

# University of Liverpool Research Archive (UoLRA)

NOTICE: this is the author's version of a work that was accepted for publication in Archives of Oral Biology. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in ARCHIVES OF ORAL BIOLOGY, [VOL 54, SUPPLEMENT 1, (2009)] <http://dx.doi.org/10.1016/j.archoralbio.2008.04.010>.

## **Accounting for Measurement Error: A Critical but Often Overlooked Process**

E.F. Harris<sup>a</sup>, R.N. Smith<sup>b</sup>

<sup>a</sup>Department of Orthodontics The Health Science Center University of Tennessee  
Memphis, Tennessee 38163 USA. <sup>b</sup>Department of Clinical Dental Sciences, University of  
Liverpool, Edwards Building Daulby Street, Liverpool, UK

**Tables:** 3

**Figures:** 4

### **Running Title:**

Technical Errors of Measurement

### **Key words:**

Measurement error

Reliability

Validation

Correlation

## **ABSTRACT**

Because of instrument imprecisions and human inconsistencies, measurements are not free of error. Technical error of measurement (TEM) is the variability encountered among dimensions when the same specimens are measured at multiple sessions. A goal of a data collection regimen is to minimize TEM. The few studies that actually quantify TEM— regardless of discipline—report that it is substantial and can affect results and inferences.

**Objective:** This paper reviews some statistical approaches for identifying and controlling TEM. Statistically, TEM is part of the residual (“unexplained”) variance in a statistical test, so accounting for TEM—which requires repeated measurements—enhances the chances of finding a statistically significant difference if one exists.

### **Methods:**

It has been the author’s intention to perform a thorough review and discuss statistical design relating to types of error and statistical approaches to error accountability. This paper address’ issues of landmark location, validity, technical and systematic error, ANOVA, scaled measures and correlation coefficients in order to guide the reader towards correctly identify true experimental differences.

### **Conclusions:**

Researchers commonly infer characteristics about populations from comparatively restricted study samples. Most inferences are statistical, and, aside from concerns about adequately accounting for known sources of variation with the research design, an important source of variability is measurement error. Variability in locating landmarks that define variables is obvious in odontometrics, cephalometrics and anthropometry, but

the same concerns about measurement accuracy and precision extend to all disciplines. With increasing accessibility to computer-assisted methods of data collection, the ease of incorporating repeated measures into statistical designs has improved. Accounting for this technical source of variation increases the chance of finding biologically true differences when they exist.

## INTRODUCTION

There is not much of research interest that can be measured without error. Consider measuring the same object like a tooth several times. Part of the inevitable variability of these measurements is due to the finite consistency and read-out precision of the instrument used to measure the object and the other is due to human inconsistency.

Multiple measurements of the same variable will not always be the same because of variability in the measurement process<sup>1,2</sup>. There are some obvious ways to reduce intra-observer repeatability<sup>3,4</sup>, such as exactly defining the landmarks that determine a measurement, enhancing observer experience (and, thus, consistency), and avoiding fatigue, but wholly eliminating this source of variation is difficult.

Suppose a specimen's true trait size (for some dimension of interest) is  $\theta_i$  and that this dimension is repeatedly measured with some device (*e.g.*, ruler, callipers, a computer program), then the observed value ( $X_{ij}$ ) will be

$$X_{ij} = \theta_i + \eta_{ij}$$

where  $\theta_i$  is the true size of the feature being measured on specimen  $i$ , and  $\eta_{ij}$  is the error of measurement<sup>5</sup>. True size is what we want to capture when we take a measurement. In practice,  $\eta$  will seldom be zero because of variability in the measuring device and how it is used (*i.e.*, variability introduced by the observer). One supposes that  $\theta_i$  is constant (*i.e.*, a theoretical construct that would be obtained if the specimen were re-measured innumerable times), while  $\eta_{ij}$  varies among data collection sessions. One goal of data collection is to minimize the  $\eta_{ij}$ . One goal of data analysis is to quantify this source of variation and, hopefully, remove its influence from interpretation of the biological differences being tested. A necessary statistical assumption here is that trait size ( $X_{ij}$ ) is

independent of the error of measurement (and this is easily testable).

Innumerable dimensions of a tooth can be constructed<sup>6,7,8,9</sup>, but none can be measured without error. In fact, there are two complementary issues: accuracy and precision<sup>3,10</sup>.

Accuracy is the closeness of measured values to the true value. Precision, in contrast, is the closeness of repeated measurements of the same quantity (Fig. 1). Importantly,

“Unless there is bias in a measuring instrument, precision will lead to accuracy”<sup>10</sup>.

Consequently, unless there is some reason to suspect precision issues, attention needs to focus on improving measurement accuracy which means reducing intra-observer error.

(Between-observer differences commonly are larger than within-observer differences

<sup>11,12</sup>, but this topic is ignored here to conserve space; moreover, most of the concepts reviewed here are directly applicable to issues of inter-observer reliability). The

reproducibility of a measurement is depressed when there are problems with accuracy and/or precision<sup>2</sup>.

## **Validity**

Another term to be introduced here is validity. Validity is defined differently in various disciplines of research; we mention just two of these, namely (1) construct validity and (2) statistical validity. Construct validity is an issue of importance in fields such as psychology and sociology, but is seldom encountered in the biological sciences where the dependent variable (*e.g.*, cell density, tissue thickness, tooth pulp volume) is amenable to direct mensuration. Roughly, construct validity involves methods such as questionnaires and preference scales to estimate an underlying latent variable such as beauty or intelligence or loneliness. One area of concern here is whether the test (a set of

questions, psychometric exercises, *etc.*) actually measures what it is intended to measure<sup>13,14</sup>. A second, statistical concern is how to formulate the optimal test instrument that best quantitates the underlying construct; this procedure now relies on multivariate techniques, notably factor analysis<sup>15,16</sup>.

In contrast, statistical validity refers to the degree to which an observed result can be relied on not to result from technical errors of measurement plus “other” statistical considerations, but “other considerations” involve the very broad topic of appropriate statistical methodologies. We can do little more in this brief space other than raise the issue that “Inappropriate statistical methods, as well as appropriate methods inappropriately used, can lead to incorrect conclusions of any research report”<sup>17</sup>. The huge growth of statistical methods in recent years has made the choice of statistical analysis let alone how to do the tests correctly harder and harder for the researcher. Each journal seems to contain an occasional article criticizing the high frequency and severity of statistical mistakes seen in that discipline’s publications<sup>18-25</sup>. These critiques, generally written by biostatisticians, identify errors ranging from the simple to the complex, and their concluding remarks typically are along the lines (1) that proper statistical analysis is complex, with a good dose of art as well as science in the analysis, so (2) the research team needs to collaborate with a statistician during all stages of the project. Unfortunately, these disclosures take the researcher no closer to understanding because, again, the field of statistical analysis has burgeoned, and the knowledge base far exceeds what people not wholly devoted to this specialty can claim competence in. Statistical validity involves a broad range of considerations, including (1) appropriate

research design, along with whether the data collected actually pertain to the question being asked; (2) appropriate level of data (nominal, ordinal, interval, ratio)<sup>26</sup>; (3) adequate sample sizes<sup>27</sup>; (4) data meeting assumptions of the inferential statistical tests<sup>10, 28</sup>; (5) that the tests themselves are appropriate and efficient; and (6) that inferences drawn from the tests are appropriate<sup>29</sup>.

