, S. M., and J. H. Howard, Jr.
1978 Memory Psychophysics for Visual Area and Length. *Memory & Cognition* 6:327– 335.
1981 Memory and Perception of Cartographic Information for Familiar and Unfamiliar Environments. *Human Factors* 23:495–503.
1984 Magnitude Estimates of Perceived and Remembered Length and Area. *Bulletin of the Psychonomic Society* 22:517–520.

Kvamme, K. L., M. T. Stark, and W. A. Longacre
1996 Alternative Procedures for Assessing Standardization in Ceramic Assemblages. *American Antiquity* 61:116–126.

Laming, D. R. J.
1997 *The Measurement of Sensation*. Oxford University Press, Oxford.

Longacre, W. A.
1999 Standardization and Specialization: What's the Link? In *Pottery and People*, edited by J. M. Skibo and G. M. Feinman, pp. 44–58. University of Utah Press, Salt Lake City.

Longacre, W. A., K. L. Kvamme, and M. Kobayashi
1988 Southwestern Pottery Standardization: An Ethnoarchaeological View from the Philippines. *Kiva* 53:101–112.

Mather, G.
1997 The Use of Image Blur as a Depth Cue. *Perception* 26:1147–1158.

Miller, G.
1996 The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review* 63:81–97.

Moyer, R. S., D. R. Bradley, M. H. Sorenson, J. C. Whiting, and D. P. Mansfield
1978 Psychophysical Functions for Perceived and Remembered Size. *Science* 200:330–332.

Norwich, K. H.
1983 On the Theory of Weber Fractions. *Perception and Psychophysics* 42:286–298.

Ogle, K. N.
1950 *Researches in Binocular Vision*. Saunders, Philadelphia.

Pagano, C. C., and K. G. Donahue
1999 Perceiving the Lengths of Rods Wielded in Different Media. *Perception and Psychophysics* 61:1336–1344.

Poulton, E. C.
1989 *Bias in Quantifying Judgments*. Lawrence Erlbaum, Hove, England.

Reh, W., and B. Scheffler
1996 Significance Tests and Confidence Intervals for Coefficients of Variation. *Computational Statistics and Data Analysis* 22:449–452.

Rice, P. M.
1991 Specialization, Standardization, and Diversity: A Retrospective. In *The Ceramic Legacy of Anna O. Shepard*, edited by R. L. Bishop and F. W. Lange, pp. 257- 279. University of Colorado, Boulder.

Ross, H. E.
1981 How Important Are Changes in Body Weight for Mass Perception? *Acta Astronautica* 8:1051–1058.
1995 Weber on Temperature and Weight Perception. In *Fechner Day 95*, edited by C. A. Possamaï, pp. 29–34. International Society for Psychophysics, Cassis, France.
1997 On the Possible Relations between Discriminability and Apparent Magnitude. *British Journal of Mathematical and Statistical Psychology* 50:187–203.

Ross, H. E., and R. L. Gregory
1964 Is the Weber Fraction a Function of Physical or Perceived Input? *Quarterly Journal of Experimental Psychology* 16:116–122.

Rottlander, R. C. A.
1966 Is Provincial-Roman Pottery Standardized? *Archaeometry* 9:76–91.
Rowe, J. H.
1978 Standardization in Inca Tapestry Tunics. In *Junius B. Bird Pre-Columbian Textile Conference,* edited by A. P. Rowe, E. P. Benson, and A. Schaffer, pp. 239–264. Dumbarton Oaks, Washington DC.
Runyon, R. P., and A. Haber
1988 *Fundamentals of Behavioral Statistics,* 6th ed. Random House, New York.
Schwartz, S. H.
1999 *Visual Perception.* 2nd ed. Appleton and Lange, Stamford, Connecticut.
Schiffer, M. B., and J. M. Skibo
1997 Explanation of Artifact Variability. *American Antiquity* 62:27–50.
Shott, M. J.
1997 Transmission Theory in the Study of Stone Tools: A Midwestern North American Example. In *Rediscovering Darwin: Evolutionary Theory and Archaeological Explanation,* edited by C. M. Barton and G. A. Clark, pp. 193–204. Archaeological Papers of the American Anthropological Association, Arlington, Virginia.
Simpson, G. C.
1947 Note on the Measurement of Variability and on Relative Variability of Teeth of Fossil Mammals. *American Journal of Science* 245:522–525.
Simpson, G. C., A. Roe, and R. C. Lewontin
1960 *Quantitative Zoology.* Harcourt, Brace, New York.
Smallman, H. S., D. I. A. MacLeod, S. He, and R. W. Kentridge
1996 Fine Grain of the Neural Representation of Human Vision. *Journal of Neuroscience* 16:1852–1859.
Stevens, J. C.
1979 Thermal Intensification of Touch Sensation: Further Extensions of the Weber Phenomenon. *Sensory Processes* 3:240–248.
Stevens, S. S.
1975 *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects.* Wiley, New York.
Teghtsoonian, R.
1971 On the Exponents in Stevens' Law and the Constant in Ekman's law. *Psychological Review* 78:71–80.
Torrence, R.
1986 *Production and Exchange of Stone Tools.* Cambridge University Press, Cambridge.
Vangel, M. G.
1996 Confidence Intervals for a Normal Coefficient of Variation. *American Statistician* 15:21–26.
Verrillo, R. T.
1981 Absolute Estimation of Line Length in Three Age Groups. *Journal of Gerontology* 36:625–627.
1982 Absolute Estimation of Line Length as a Function of Sex. *Bulletin of the Psychonomic Society* 19:334–335.
1983 Stability of Line-Length Estimates Using the Method of Absolute Magnitude Estimation. *Perception & Psychophysics* 33:261–265.
Weber, E. H.
1834 *De Pulen, Resorptione, Auditu et Tactu: Annotationes Anatomicae et Physiologicae.* Kohler, Leipzig, Germany.
White, J. P., and D. H. Thomas
1972 What Mean These Stones? Ethno-Taxonomic Models and Archaeological Interpretations in the New Guinea highlands. In *Models in Archaeology,* edited by D. L. Clarke, pp. 275–308. Methuen, London.

Notes

1. This ensues becuase the mean of a uniform distribution on [0,X] is X/2 and the standard deviation is $X/\sqrt{12}$. Thus, the *CV* is $2/\sqrt{12}$, or $1/\sqrt{3}$, or 57 percent. Note that this value only applies when the lower limit of the interval is set to 0. When the interval is more narrow, starting at a value greater than 0, a smaller *CV* ensues.

2. Normal distribution in human perception and manufacture are more likely. However, modeling the Weber fraction as a normal variable requires *a priori* definition of a standard deviation or error value, which is what we are trying to model in the first place. As such, a uniform distribution is used.

Received April 22, 1999; Revised March 13, 2000; Accepted July 18, 2000.