Acta Zoologica Academiae Scientiarum Hungaricae 44 (1–2), pp. 73–96, 1998

MEASUREMENT ERROR IN GEOMETRIC MORPHOMETRICS: EMPIRICAL STRATEGIES TO ASSESS AND REDUCE ITS IMPACT ON MEASURES OF SHAPE*

ARNQVIST, G.** and T. MÅRTENSSON

Department of Animal Ecology, University of Umeå
S-901 87 Umeå Sweden
E-mail: Goran.Arnqvist@animecol.umu.se

Random measurement error is ubiquitous in morphometric data, and it can cause serious statistical problems. We stress that measurement error is a potential problem primarily when true phenotypic variation in shape is relatively small, such as in studies of intraspecific variation in shape. A model for the partitioning of measurement error in landmark based morphometrics is presented. The impact of measurement error can be reduced in a number of ways, depending on the methods used to collect, process and analyse data, and we give some practical advice. We also recommend that repeated measures of all individuals are taken routinely in morphometric studies where measurement error may be a potential problem. This enables both a quantification, by estimating repeatabilities from analyses of variance, and a reduction, by averaging repeated measures, of the relative impact of measurement error. We perform an analysis of shape variation in a uniform sample of young perch (*Perca fluviatilis*), solely aimed at illustrating how different components of measurement error can be quantified, and demonstrate (a) that estimates of repeatability will only be informative of the error components that are actually repeated in each repeated measure, (b) that the relative impact of different components of measurement error can be partitioned and assessed by planned hierarchical repeated measurement protocols followed by nested analyses of variance, (c) that measurement error is unevenly distributed among different shape variables and (d) that the relative magnitude of ME in a given shape variable can be reduced to an estimable extent by averaging several repeated measures.

Key words: shape analysis, repeatability, repeated measures, measurement error, *Perca fluviatilis*

INTRODUCTION

Whenever a value is assigned to a physical quantity, it is associated with uncertainty to some extent; an initially unknown magnitude of measurement error (henceforth, ME) is ubiquitous (FULLER 1987, RABINOVICH 1995). Recent developments in the methodology of data acquisition in morphometrics has led to an increasing level of sophistication of our technical equipment (FINK 1990, MACLEOD 1990, MARCUS *et al.* 1996). However, every biometric measure, no mat-

* Symposium presentation, 5th International Congress of Systematic and Evolutionary Biology, 1996, Budapest

** Corresponding author

ter how sophisticated and/or accurate our methods are, will be associated with ME. It is thus important to understand the effects that ME has on our data and our analyses. In order to do so, we need a framework to classify different sources of potential ME as well as strategies to "cope" with such error. These are the main objectives of the current contribution.

Measurement error is defined as the deviation of the result of a measurement from the true value of the measured quantity. There are two main types of ME; random and systematic (RABINOVICH 1995). The effect of random ME is an increase in the variance of the estimated parameters, by introducing error which is randomly distributed with regards to the true value of the measured quantity. Systematic ME, in contrast, will cause systematic bias in the magnitude of the estimate, by introducing error which is non-randomly (i. e., systematically) distributed with regards to the true value of the measured quantity. Most statistical models of parameter estimation and statistical inference incorporate ME (at least in the dependent variable) that is purely random, and many of the methods used to evaluate geometric morphometric data assume isotropic normal errors in landmark locations (GOODALL & MARDIA 1993, MARDIA & DRYDEN 1994, BOOKSTEIN 1996*a*).

Systematic and random ME entail quite different and distinct problems for the empirical investigators. First, the data that we are analysing for a given specimen will only be an approximation of the true data. This will be a potential problem if systematic ME is present, since systematic deviations from the true values essentially mean that we then analysing shapes that are systematically false approximations of the true underlying shape (see for example Fig. 2). While there are sometimes methods for quantifying and compensating for systematic ME (RABINOVICH 1995), these are often case specific and complex. Second, random variations in the magnitude of error across specimens generates problems when statistical inferences are made (see below). Since systematic ME will occur only if the particular methods or instruments used for gathering data are flawed, while random ME is ubiquitous and often a potential problem, we focus strictly on the more general problem of random ME here. It is important to remember, however, that presence of systematic ME will violated the assumption of isotropic normal errors and may lead to deviations from multivariate normality in the tangent space (BOOKSTEIN 1991 1996*a*, GOODALL & MARDIA 1993, MARDIA & DRYDEN 1994, KENT & MARDIA 1997).

WHY, AND WHEN, IS MEASUREMENT ERROR A PROBLEM?