## **Technical Measurement Error**

The only way to quantify TEM is by taking repeated measurements on the same objects. It generally is assumed that the mean of a series of repeated measurements is the best available estimate of an object's true size. It has been conventional<sup>30</sup> to take just two measurements per specimen (one measurement and a repeated measurement), but this actually is just the lower limit, and it can be useful to increase the repetitions<sup>1,31</sup> to better assess this source of unintended variability.

Why bother with repeatability error? There are several reasons, but an important perspective is that repeatability errors are part of the residual term in most any statistical test<sup>5,10</sup>. Reducing repeatability error increases the among-to-within ratio of variances, thereby increasing the chances of finding a statistically significant difference if one exists.

Repeatability errors can be random or systematic<sup>32</sup>. Examples of systematic errors can be constant (due to personal biases or, perhaps, handedness) or they can occur across time (e.g., progressive accuracy with greater experience, or a shift in measurer's style, or landmark interpretations) or among instruments. For example, radiographic

(cephalometric) studies are susceptible to systematic errors because the X-ray source is at a finite distance from the object and film, so there is some magnification. The amounts of magnification can differ systematically between machines<sup>33,34</sup>. Not correcting for magnification—on the order of 6-8% for most cephalometric arrangements<sup>35,36</sup>—systematically overestimates true size, and comparing linear distances from X-rays taken on different machines is likely to produce systematic biases if magnification is not taken into account. So too, as a child's head grows with age, his mid-sagittal plane is moved farther from the film, which systematically increases image magnification, and peripheral structures are enlarged more than those centred in the X-ray beam. The use of helical computer tomography and other sophisticated three-dimensional systems will help control for magnification error, but errors in landmark location seem unavoidable (but are improved with greater pixel resolution)<sup>37</sup>.

Systematic errors hardly are limited to cephalometrics, however. Systematic errors commonly are detectable with statistics that test for differences in sample means (*e.g.*, t-tests, ANOVA, sign tests), but variation due to random influences requires less-obvious sorts of analysis. The present discussion focuses primarily on random sources of variation, notably due to variability in an observer's assessments<sup>38,39</sup>.

The comments in this overview apply to continuous (interval and ratio scale) variables, not those recorded using nominal or ordinal scales where other statistical methods of concordance are better suited<sup>40,41</sup>. So, for example, when the basal area of a molar cusp is measured as a ratio-scale variable<sup>42</sup>, repeatability error can be described using methods reviewed in this paper. When, instead, cusp size is recorded on an interval scale<sup>43</sup>, other

methods of internal agreement are needed.

Cameron<sup>44</sup> points out that the literature on TEM can broadly be categorized into two complementary approaches: One is technical, where the goal is to standardize data collection methods by specifying operational definitions of landmarks, standardizing operator styles and instrumentation and other sources of variability. There is a useful literature describing TEM as it relates to anthropometrics (body measurements), where observers' styles of data collection can be a significant source of variability<sup>4,38,39,45,46</sup>. Critical too are stylistic differences that can affect sample means<sup>2</sup>. The latter is a particular problem if, as often occurs, different observers measure different groups, so the cause of a difference—whether it is due to measurement bias and/or a true biological difference cannot be disentangled.

TEM is particularly influential in longitudinal studies because the sizes of measurement errors are large vis-à-vis the amounts of growth. The closer together the measurements are taken, the greater the chance that TEM can confound the biological differences<sup>47,48</sup>. Perhaps the extreme example is where measurements are taken daily in order to study the episodic (“saltation and stasis”) nature of growth<sup>49,50</sup>.

TEM inflates the dispersion of the measured dependent variable, as reflected in the sample characteristics such as variance, standard deviation, and standard error of the mean. Greater variability thus depresses the chances of finding a statistically significant difference if one exists, thereby increases chances of a Type II error (*i.e.*, accepting the null hypothesis of no difference when it is false). When variance due to TEM still seems

a problem, but it cannot be further reduced technically, one solution is to increase sample size<sup>32,51</sup>. Power analysis—the estimation of sample sizes needed to be reasonably confident of rejecting the null hypothesis when it is false—can be particularly useful at the beginning of a study to evaluate an experiment's practicality with the sample sizes to be studied<sup>27</sup>. Prior studies that used comparable data collection methods probably incorporate TEM as part of their own parameter estimates, so they can practicably be used to estimate needed sample sizes.

## **Correlation Coefficients**

Researchers frequently understand that measurements incorporate some degree of error due to human inaccuracies, but it is not always evident how best to deal with this. A common, intuitive solution is to calculate a Pearson product-moment correlation coefficient ( $r$ ). After all, the reasoning seems to be, correlation measures the association between paired sets of data. Characteristically, though, the resulting correlation is always fairly high unless measurement error is huge. The effect of random (but not systematic) errors is to reduce the strength of the correlation coefficient, but there are several other shortcomings of the correlation coefficient. First,  $r$  is a measure of the mutual relationship between two variables; it does not measure the strength of agreement<sup>52,53</sup>.

Strength of a correlation also depends on the range of the variables:  $r$  can be increased simply by choosing a greater range of variables. This is evident when examining a bivariate plot as in Fig. 2. Viewing the whole graph makes the arrangement of points look pretty linear. Focusing, instead, on any small portion of the graph and the perceived

association is much weaker because of the considerable local scatter. A researcher can, then, enhance the strength of the correlation by being sure to include a large range of points along the axes. Also, Pearson's  $r$  only is sensitive to the linear association between two variables, which can be an unnecessarily restrictive assumption when the degree of reliability is of interest. That is, the correlation coefficient measures the trend throughout the range of the distribution of one variable to be consistently accompanied by a change in the other variable. A more complex model is needed to test for curvilinear associations<sup>54</sup>. Correlation does not measure agreement. Consider two sets of measurements, one consisting of the "correct" measurements and the second equal to the first but with a 1-inch offset (Fig. 3). The correlation is not affected because the measurements still characterize the same straight line, but the agreement is now horrendous.

Additionally, given repeated pairs of measurements, there is no logic in labelling one as the X (independent) and the other as the Y (dependent) variable. This assumption, in itself, negates the use of  $r$  as a measure of reliability. Consider that a correlation coefficient is computed between two columns of data: if the paired values of some rows are swapped, the correlation will be different. There is no fixed correlation between two data series;  $r$  is a variable measure of agreement subject to manipulation<sup>53</sup>.

Use of the intra-class correlation coefficient ( $r_i$ ) provides a solution to the latter problem of treating variables from one measurement session as independent and a second session as dependent<sup>54</sup>. The  $r_i$  is the average correlation regardless of the ordering of pairs of data, so that shortcoming of Pearson's (interclass) correlation is avoided. Moreover, the  $r_i$  can be used to evaluate any number of measurement sessions, not just two. Indeed, the

requisite assumption of  $r_i$  is that the variables of each session have equal means and variances<sup>5</sup>. Thus, this is a better model conceptually, and it also means that just two rather than four parameters need to be estimated (because a common mean and common variance are assumed for the two variables), so the test is more reliable. However,  $r_i$  is less well known to researchers, and few statistical packages calculate  $r_i$  directly. Instead,  $r_i$  commonly has to be calculated from the output of a model II (repeated measures) ANOVA<sup>10</sup>, and this is discussed in a later section. Shortcomings of the intra-class correlation are similar to those of other correlation coefficients:

(1) it is unit-less, so it imparts no information about the magnitude of the differences between measurements (*cf.* Fig. 3) and (2) strength of  $r_i$  can be manipulated by altering the range of observations, which means it is subject to sampling fluctuations that are “hidden” unless its confidence limits also are reported.

## Systemic Errors

Systematic errors are due, for example, to different devices or different kinds of data. We mention here a couple of systematic tooth size differences, just because they are familiar to this audience. Measuring extracted teeth yields systematically larger mesiodistal dimensions than when the teeth are *in situ*. The beaks of callipers, even if machined to fine points, typically cannot get fully into the embrasures to yield a tooth’s maximum MD diameter. The situation is worse if plaster dental casts are used, because the impression material does not preserve the infinitesimal space between tightly approximated teeth. Measurements of isolated teeth therefore are systematically larger mesiodistally than measurements from casts of the same teeth. Another systematic

hindrance in measuring teeth in the living is that the gingival margins commonly overlap the CEJ (cemento-enamel junction), yet the maximum bucco-lingual dimensions of human incisors are commonly subgingival<sup>55</sup>. Gingival recession tends to be age-progressive, so it is common to find larger BL dimensions in older people<sup>56</sup>. Rather than proving that people with bigger incisors live longer, the trend with age needs to be recognized as a systemic source of measurement error because the fiducial points change with time along the base of the crown.

In concept, systematic differences are easy to detect statistically, but this assumes that the nature of the effect is itself simple. For example, differences between measuring sessions could be due to the use of different callipers with differently shaped beaks. Test for a mean difference between sessions can be done with a paired t-test (or a repeated-measures ANOVA if there are more than two sessions). It is important not to rely on the more-familiar group comparison form of the t-test because it is relatively inefficient. With a paired t-test, the standard error (denominator) of the test is smaller because the pairs of data are correlated, so the chances of finding a difference if one exists is greater. The paired t-test is equivalent to testing whether the mean difference between pairs differs from zero. Comparably, the ANOVA design also should capitalize on the repeated nature of the data when testing for mean differences among sessions.

### **Quantifying Technical Error of Measurement**

Historically, Gunnar Dahlberg was the first to provide a formula for repeatability error. Dahlberg first published the formula in 1926<sup>57</sup>, but his statistics textbook published in 1940 is more readily available<sup>30</sup>. The well-known equation is

$$d = \sqrt{\frac{\sum_{i=1}^n (X_{1i} - X_{2i})^2}{2n}}$$

where  $d$  is the standard deviation calculated from two sets of  $n$  repeated measurements (set 1 and set 2) taken from  $i = 1$  to  $n$  specimens. Consider the data in **Table 1** where 20 specimens were measured three times. Dahlberg's  $d$  for T1 and T2 is 0.084, which is just less than a tenth of a millimetre. Dahlberg's  $d$  is the standard deviation of the sample of double determinations, not the average difference. This is a measure of the variability, both random and systematic, due to technical inconsistencies. We assume (1) that this value is the same for all specimens otherwise there is no point in estimating it <sup>1</sup> and (2) by randomly selecting cases to re-measure, we assume that the estimate of method error can be extrapolated to the whole sample.

Solow <sup>58</sup> suggests that  $d$  be termed the method error, though TEM (technical error of measurement) coined by Johnston <sup>45</sup> has become the favoured term in many circles. Dahlberg's  $d$  (occasionally labelled s i) is sometimes labelled Solow's error statistic because Solow described the formula without citing any source. The denominator is  $2n$  because the variance of two observations is one-half their squared difference, so  $n$  in the denominator is the number of paired observations and the "2" halves the value of the numerator <sup>32,59</sup>. The square root sign in the equation yields the standard deviation from the variance. The "Dahlberg formula" is commonly encountered in research publications, though its supposed formulation is rather frequently miss-stated. Mueller and Martorell <sup>46</sup> show how the formula can be generalized (A) to multiple observers (inter-observer repeatability) and/or (B) to more than two sets of measurements.

A number of statistical notes by the statisticians Bland and Altman<sup>1,31,53,60</sup> collectively provide a useful guide to the estimation of repeatability error, and these are mostly available *gratis* on the internet. Bland and Altman promote the use of  $s_w$ , which is the standard deviation of the repeatability error and, synonymously, the within-subject standard deviation<sup>1</sup>.

It is worth reviewing the simplest structure for estimating repeatability, where there are two measurement sessions, because this structure is so commonly encountered. Consider just the first two data columns (T1 and T2) in Table 1. ANOVA produces the results in Table 2. The mean within-subject variance is 0.007, so the within-subject standard deviation is the square-root of this ( $s_w$ ), namely 0.084 mm. This means that most (+ 1 sd) technical errors are expected to be in the range of -0.8 and +0.8 mm around the mean because 1 sd bounds 68.2% of the distribution when the sample is normally distributed. Comparably, + 2 sd bounds about 95% of the distribution of errors. What if, instead, we want to characterize the sampling distribution of the mean error (not the distribution of data points themselves).

What Bland and Altman<sup>1</sup> point out is that  $\sqrt{2}$  times  $s_w$  yields the standard error of the mean (because a standard error is  $\sqrt{sd/n}$  and  $s_w$  is the sd). Therefore,  $\sqrt{2} \times 1.96 \times s_w$  defines the 95% confidence limits of the method error. This range can be useful if different measurement techniques are compared<sup>38</sup>. If, graphically, the 95% confidence limits of two methods do not overlap, that is suggestive that the methods' means differ significantly (at alpha = 0.05). Also, a significantly more accurate measurement method

would be supposed to have smaller confidence limits than the other.