In a world of error-free measurements, all our data would be true. Unfortunately, our data are always contaminated with ME, and there is no reason to be-

lieve *a priori* that geometric morphometrics generally suffer less from problems with ME than does traditional morphometrics. Two classes of problems can arise from ME. The first, and most general, problem with ME is that it increases our measure of total phenotypic variance in shape, by introducing a component of residual "noise" to our data. Since statistical models are typically based on the relationship between "explainable" and "residual" variation in morphology (ME is a component of the latter), ME dilutes trends and patterns in data by increasing residual variance. Hence, while ME in morphometric data does not necessarily violate any assumptions of statistical models, it is a potentially very serious problem for the practitioner since it reduces the statistical power of our analyses by increasing the type II error rate (i. e., our inability to reject false null hypotheses) (FRANCIS & MATLIN 1986, COHEN 1988, BAILEY & BYRNES 1990, LEE 1990, YEZERINAC *et al.* 1992). The degree to which investigators need to worry about ME in this respect, will be inversely related to the relative magnitude of true between-individual variance in shape in the sample. Thus, comparative studies of shape variation in different species (or higher order taxa) or ontogenetic stages rarely suffer from serious problems with ME. In contrast, in studies where more subtle morphological variation in shape within species, or even populations, is being analysed, ME can be a serious problem. Since the development of the morphometric synthesis now has reached a mature stage (BOOKSTEIN 1996*a*, MARCUS *et al.* 1996) and since ecological morphology is currently receiving an increasing amount of attention (WAINWRIGHT & REILLY 1994), we anticipate that the amount of applications of the latter kind will increase. Thus, it is of great importance that we acknowledge ME, and create a framework for coping with problems caused by it (HIMES 1989, MARKS *et al.* 1989, BAILEY and BYRNES 1990, LOUGHEED *et al.* 1991, YEZERINAC *et al.* 1992).

Secondly, statistical models commonly make two critical assumptions with regards to ME, that may be violated more often than we would wish (FLEISS & SHROUT 1977). (a) It is generally assumed that the magnitude of a variable of interest and the error with which is measured are uncorrelated. The most classical violation of this assumption includes the commonly observed covariance between the true size of a trait and ME (PANKAKOSKI 1987, YEZERINAC *et al.* 1992). This kind of association has been shown, e. g., to generate artifactual relationships between within- and among-population variation across traits (see ROHLF *et al.* 1983) and to generate apparent patterns of decreasing variance in traits during ontogeny as individuals grow larger (see LEE 1990). (b) It is also generally assumed that, when more than one variable is measured, the errors across variables are uncorrelated. Violations of this assumption may be common when there are systematic components of ME involved, for example when more than one person is involved in recording data (LEE 1990, YEZERINAC *et al.* 1992, EASON *et al.* 1996). When the errors across variables are correlated, it will gener-

ally be impossible to determine the extent to which an observed relationship between two variables reflects associations between the true scores and the extent to which it reflects associations between the errors (FLEISS & SHROUT 1977). Violation of this assumption has been shown to generate serious problems in multivariate procedures, such as discriminant analysis (JAMISON & ZEGURA 1974, FRANCIS & MATTLIN 1986, LEE 1982, 1990).

THE ORIGIN OF MEASUREMENT ERROR IN LANDMARK DATA – IDENTIFYING AND CLASSIFYING SOURCES OF ERROR

In order to develop strategies to deal with ME, we first need to identify its sources. A general partitioning of the components of total ME (ζ) is

$$\zeta = \zeta_m + \zeta_i + \zeta_p \qquad [1]$$

where ζ_m is methodological error, ζ_i is instrumental error and ζ_p is personal error (RABINOVICH 1995). A variety of methods and instruments can be employed to capture morphometric data, and the sources of error will of course vary with the methods (see below). We focus below primarily on landmark data, i. e. two- or three-dimensional co-ordinates of specific points on a biological specimen, since this is the most common type of data in geometric morphometrics (BOOKSTEIN 1991, 1996*a*). For such data, the primary origin of error is erroneous locations of landmarks, but this error will then secondarily cascade through the subsequent geometric and statistical analyses to result in ME in shape variables.

The accuracy and precision of recorded data will in theory depend partly on which methods are being used to gather landmark data. In order to recognise this, we first need to identify the sources of error in greater detail than the general partitioning above. Such a systematic functional partitioning of principal sources of ME in landmark data is proposed in Fig. 1, and will be commented on below.

1. Virtually all morphometric data are gathered from specimens that are prepared in one way or the other. What we call error due to specimen preparation occurs as result of slight variations in the way different specimens are prepared for data gathering. Examples of "preparations" include preservation of specimens, skeletonization, histological preparations, dyeing techniques and various forms of presentation/mounting strategies. The latter includes distortions in the "state" in which specimens are represented (e. g., the degree of compression, expansion or flexure), which may be a big source of error in specimens with soft tissue (LEE 1982, CARPENTER 1996).