Bland and Altman do not mention it, but when, as is common, there are two measurements per subject (and no missing data), their statistic  $s_w$  can be rewritten as a variance, where  $d \sum d_i^2 / 2n$  is  $(X_{1i} - X_{2i})$ , showing that it is identical to the Dahlberg statistic. The attractiveness of  $s_w$ —and the underlying repeated-measures ANOVA design—is that it can be extended to any number of measurement sessions, more complex models can be tested<sup>5,61</sup>, and sample sizes among sessions need not be equal.

Mention should also be made to the well-known Bland-Altman plot<sup>1,60,62,63</sup>. This graphical approach is easy to conceptualize (Fig. 4); the means of the repeated measures are arrayed along the X-axis and the differences between the corresponding pairs of measurements are plotted on the vertical axis. If the two measurement sessions measured the specimens the same in an unbiased manner, the plot would show a random scatter of differences around a mean of zero. The plot provides a visual sense of whether repeatability error depends on trait size, where either smaller or larger specimens might be at risk of greater intra-observer discrepancies. The Bland-Altman plot was devised to evaluate two measurement techniques, but it likewise provides a simple test to evaluate trends among any sort of repeated measures. If desired, a suspected trend can be tested by regressing the differences on the mean sizes to assess whether trait size is predictive of the magnitude of TEM<sup>63,64</sup>.

## **Model II ANOVA**

There are any number of elegant repeated-measures ANOVA models that can be

designed to accommodate the analysis of method errors<sup>5,47,51</sup>. Shrout and Fleiss<sup>65</sup> detail the analyses of three typical reliability designs: (1) each specimen is measured by a different set of observers, (2) a sample of observers measure specimens but just the specimens' means are analyzed, or (3) specimens are measured by multiple observers, but they are only observers of interest (model I). Having said this, the most commonly encountered situation in the biological sciences seems to be accommodated by a single classification ANOVA model. Take, for example, the table of mesiodistal tooth crown dimensions in Table 1 for a hypothetical set of maxillary central incisors. Measurements were made of 20 specimens on three occasions for a total of 60 observations. The concept is that these teeth were chosen at random from a larger odontometric study. We care little about these 20 specimens themselves; instead, we want to use the results from these 20 specimens as representative of findings applicable to the larger sample. Analogously, we have little interest in the three measurement sessions *per se*; instead, these repetitions are perceived as sessions chosen at random from among the indefinitely large number of repetitions that could have been performed.

The required arrangement of the actual data differs among statistical programs. The JMP package (SAS Corp, Cary, NC) was used to produce Table 2. Just as in prior sections, the goal oftentimes is to obtain the within-subject standard deviation, which Bland and Altman<sup>1</sup> term  $s_w$  for the within-specimen error. Other information is also available, such as how large the variance components are within and among subjects and, importantly, their relative magnitudes which includes the intra-class correlation (ri).

Of course, there will be size variability among the teeth due to some interplay of genetic and environmental differences among individuals. The among-specimen variance is one component of the ANOVA model. If the repeated measurements for each tooth were identical, all of the variance would be among specimens; there would be no additional within-specimen variance. Predictably, though, there are some measurement inconsistencies, and this within-specimen variance adds to the total variance. A common statistical question is what is the relative magnitude of the method error?

Expected specimen mean square (Table 3) is composed of a weighted combination of the variances within and among specimens, so the among-specimen variance  $s_A^2$  alone is  $(0.815 - 0.005)/2 = 0.405$ . This value ( $s_A^2$ ), divorced from measurement error, is the variance of the true measures estimated from the population from which the n specimens were selected.

The intra-class correlation coefficient is the ratio of the among-to-total variances,

$$r_i = \frac{s_A^2}{s^2 + s_A^2} = \frac{0.405}{0.005 + 0.405} = 0.988$$

With an equal number of observations per specimen,  $r_i$  can also be calculated as

$$R_i = \frac{E_s - E}{E_s + (n-1)E} = \frac{0.815 - 0.005}{0.815 + (2-1)0.005} = \frac{0.800}{0.820} = 0.98$$

which is identical accounting for round-off error. (E is the subjects mean square;  $E_s$  is the trials mean square.)

However, the correlation coefficient is not linearly related to the proportion of explained variance, but the square of this coefficient ( $r_i^2$ ) is. Therefore,  $r_i^2$ , known as the coefficient of determination, is the reliability of the procedure. In a complementary fashion,  $1 - r_i^2$  is labelled the coefficient of non-determination. In other words,  $r_i^2$  is the reliability of the procedure based on the estimate of true scores ( $s_A^2$ ) as a ratio of the true scores plus the error of measurement. (Since we can never know the true size parameter, all efforts are estimates.) In the absence of any method error,  $r_i^2$  will be 1.0. If desired, one can test the statistical significance of  $r_i$  against the null hypothesis that  $r_i^2 = 0$  by using the conventional ratio of variances. In this example,  $F = 0.815/0.005 = 163$ , which is highly significant at any reasonable level of alpha.

Buschang and colleagues<sup>66</sup> illustrate another informative approach to estimating reliability using comparisons of complementary ANOVA models<sup>67</sup>. Quite briefly, reliability, as above, is defined as the ratio of explained variance to total variance. In its simplest form, with just one dependent variable, the full model is

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

While the restricted model ignoring the dependent variables is

$$Y = \beta_0 + \epsilon$$

Consider that the one independent variable here is a measure of TEM obtained by repeated observations. If  $\beta_1 X_1$  is a significant amount of the total variance, then the error sum of squares (SSE<sub>F</sub>) for the full model would be smaller than SSE<sub>R</sub> for the restricted model, and we would expect the difference (SSE<sub>R</sub> – SSE<sub>F</sub>) to be positive and large because the

variance due to  $\beta_1 X_1$  has accounted for (and, thus extracted) a discernible portion of the variance from  $SSE_F$ . If, in contrast, the difference ( $SSE_R - SSE_F$ ) is quite small, then the term  $\beta_1 X_1$  is of no particular help in “explaining” the model and can be ignored.

As described by Kirk<sup>67</sup>, this difference  $SSE_R - SSE_F$  can be rewritten (when accounting for the appropriate degrees of freedom) as the ratio of mean squares due to regression of  $Y$  on  $X$  compared to the error (unexplained) mean squares, namely  $MS_R / MS_E$ . This ratio of mean squares can be tested as an F-ratio, and it also can be expressed as a coefficient of determination ( $r^2$ ), which is the proportion of the total variance among the  $Y$  values accounted for (“explained” in the statistical sense) by the independent variable  $\beta_1 X_1$ . This coefficient of determination is termed reliability, and it is the proportion of the overall variance due to true, biological variation. So, if TEM were absent,  $r^2$  would be 1.0, and the higher the  $r^2$  (range of 0 to 1), the smaller the effect of TEM.

In practice (as illustrated by Buschang<sup>66</sup>), several independent variables (such as race, sex, age) would be tested along with the measure of repeatability error to yield a more complete interpretation of the sources of variation in the dependent variable. Likewise, the coefficient of determination attributable to each independent variable would be estimated against the hypothesis that the restricted model without one or more of the independent variables explains just as much of the total variance. Nowadays, mixed model ANOVA designs, with one independent variable being the TEM, can evaluate the sources of variance more efficiently<sup>51</sup>.