2. Structures which are truly three dimensional are often reduced to two dimensions in landmark based studies, by two dimensional views of, or cuts through, specimens. In such cases, dimensionality reduction error will arise as a

result of lack of perfect orthogonality between the major axes of the specimen (x,y) to that of the dimension which is being reduced (z), i. e. any variation across specimens in the orientation (alignment) when data is captured. A second, and independent, form of dimensionality reduction error (which is systematic and will

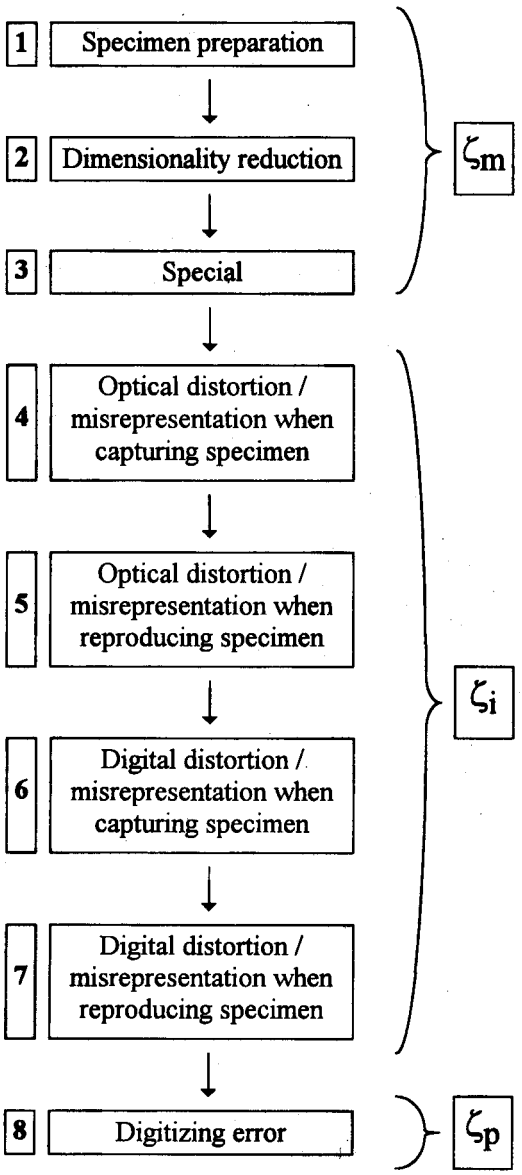


Fig. 1. A sequential partitioning of the main components of measurement error in landmark based morphometrics. ζ_m denotes methodological error, ζ_i instrumental error and ζ_p personal error. See text for descriptions of each of the components

thus not be treated further here) always occurs when whole specimens are viewed through a camera or microscope, by registration of false x,y co-ordinates of landmarks due to differences in their true location in the reduced z dimension (Fig. 2) (see BOOKSTEIN 1991, ROTH 1993, for discussions of dimensionality reduction). Dimensionality reduction error does not occur when three dimensional landmark points are collected directly (ROTH 1993, DEAN 1996).

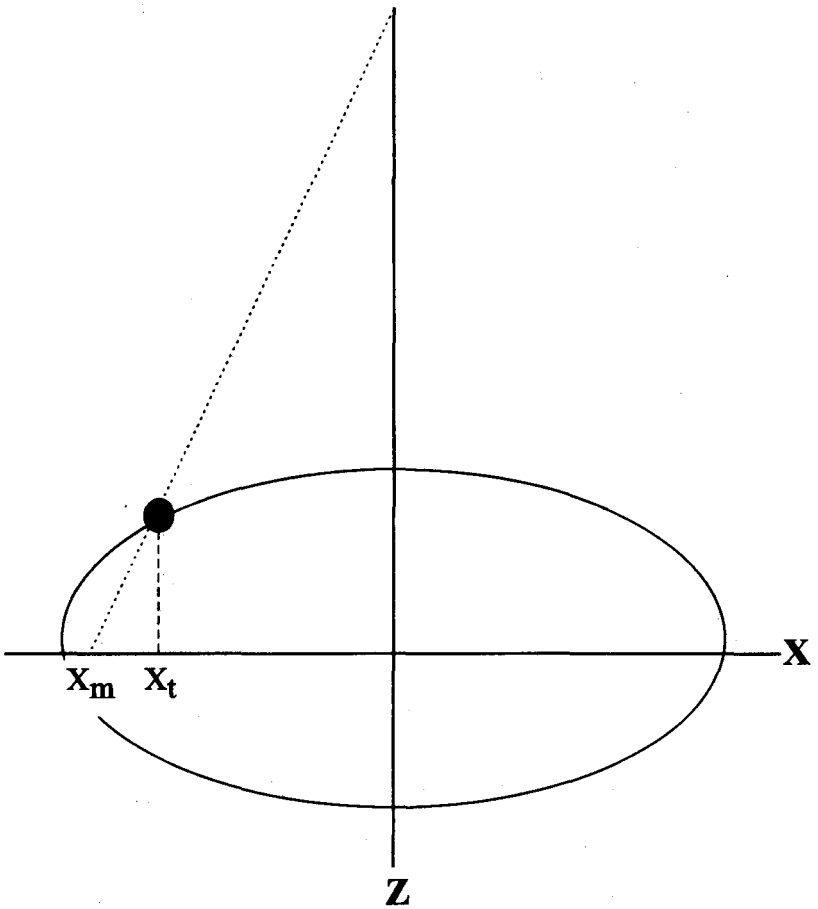


Fig. 2. An illustration of a systematic measurement error in landmark data. This particular type of error arises as a result of dimensionality reduction, when three-dimensional objects are viewed from a single point (camera or microscope). In the illustrated case, a landmark situated on the surface of an oval shaped specimen is being measured. The error equals the difference between the landmark's true and measured location in the x-dimension ($\vartheta = x_t - x_m$). This type of error in landmark location will be systematic, and will occur whenever the true three dimensional location of a landmark is not in the two dimensional plane in which it is projected, unless it is located exactly along the z-axis (i. e., whenever $z \neq 0$; provided that $x,y \neq 0,0$)

3. When special methodological routines are employed to attain or prepare data prior to analysis, this may introduce special methodological error. An example of such a routine, which may introduce considerable error, is when parts of specimens are recorded separately and then merged into complete landmark configurations prior to analysis (see REIG 1998).

4. Landmark data are typically acquired from images of specimens that have been captured in some way. What we call error due to optical distortion/misrepresentation when capturing specimens occurs as a result of imperfections of the optical system (lenses) used when an image of the specimen is being captured and transformed into an analysable form. Examples include distortions due to the characteristics of the lenses in microscope and camera systems (when specimens are viewed directly, or are being transferred to a photographic film or a plane in a video camera).

5. Images of specimens are often being reproduced prior to data acquisition. Error due to optical distortion/misrepresentation when reproducing specimens occurs when already captured images of specimens are being reproduced, and result from imperfections in the optical systems used (photo enlargers and slide projectors). Examples include enlargements of images when transferred from small photographic images (i. e. negatives or colour slides) to larger ones (photographic prints or slide projections) prior to data recording.

6. When a specimen is captured and transformed into digital form, error due to digital distortion/misrepresentation will occur. These types of errors are often complex, resulting from a series of potential imperfections in the digital system used, and they can be considerable (ARNQVIST, unpubl). Examples includes various properties of digital framegrabbers and scanners, but also the more problematic and general sources of error due to image distortion resulting from non-square pixels and the various algorithms used to compensate for this (see MACLEOD 1990). In general, digitising tablets suffer less from this component or error compared to other digital techniques (FINK 1990).

7. To acquire landmarks from digital images, they are reproduced on video monitors or computer screens. Digital distortion/misrepresentation when reproducing specimens occurs as a result of imperfections in the system employed in reproducing the digital data to a visual form (MACLEOD 1990), but also from various manipulations aimed at "enhancing" the reproduced images whenever such software options are used (see ROHLF 1990).

8. When the relative position of landmarks are localised and their co-ordinates are recorded, digitising error will occur as a result of imperfect and/or inconsistent localisation of landmarks. The magnitude of this personal component of error will vary greatly depending on, for example, the type of landmark (FRANCIS & MATLIN 1986, BOOKSTEIN 1991, LOUGHEED *et al.* 1991, DEMETER *et al.* 1996, REIG 1998), the accuracy of the equipment (DEAN 1996), the

quality/resolution of the image/projection and the personal characteristics of the investigator. The sequential and additive input of different components of ME can be illustrated by the following example. Numbers refers to components of error. An investigator wishes to assess potential shape differences due to habitat occupation in a fish species. Samples of fish are taken in the pelagic and benthic zones of a local lake. First, each fish is placed on a flat board in a certain position (1). A picture of each fish is taken (4) from a right angle to the board with a camera mounted on a tripod (2). Photographic prints are produced from the negatives (5), and these are then transformed into digital format by means of a digital scanner (6). Each picture is then reproduced and "enhanced" on a computer screen (7), and landmarks are entered with a pointing device (8). In this case, total ME will equal the sum of all seven components of error.