## Scaled Measures of Error

How big is the repeatability error? By itself, a value of  $d$ ,  $s_w$ , or so some other measure tells us little about its influence because the error is unrelated to the dimension measured. For example, a mean technical error of 1 mm probably is inconsequential when measuring a person's stature, but a mean error of 1 mm is considerable for a tooth that is a centimetre or less in size. Various measures of relative technical error have been developed (reviewed in Utermohle and Zegura 1982, 1983<sup>11,68</sup>); most of these scale the error by mean size of the measurement.

Looking again at the standard deviation of the TEM ( $d$ ,  $s_w$ ), this value can be recast as  $d/\bar{x}$ , where  $\bar{x}$  is the mean of the variable. This can also be post-multiplied by 100 to provide a percentage of the method error relative to the size of the variable. This is termed relative technical error of the method<sup>39</sup>, which is unit-less. However, as mentioned,  $d$  is the standard deviation of errors, not the mean error as implied by the relative TEM. For the comparison between T1 and T2 (Table 1), relative TEM would be  $0.84 / 8.57 = 0.009$  or about 1% (where 8.57 is the mean of all 40 observations).

Promoters suggest that this yields a more intuitive sense of how large the method error is.

Kieser<sup>12,69</sup> expressed average repeatability differences as a function of the mean and showed that reproducibility of tooth crown dimensions varies considerably among tooth types, evidently because the defining landmarks are more or less accessible and/or well delimited. They report that relative difficulty in obtaining true maximum crown lengths, especially for the buccal teeth, makes MD dimensions more error prone to measurement error than BL breadths.

Sokal and Rohlf<sup>10</sup> suggest that the mean absolute difference of repeated measures is a useful indicator of how variable a measurement technique is, though it (and other scaled measurements) does not have the generalisability of ANOVA methods,

$$\frac{\sum_{i=1}^n |X_{1i} - X_{2i}|}{n}$$

This is more useful than the mean difference since positive and negative results do not average out to a misleadingly small average.

## DISCUSSION

Repeatability errors can be minimized by properly regimented data collection methods, but they probably cannot be eliminated. We suppose that the true trait size is itself invariant, which seems a safe assumption for mineralized tissues observed at one point in time because, for example, they are basically incompressible. On the other hand, Utermohle and Zegura<sup>11</sup> provide a cautionary tail that something as seemingly immutable as a dry skull changes size under varying ambient conditions. Kieser<sup>12,69</sup> are among the few to have quantified TEM for odontometrics, and they reported that TEM introduces a “large and noteworthy error component” into data collection, that TEM tends to increase with time between measurement sessions, and that it may be larger for MD than BL dimensions, perhaps because of greater obstruction when obtaining the MD measurements.

The dual purposes of the present study are (1) to review some methods for quantifying the extent of the method error and (2) to suggest some statistical designs that help control this confounding source of variation. We wholeheartedly agree with Houston’s<sup>32</sup>

perspective that, “Error analysis is tedious and may seem to be unrewarding,” but it needs to be viewed as a necessary step in exploratory data analysis<sup>70</sup>. Repeated measures of the same specimens provide the most informative data on the extent of method error, but it is not a panacea. As an example, Lundström<sup>71</sup> points out one basic shortcoming: What if measurements are repeated accurately but wrongly? If dental crowding prevents correct calliper placement, repeated measurements may be very consistent, but wrong. Analogously, stylistic differences (and handedness) of an observer may account for some of the directional asymmetries recorded for tooth dimensions<sup>72</sup>.

This overview of methods shows that researchers (and disciplines) have pursued two complementary approaches to method error: One is characterized by the intra-class correlation coefficient, where the strength of the association is quantified among measurement sessions, with larger  $r_i$  disclosing smaller random variability caused by method error. The correlation coefficient is dimensionless, so it does not indicate “how” close the repeated observations are. Other researchers opt for measuring method error in actual units; for example, Dahlberg’s  $d$  retains the units of measurement, though the  $sd$ , not the mean  $TEM$ , is obtained. Baumgartner<sup>73</sup> labels these two sorts of repeatability measures as relative and absolute, respectively. As is common with such complementary approaches, the “best” (most informative) solution is to investigate both.

Several researchers suggest that repeatability studies are useful as a preliminary step to the selection of the most reliable variables, namely those with the least method error<sup>11</sup>. This certainly is reasonable, but it contradicts what seems to be a more fundamental issue. Dimensions to be used in a study ought to be chosen because they are informative

vis-à-vis the research question, not simply because they are reproducible. Maximum crown dimensions are a case in point. Mesiodistal and bucco-lingual dimensions entered the research arena simply because they are readily obtainable with callipers, not because of any biological imperative<sup>3</sup>.

### **Landmark Identification**

Studies of TEM in biological settings often report that landmark location is the major source of variability, probably because this step depends the most on human judgment. Much work in this area has focused on cephalometric studies, partly because there are several sources of TEM along the data collection track<sup>74,75</sup> but also because treatment decisions and the evaluation of treatment outcomes can depend on them<sup>32,76,77,78</sup>.

Analogue radiographs are rapidly giving way in dentistry to digital images, which eliminates some sources of error but introduces others. Landmark location persists as an important source of TEM, though computer algorithms for edge detection<sup>79,80</sup> may one day minimize human subjectivity in landmark identification, but variability in image quality will persist. Cephalometric landmarks—and those in several other disciplines (craniometry, anthropometry, odontometrics, *etc.*)—are defined as naturally-occurring maxima and minima of the structures themselves. For example, Menton is the most-inferior (caudal) point on the mandibular symphysis, and Nasion is the dorsal-most point at the intersection of the frontal and nasal bone. These extremes of bony contours are not visibly discrete points; they depend on orientation of the head in each of the three planes of space<sup>81,82</sup>; they remodel with growth; they exhibit different morphologies among individuals; and their locations depend on subjective determinations that are coloured, in

turn, by image quality, bone density, operator experience, landmark definition (theoretically and operationally), and other factors. Much of the same is true of landmark identification in other disciplines<sup>45,69</sup>.

Baumrind and Frantz<sup>83</sup> show that the variability in location and the “envelope of error” (*i.e.*, shape of the distribution of locations) differ among landmarks. Landmarks based on sharper skeleton-dental curves tend to be identified most accurately. Shape of the envelope of error depends on the axis of curvature. For example, visually locating the midpoint on the right edge of this page can be done with considerable accuracy in the horizontal axis (because the edge is straight and runs up-and-down) but will vary much more along the vertical axis. The incisal edge of an incisor (sharp curvature) can be located more accurately on a radiograph than, say, Gnathion (inferior-anterior most point on mandibular symphysis) because of its more gradual curvature.