MINIMISING THE IMPACT OF MEASUREMENT ERROR

Due to its negative effects on statistical evaluations of morphometric data, the impact of ME should, of course, be minimised. There are several things that should be kept in mind in order to do so.

(A) The route between the actual specimen itself and the data should in principle be made as "short" as possible. Landmark data can be gathered in several ways (see ROHLF & BOOKSTEIN 1990, MARCUS *et al.* 1996), the most common being: (a) landmarks located on the specimen itself, by placing the specimen directly on a digitising tablet, (b) landmarks located on a projection of the specimen, by projecting the specimen onto a digitising tablet through a camera lucida, (c) landmarks located on a projection of the specimen, by placing/projecting a photographic reproduction of the specimen (print or slide) onto a digitising tablet, (d) landmarks located on a digital image of the specimen, by capturing a digital representation of the specimen itself through a video-camera or other digital instrument and projecting this onto a video screen, (e) landmarks located on a digital image of the specimen, by capturing a digital representation of a photographic reproduction of the specimen through a video-camera, scanner or other digital instrument and projecting this onto a computer screen. Since the total amount of ME in data is related to the number of methodological steps involved (Fig. 1), steps should be avoided if possible. One important conclusion of this is that there might actually be a positive relationship between the technical sophistication of the methods of data collection and the relative amount of ME present! For example, given that everything else is equal, method (a) above is strongly preferable over method (e).

(B) Methods of specimen preparation, presentation and digitising should be very thoroughly standardised across specimens. Error due to specimen prepara-

tion is often a major source of ME (LEE 1982, CARPENTER 1996, and example below), so the importance of standardisation in this stage cannot be overestimated. Important implications are also that one and the same person should do all the specimen preparations as well as the actual digitising, and that the same equipment and instruments should be used throughout the data collection process within single studies (cf., LEE 1990, YEZERINAC *et al.* 1992, EASON *et al.* 1996).

(C) It has been suggested that certain instruments are more prone to error than others. For example, FINK (1990) recommended that optical slide projectors not be used to project slides of specimens, since these in general have lower optical quality than do other optical instruments. Similarly, digitising tablets in general suffer less from digital distortion/misrepresentation and have higher precision than do visual digital systems (i. e., image analysis systems) (MACLEOD 1990). Thus, the choice of methodology will affect the amount of error present. The instrumental components of measurement error can be assessed, and sometimes partly compensated for, by careful calibration of the equipment. This can be done by, for example, measuring a distance of precisely known length in varied positions and locations (MACLEOD 1990, BECERRA 1993, BECERRA *et al.* 1996, DEMETER *et al.* 1996).

(D) The amount of optical and digital distortion/misrepresentation error that burden data to some extent depends on the quality *per se* of the instruments and equipment used. Thus, the instruments and hardware used to gather data should be chosen carefully. This may be especially true for various digital components, such as different framegrabbers, scanners and digitising tablets which may vary a lot in quality, precision, accuracy and suitability for quantitative morphometrics (FINK 1990, MACLEOD 1990, ROHLF 1990, BECERRA *et al.* 1993, KOHN *et al.* 1995, DEAN 1996, GARCÍA-VALDECASAS 1996). Apart from problems with distortion/misrepresentation, the accuracy and resolution of both capturing equipment (e. g., digitising tablets) and reproducing equipment (e. g., computer monitors) will affect the amount of error present in data.

(E) Using reduced data sets can sometimes decrease the relative amount of error. The impact of ME in a study will depend on the amount of true variation in shape relative to that due to ME. It may sometimes be desirable to use only a restricted subset of all the landmarks collected for the final analysis and statistical evaluation, even if this potentially compromises the information content of a given data set. Choosing which landmarks to use will, however, often be a tedious task. For example, while landmarks located closely together may potentially give valuable information of small scaled and localisable true shape variation, they will also tend to duplicate information of more large scaled true shape variation and hence be redundant to some extent. Information on the absolute precision of landmark locations (ROTH 1993, REIG 1998) can be of great help in choosing landmarks, as can various superimposition methods (ROHLF & SLICE

1990, SLICE 1996) and multivariate techniques (RUSAKOV 1996) aimed at finding regions of independent variation in data.

(F) The most general, and often a very effective, way to reduce the amount of ME in geometric morphometrics is to take repeated measures of each specimen, and then base subsequent statistical inferences on the average shape scores for each individual (HIMES 1989, BAILEY & BYRNES 1990, RABINOVICH 1995) (see below for a discussion of the number of repeated measures that should be taken for each individual specimen). Access to repeated measures also enables assessments of the absolute and relative magnitudes of measurement error (see below).

ASSESSING THE IMPACT OF MEASUREMENT ERROR: TAKING REPEATED MEASURES

Despite its potentially very serious impact, ME has very rarely been explicitly dealt with in morphometric studies (BAILEY & BYRNES 1990). Since there are now established strategies with which to assess and reduce ME, the general awareness of the problem will hopefully increase. It is not possible to assess the amount of ME in a single measure of a single specimen. In contrast, whenever repeated measures (≥ 2) of each of a series of specimens are available, it is possible to assess the magnitude of ME. When taking repeated measures, one should make a great effort to repeat all, or at least as many as possible, of the methodological steps involved in data gathering (i. e., start from "scratch" with each repeated measurement). The importance of this has not previously been recognised (cf., SLICE 1993*a*, DEMETER *et al.* 1996, but see LEE 1982).

Sometimes this can not be done, as in cases when specimens are irreversibly prepared (e. g., skeletonizations, histological preparations) or when only a single reproduction of each specimen is available. It is important to realise that assessments of ME in such cases will not reflect total ME: it will only embody the components of error (Fig. 1) that are in fact repeated across repeated measurements, and will hence reflect only the minimum amount of error present (see example below).

Depending on circumstances, thus, repeated measures will be informative of different components of error (RABINOVICH 1995). The *reproducibility* of measurements reflects the closeness of results of measurements performed under different conditions, with different methods and with different equipment. The reproducibility indicates the magnitude of both random and systematic ME, and is important when systematic errors are suspected to influence the results (LEE 1990, YEZERINAC *et al.* 1992, DEMETER *et al.* 1996, EASON *et al.* 1996). The *repeatability* of measurements reflects the closeness of results of measurements

performed under identical conditions, with the same methods and with the same equipment. The repeatability is informative primarily of the amount of random error present in data (but see below), and is thus the focus of this section. There are two alternative, though not mutually exclusive, ways of assessing ME.

Absolute measures of measurement error

The most common method used so far to assess ME in geometric morphometrics is to calculate absolute accuracy or precision. This is expressed either as an average distance from the mean, or a measure of dispersion such as the standard deviation, among a set of repeated measures of a linear distance or a particular landmark location (e. g., BECERRA *et al.* 1993, LOY *et al.* 1993, DEAN 1996, DEMETER *et al.* 1996, REIG 1998). These measures (e. g., the root mean square – RMS) are absolute in the sense that they can be expressed in an absolute metric unit (e. g., mm), and can be very valuable and appropriate (a) when different data acquisition techniques, or equipment, are compared or (b) when selecting which subset of landmarks, or distances, to include in morphometric analysis (BAILEY & BYRNES 1990, ROTH 1993, REIG 1998, but see KOHN *et al.* 1995).

As mentioned above, the statistical problems generated by ME in empirical morphometrics are related to the amount of true variance in the variables of interest. Because of this, absolute measures of ME are often inadequate to asses the impact of ME in a given study, and investigators are typically left to subjectively deem absolute ME as being either "problematic" or "negligible" (e. g., LOY *et al.* 1993). Most would intuitively agree, for example, that an instrument with an accuracy of 0. 5 mm root mean square (RMS) would be perfectly relevant for a comparative study of cranial shape among primates, and equally inadequate for a study of intraspecific variation of cranial shape in a small rodent species (DEAN 1996). But would this instrument be adequate for a comparison of the cranial shape of two closely related species of eagles? This problem is overcome by the use of relative measures of ME, since they provide quantitative measures of the relative magnitude of ME in shape variables for specific data sets.

Relative measures of measurement error

Several authors have stressed that a meaningful measure of the impact of ME in morphometric studies must relate the amount of ME in a variable to true variation among individuals in the sample (HAGGARD 1958, FLEISS & SHROUT 1977, PALMER & STROBECK 1986, SCHLUTER & SMITH 1986, HIMES 1989, MARKS *et al.* 1989, BAILEY & BYRNES 1990, LEE 1990, LOUGHEED *et al.* 1991, YEZERINAC *et al.* 1992, KOHN *et al.* 1995). This is done by simply performing a model II one-way analysis of variance on repeated measures from each of a series of individuals, with individual as a categorical factor. From such an ana-

lysis, we attain a ratio R of the variance due to differences among individuals to the total variance:

$$R = S^2_A / (S^2_W + S^2_A) \quad [2]$$

where S^2_A is the among-individuals variance component and S^2_W is the within-individuals variance component. Variance components are calculated from the analysis of variance table as:

$$S^2_W = MS_{within} \quad [3]$$

and

$$S^2_A = (MS_{among} - MS_{within}) / n \quad [4]$$

where n is the number of repeated measures per individual (≥ 2) (see LESSELS & BOAG 1987, for cases where the number of repeated measures varies between individuals). This variance ratio has formally been given the somewhat misleading term "intraclass correlation coefficient", since it equals the Pearson correlation coefficient in the simplest case where only two repeated measures have been taken (SOKAL & ROHLF 1995). It has a long history as a measure of ME in the anthropometric literature (e. g., HAGGARD 1958, LORD & NOVICK 1968, FLEISS & SHROUT 1977) as well as in quantitative genetics (FALCONER 1960, BECKER 1984, FALCONER & MACKAY 1996), and has more recently been "rediscovered" as a measure of the impact of ME in morphometrics (BAILEY & BYRNES 1990, LOUGHEED *et al.* 1991, YEZERINAC *et al.* 1992). The variance ratio has been termed reliability in the first, repeatability in the second, and %ME in the third discipline. We suggest that the term repeatability is maintained for the use of R in morphometrics, to agree with its frequent use in other domains of evolutionary biology.

The most attractive characteristic of the repeatability is that it, in contrast to absolute measures, directly relates the magnitude of ME in a particular variable to the magnitude of true morphological variation. Repeatability parameterises the proportion of variance due to true variation between individuals, and ranges between 0 and 1; in the former case all variance is attributable to variance within individuals (i. e., 100% ME), and in the latter all variance is found between individuals (i. e., 0% ME). For the empiricist interested in statistical inferences, it is thus a direct and adequate measure of the relative impact of ME for a given variable in a specific sample.

As mentioned above, the relative magnitude of ME in a given variable decreases when multiple scores of each individual is averaged. The relationship between number of repeated measures per individual, n, and the repeatability after averaging the n measures, R_n, is

$$R_n = \frac{n R}{1 + (n - 1) R} \quad [5]$$

where R is the estimated repeatability of single measures (eq. [2] above) (FLEISS & SHROUT 1977, HIMES 1989, ARNOLD 1994, FALCONER & MACKAY 1996). In general, the lower the repeatability is of a variable, the more can be gained by averaging repeated measures (Fig. 3). The number of repeated measures necessary to achieve a desired level of repeatability after averaging is given by

$$n = \frac{R_n(1-R)}{R(1-R_n)} \qquad [6]$$

These relationships are derived from the Spearman-Brown prophesy formula (FLEISS & SHROUT 1977, HIMES 1989), and can be very helpful when deciding how many repeated measures one should take from each individual.

A STRATEGY TO COPE WITH MEASUREMENT ERROR IN GEOMETRIC MORPHOMETRICS

It is important to stress, again, that the problem with ME in empirical studies is a relative one. In some cases, the relative impact of ME can be safely

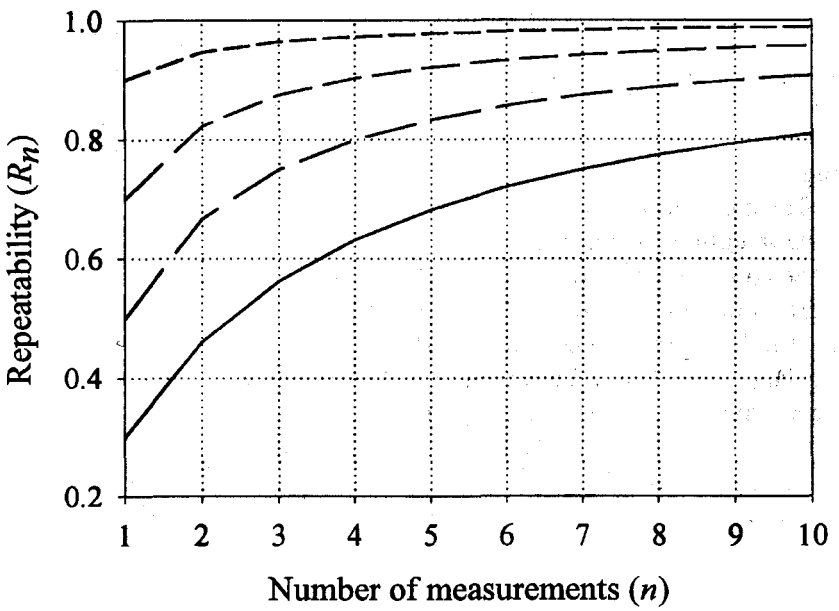


Fig. 3. Estimates of the repeatability that results from averaging several repeated measures (R_n) as a function of the number of repeated measures taken on each individual, for four different values of single measure repeatability (R = 0.3, 0.5, 0.7 and 0.9) (see eq. [5]). This illustrates that the relative proportion of true between-individual variation in a variable can be dramatically increased by averaging several repeated measures

assumed *a priori* to be very low, such as when the method of data acquisition is known to be very accurate and the magnitude of true between-individual variation in shape is known to be very large. However, in cases where this can not be safely assumed, such as in many studies of intraspecific variation, the quality of a study increases significantly when repeated measures are taken (BAILEY & BYRNES 1990). As mentioned above, the benefits of taking repeated measures are two-fold. First, it enables assessments of the relative magnitude of ME in various variables in the sample, by making the estimation of repeatabilities possible (eq. [2]). Second, it increases the statistical power (COHEN 1988), by decreasing the impact of ME to an estimable degree (eq. [5]) when averages of repeated measures for each individual are used for inferential statistical evaluation.