Notably, variability in landmark identification is ramified when distances (2 landmarks), angles (either 3 or 4 landmarks), or areas (multiple landmarks) are determined because the errors are cumulative<sup>84,85</sup>.

## **Optimal Study**

The methods reviewed here generally involve repeated measures on a sample of the data being analyzed. This is fine so far as it goes, but why not repeat all of the measurements? Lack of time and/or resources is the obvious response, but this is not a compelling reason, especially since repeatability studies<sup>11,12,32,69,83</sup> show consistently and dramatically that method error is considerable. Moreover, the statistical software to control for TEM is

increasingly accessible. So, rather than estimating TEM from a subgroup of the sample, TEM should become an integral part of the analytic model, where it can be quantified and removed from the interesting parts of the analysis. Accounting for TEM involves a mixed model statistical design<sup>5,51</sup>. At its simplest, two or more repeated measurements of each subject are taken and these repetitions are a random effect in the statistical model, while the fixed effect consists of the groups under consideration (such as sex, or genotype, or treatment group). The goal is to remove the effects of TEM from the residual variance, thus enhancing the ratio of explained-to-unexplained mean squares, and it also provides a means of quantifying the TEM as a component of the total variance. Contemporary statistical packages make this relatively easy, but it is hardly a new idea. Gaito and Gifford<sup>86</sup> addressed exactly this issue in 1958, and they provide worked examples of ANOVA models showing how TEM can be separated from the residual term.

Some other efforts are notable in this regard. Palmer<sup>87,88</sup>, and Swaddle<sup>89</sup> describe mixed-model ANOVA where measurement error can be extracted statistically from the study of left-right asymmetry. Van Dongen<sup>61</sup> provide a generalized method of controlling for measurement error while testing for several fixed effects.

## Conclusion

In sum, researchers commonly infer characteristics about populations from comparatively restricted (small) study samples. Most inferences are statistical, and, aside from concerns about adequately accounting for known sources of variation with the research design, an important source of variability is measurement error. Variability in locating landmarks

that define variables is obvious in odontometrics, cephalometrics<sup>83</sup>, and anthropometry<sup>90,91</sup>, but the same concerns about measurement accuracy and precision extend to all disciplines. With increasing accessibility to computer-assisted methods of data collection, the ease of incorporating repeated measures into statistical designs has improved. Accounting for this technical source of variation increases the chance of finding biologically true differences when they exist.

## REFERENCES

01. Bland JM, Altman DG. Statistical notes: measurement error. *BMJ* 1996a; **313**:744.
02. WHO Multicentre Growth Reference Study Group. Reliability of anthropometric measurements in the WHO Multicentre Growth Reference Study. *Acta Paediatrica* 2006; **450**:38-46.
03. Simpson GG, Roe A, Lewontin RC. Quantitative Zoology. New York: Harcourt, Brace and Company, 1960.
04. Bowles FP. Measurement and instrumentation in physical anthropology. *Yrbk Phys Anthropol* 1974 1976; **18**:174-190.
05. Winer BJ, Brown DR, Michels KM. Statistical Principles in Experimental Design, 3rd Edition. New York: McGraw-Hill Book Company, 1991.
06. de Terra M. Beitrage zu einer Odontographie den Menschenrassen. Berlin: Berlinishche Verlagsanstalt, 1905.
07. Wood BA, Abbott SA, Graham SH. Analysis of the dental morphology of Plio-Pleistocene hominids. II. Mandibular molars—study of cusp areas, fissure pattern and cross sectional shape of the crown. *J Anat* 1983; **137**(Pt 2):287-314.
08. Zilberman U, Smith P, Alvesalo L. Crown components of mandibular molar teeth in 45,X females (Turner syndrome). *Arch Oral Biol* 2000; **45**:217-225.
09. Hillson S, Fitzgerald C, Flinn H. Alternative dental measurements: proposals and relationships with other measurements. *Am J Phys Anthropol* 2005; **126**:413-26.
10. Sokal RR, Rohlf FJ. Biometry: The Principles and Practice of Statistics in Biological Research, 3rd Edition. San Francisco: WH Freeman and Company, 1995.
11. Utermohle CJ, Zegura SL, Heathcote GM. Multiple observers, humidity, and choice of precision statistics: factors influencing craniometric data quality. *Am J Phys Anthropol* 1983; **61**:85-95.
12. Kieser JA, Groeneveld HT, McKee J, Cameron N. Measurement error in human dental mensuration. *Ann Hum Biol* 1990; **17**:523-528.

13. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psych Bull* 1955; **52**: 281-302.
14. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitraitmultimethod matrix. *Psych Bull* 1959; **56**: 81-105.
  
15. Gorsuch RL. Factor Analysis, 2nd Edition. Hillsdale, NJ: Erlbaum, 1983.
16. Morrison DF. Multivariate Statistical Methods. New York: McGraw-Hill, 1990.
17. Golbeck AL. Evaluating statistical validity of research reports: a guide for managers, planners, and researchers. General Technical Report PSW-87. Berkeley: U.S. Department of Agriculture, Pacific Southwest Forest and Range Experimental Station, 1986.
18. Bryant TN. The presentation of statistics. *Pediatr Allergy Immunol* 1998; **9** :108-115.
19. Bryant TN. Presenting graphical information. *Pediatr Allergy Immunol* 1999; **10** :4-13.
20. Lorton L, Rethman MP. Statistics: curse of the writing class. *J Endod* 1990; **16**: 13-18.
21. Jamart J. Statistical tests in medical research. *Acta Oncol* 1992; **31**:723-727.
22. Hart A. Towards better research: a discussion of some common mistakes in statistical analyses. *Complement Ther Med* 2000; **8**: 37-42.
23. Kusuoka H, Hoffman JI. Advice on statistical analysis for circulation research. *Circ Res* 2002; **91**:662-671.
24. Martínez-Sellés M, Prieto L, Herranz I. Frequent mistakes in the statistical inference of biomedical data. *Ital Heart J* 2005; **6**:90-95.
25. Strasak AM, Zaman Q, Pfeiffer KP, Göbel G, Ulmer H. Statistical errors in medical research—a review of common pitfalls. *Swiss Med Wkly* 2007; **137**:44-49.
26. Ellis B. Basic Concepts of Measurement. London: Cambridge University Press, 1966.

27. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd Edition. Hillsdale, NJ: Erlbaum Associates, Inc, 1988.
28. Zar JH. Biostatistical Analysis, 4th Edition. Upper Saddle River, NJ: Prentice Hall, 1998.
29. Sterne JA. Teaching hypothesis tests—time for significant change? Stat Med 2002; **21**:985-994.
30. Dahlberg G. Statistical Methods for Medical and Biological Students. London: George Allen and Unwin, Ltd, 1940.
31. Bland JM, Altman DG. Statistical notes: measurement error proportional to the mean. BMJ 1996c; **313** :106.
32. Houston WJ. The analysis of errors in orthodontic measurements. Am J Orthod 1983; **83**:382-90.
33. Hendee WR, Chaney EL, Rossi RP. Radiologic Physics, Equipment and Quality Control. Chicago: Year Book Medical Publishers, Inc., 1977.
34. Athanasiou AE, Editor. Orthodontic Cephalometry. St Louis: Mosby-Wolfe, 1995.
35. Bergersen EO. Enlargement and distortion in cephalometric radiography: compensation tables for linear measurements. Angle Orthod 1980; **50** :230–244.
36. Dibbets JMH, Nolte K. Regional size differences in four commonly used cephalometric atlases: the Ann Arbor, Cleveland (Bolton), London (UK), and Philadelphia atlases compared. \_Orthod Craniofacial Res 2002; **5**:51–58.
37. Togashi K, Kitaura H, Yonetsu K, Yoshida N, Nakamura T. Three-dimensional cephalometry using helical computer tomography: measurement error caused by head inclination. Angle Orthod 2002; **72**:513-520.
38. Hopkins WG. Measures of reliability in sports medicine and science. Sports Med 2000; **30**:1-15.
39. Perini TA, de Oliveira GL, Ornelia JS, de Oliveira FP. Technical error of measurement in anthropometry. Rev Bras Med Esporte 2005; **11**:86-90.
40. Fleiss JL. Statistical Methods for Rates and Proportions, 2nd Edition. New York: John Wiley & Sons, 1981.

41. Agresti A. An Introduction to Categorical Data Analysis. New York: John Wiley & Sons, Inc., 1996.
42. Kondo S, Townsend GC. Associations between Carabelli trait and cusp areas in human permanent maxillary first molars. *Am J Phys Anthropol* 2006; **129**:196-203.
43. Turner CG II, Nichol CR, Scott GR. Scoring procedures for key morphological traits of the permanent dentition: the Arizona State University dental anthropology system. In: Kelley MA, Larsen CS, Editors. *Advances in Dental Anthropology*. New York: Wiley-Liss, 1991, p 13-31.
44. Cameron N. *The Measurement of Human Growth*. London: Croom Helm Ltd, 1984.
45. Johnston FE, Hamill PVV, Lameshow S. Skinfold thickness of children 6-11 years: United States. *Vital Health Stat Series 11*, No. 120, 1972.
46. Mueller WH, Martorell R. Reliability and accuracy of measurements. In: Lohman TG, Roche AF, Martorell R, Editors. *Anthropometric Standardization: Reference Manual*. Champaign, IL: Human Kinetics Books, 1988, p 83-86.
47. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag, 2000.
48. Singer JD, Willett JB. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford: Oxford University Press, 2003.
49. Lampl M, Editor. *Saltation and stasis in Human Growth and Development: Evidence, Methods and Theory*. London: Smith-Gordon, 1999.
50. Lampl M. Saltation and stasis. In: Cameron N, Editor. *Human Growth and Development*. New York: Academic Press, 2002, p. 253-270.
51. Littell RC, Milliken GA, Stroup WW, Wolginger RD, Schabenberger O. *SAS<sup>®</sup> for Mixed Models*, 2nd Edition. Cary, NC: SAS Institute Inc, 2006.
52. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *The Statistician* 1983; **32** :307-317.
53. Bland JM, Altman DG. Statistics notes: measurement error and correlation coefficients. *BMJ* 1996b; **313** :41-42.

54. Cohen J, Cohen P. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. New York: John Wiley & Sons, 1975.
55. Peck S, Peck H. Crown dimensions and mandibular incisor alignment. *Angle Orthod* 1972; **42**:148-53.
56. Harris EF, Potter RH, Lin J. Secular trend in tooth size in urban Chinese assessed from two-generation family data. *Am J Phys Anthropol* 2001; **115**:312-318.
57. Dahlberg G. Twin Births and Twins from a Hereditary Point of View. Stockholm: University Press, 1926.
58. Solow B. The pattern of craniofacial associations. *Acta Odonol Scand* 1966; **24**:1-174.
59. Knapp TR. Technical error of measurement: a methodological critique. *Am J Phys Anthropol* 1992; **87**:235-236.
60. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1** :307-310.
61. Van Dongen S, Molenberghs G, Matthysen E. The statistical analysis of fluctuating asymmetry: REML estimation of a mixed regression model. *J Evol Biol* 1999; **12**:94102.
62. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* 1995; **346**:1085-1087.
63. Mantha S, Roizen MF, Fleisher LA, Thisted R, Foss J. Comparing methods of clinical measurement: reporting standards for Bland and Altman analysis. *Anesth Analg* 2000; **90**:593-602.
64. Dewitte K, Fierens C, Stöckl D, Thienpont LM. Application of the Bland-Altman plot for interpretation of method-comparison studies: a critical investigation of its practice. *Clin Chem* 2002; **48**:799-801.
65. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psych Bull* 1979; **86**:420-428.
66. Buschang PH, Tanguay R, Demirjian A. Cephalometric reliability—a full ANOVA model for the estimation of true and error variance. *Angle Orthod* 1987; **57**:168-175.

67. Kirk RE. Experimental Design: Procedures for the Behavioral Sciences, 2nd Edition. Monterey, CA: Brooks/Cole Publishing Company, 1982.
68. Utermohle CJ, Zegura SL. Intra-and interobserver error in craniometry: a cautionary tale. *Am J Phys Anthropol* 1982; **57**:303-310.
69. Kieser JA, Groeneveld HT. The reliability of human odontometric data. *J Dent Assoc S Afr* 1991; **46**:267-270.
70. Tukey JW. Exploratory data analysis. Reading, Mass: Addison-Wesley, 1977.
71. Lundström A. Tooth Size and Occlusion in Twins. New York: S. Karger, 1948.
72. Harris EF. Laterality in human odontometrics: analysis of a contemporary American White series. In: Lukacs JR, Editor. Culture, Ecology and Dental Anthropology. Chawri-Bazar, Delhi: Kamla-Raj Enterprises, 1992, p 157-170.
73. Baumgartner TA. Norm-referenced measurement: reliability. In: Safran MJ, Wood TM, Editors. Measurement Concepts in Physical Education and Exercise Science. Champaign, IL: Human Kinetics, 1989, p 45-72.
74. Adams JW. Correction of error in cephalometric roentgenograms. *Angle Orthod* 1940; **10**: 3-13.
75. Björk A, Solow B. Measurements on radiographs. *J Dent Res* 1962; **41** :672-683.
76. Midtgård J, Björk G, Linder-Aronson S. Reproducibility of cephalometric landmarks and errors of measurements of cephalometric cranial distances. *Angle Orthod* 1974; **44**:56-61.
77. Baumrind S, Miller D, Molthen R. The reliability of head film measurements. 3. Tracing superimposition. *Am J Orthod* 1976; **70** :617-644.
78. Broch J, Slagsvold O, Røsler M. Error in landmark identification in lateral radiographic headplates. *Eur J Orthod* 1981; **3**:9-13.
79. Liu JK, Chen YT, Cheng KS. Accuracy of computerized automatic identification of cephalometric landmarks. *Am J Orthod Dentofacial Orthop* 2000; **118**:535-540.

80. Kazandjian S, Kiliaridis S, Mavropoulos A. Validity and reliability of a new edge-based computerized method for identification of cephalometric landmarks. *Angle Orthod* 2006; **76**:619-624.
81. Cooke MS, Wei SH. Cephalometric errors: a comparison between repeat measurements and retaken radiographs. *Aust Dent J* 1991; **36**:38-43.
82. Mori Y, Miyajima T, Minami K, Sakuda M. An accurate three-dimensional cephalometric system: a solution for the correction of cephalic malpositioning. *J Orthod* 2001; **28**:143-149.
83. Baumrind S, Frantz RC. The reliability of head film measurements. 1. Landmark identification. *Am J Orthod* 1971a; **60**:111-127.
84. Baumrind S, Frantz RC. The reliability of head film measurements. 2. Conventional angular and linear measures. *Am J Orthod* 1971b; **60**:505-517.
85. Kamoen A, Dermaut L, Verbeeck R. The clinical significance of error measurement in the interpretation of treatment results. *Eur J Orthod* 2001; **23**:569-578.
86. Gaito J, Gifford EC. Components of variance in anthropometry. *Hum Biol* 1958; **30**: 120-127.
87. Palmer AR. Fluctuating asymmetry analyses: a primer. In: Markow TA, Editor. *Developmental Instability: Its Origins and Evolutionary Implications*. Dordrecht: Kluwer Academic Publishers, 1994, p 335-364.
88. Palmer AR, Strobeck C. Fluctuating asymmetry analyses revisited. In: Polak M, Editor. *Developmental Instability: Causes and Consequences*. Oxford: Oxford University Press, 2003, p 279-319.
89. Swaddle JP, Witter MS, Cuthill IC. The analysis of fluctuating asymmetry. *Anim Behav* 1994; **48**:986-989.
90. Spielman RS, Da Rocha FJ, Weitkamp LR, Ward RH, Neel JV, Chagnon NA. The genetic structure of a tribal population, the Yanomama Indians. VII. Anthropometric differences among Yanomama villages. *Am J Phys Anthropol* 1972; **37**:345-356.
91. Moss JP. The use of three-dimensional imaging in orthodontics. *Eur J Orthod* 2006; **28**:416-425.

## FIGURE LEGENDS

Fig. 1. A metaphor of a “bull’s eye” characterizes the concepts of precision and accuracy.

(A) The mean of the measurements is close to the center of the bull’s eye, which is the true value. These measurements have low repeatability, though, because of their scatter and individual departures from the true value. (B) The measurements are close together (good precision), but all are about equally biased from the true value. For example, calipers might be out of kilter, so all measurements are exaggerated by, say, 0.1 mm. (C) Here the measurements are all close to the measurement (high accuracy) and close to one another (high precision).

Fig. 2. Plot of 310 double determinations of the left maxillary central incisors in a sample of American whites (Harris *unpubl.*). Measurements were made independently on two separate occasions several years apart using different calipers. The “checkerboard” appearance of the dots in the dense ellipse along the main diagonal occurs because measurements were truncated to 0.1 mm, and many cases are superimposed.

Fig. 3. The same data used in Figure 2 are plotted, but here 1 inch (2.54 cm) has been added to all of the second-determinations (Y-axis) to illustrate (A) that a systematic bias does not affect the correlation but, of course, (B) a TEM of 1 inch difference is hardly acceptable.

Fig. 4. Example of a Bland-Altman graph where mean size  $(X_{1i} + X_{2i})/2$  on the X-axis is plotted again that pair of measurements difference  $X_{1i} - X_{2i}$ . If, as assumed, trait size is independent of measurement accuracy, the array of dots will be randomly arrayed and centered vertically on zero. In these contrived data, greater inconsistencies occur at the smaller trait sizes (the least-squares regression line is shown). If these data were real, one interpretation might be that smaller specimens are harder to measure accurately. In this example, the regression coefficient ( $b = -$

0.08) is significantly different from zero ( $P = 0.0001$ ), which confirms the visual perception that the error is inversely proportional to the mean.

Table 1. Three hypothetical sets of measurements of the mesiodistal widths of 20 maxillary central incisors.<sup>1</sup>

<b>Case</b>	<b>T1</b>	<b>T2</b>	<b>T3</b>	<b>mean</b>	<b>sd</b>
A	7.9	7.9	7.8	7.9	0.06
B	7.9	8.1	8.0	8.0	0.10
C	8.1	7.9	7.9	8.0	0.12
D	8.3	8.3	8.4	8.3	0.06
E	8.3	8.5	8.4	8.4	0.10
F	8.4	8.5	8.5	8.5	0.06
G	8.5	8.5	8.4	8.5	0.06
H	8.5	8.4	8.4	8.4	0.06
I	8.5	8.4	8.5	8.5	0.06
J	8.6	8.5	8.6	8.6	0.06
K	8.6	8.4	8.6	8.5	0.12
L	8.6	8.4	8.5	8.5	0.10
M	8.6	8.7	8.7	8.7	0.06
N	8.7	8.7	8.7	8.7	0.00
O	8.7	8.6	8.7	8.7	0.06
P	8.7	8.7	8.6	8.7	0.06
Q	8.7	8.7	8.8	8.7	0.06
R	8.7	8.6	8.7	8.7	0.06
S	9.0	8.9	9.0	9.0	0.06
T	10.5	10.5	10.4	10.5	0.06

<sup>1</sup>The 20 teeth are coded A through T, while the 3 sets of repeated measurements are T1, T2, and T3. The standard deviation of each row is in the column labeled sd.

Table 2. ANOVA results analyzing repeatability for the first two data columns in Table 1.<sup>1</sup>

Source	df	SSQ	MS
Subjects	19	10.395	0.547
Trials	20	0.140	0.007

<sup>1</sup>In this model, 20 teeth (subjects) were each measured twice (trials 1 and 2); trials is nested within subjects. Abbreviations are degrees of freedom (df), sum of squares (SSQ), and mean square (MS).

Table 3. ANOVA results analyzing repeatability for the three measurements sessions in Table 1.<sup>1</sup>

Source	df	SSQ	MS	Expected MS
Specimens	19	15.481	0.815	$\sigma^2 + n\sigma_A^2$
Trials	40	0.207	0.005	$\sigma^2$

<sup>1</sup> 20 teeth (subjects) were each measured three times; trials is nested within subjects. Expected mean squares are denoted in the right-hand column, where n is the number of specimens.