If possible, we recommend that repeated measures are taken routinely for all individuals in the sample, but repeated measures of a subset of all individuals included at least enables estimation of repeatabilities. With regards to the number of repeated measures that should be taken on each individual, we refer to equation [6]. It should be mentioned, however, that it is rarely worth while taking more than four repeated measures (HIMES 1989, FALCONER & MACKAY 1996), unless single measure repeatability is very low (Fig. 3) or one is interested in a further partitioning of ME (see below). If the total number of measurements is logistically constrained, for example if each measure is "costly", a trade-off between the number of individuals measured and the number of repeated measures per individual may occur. BAILEY & BYRNES (1990) gives some valuable guidelines as to how such trade-offs are optimised. In general, increasing the number of individuals measured (if possible), rather than the number of measures per individuals, will often be the preferred strategy.

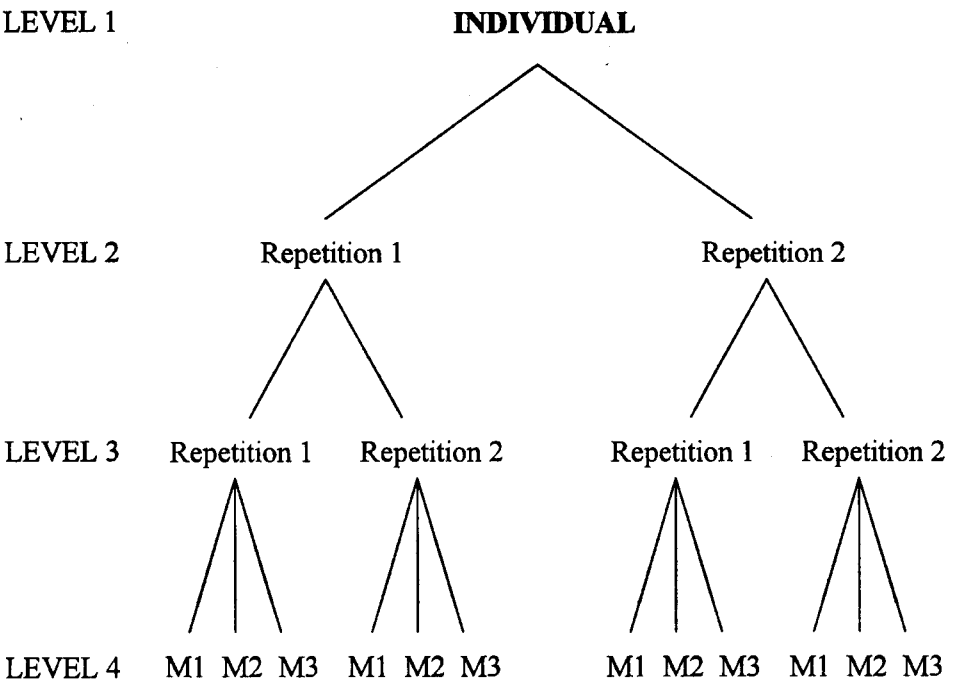
When repeated measures of individuals (multiple repeated landmark configurations per individual) are present one can either (1) use average landmark co-ordinates (e. g., means produced by GLS) for each individual for all further analytical purposes, or one can (2) use all repeated measures of all individuals in the morphometric analysis and then calculate repeatability of the shape variables. In the latter case, subsequent statistical inferences (e. g., various tests of shape differences between groups) can then be based on average shape (e. g., mean partial warp or relative warp score) for each specimen. While the former strategy does increase the quality of the data, we suggest that the latter strategy is most useful for morphometric analyses (see also the example below). The primary advantage of method (2) over (1) is a higher biological and statistical direct relevance of the repeatability estimates. While it may be possible in theory to understand how ME in single landmarks affects the relative magnitude of ME in certain multivariate shape variables (see GOODALL & MARDIA 1993, MARDIA & DRYDEN 1994, DRYDEN *et al.* 1997, KENT & MARDIA 1997, and references therein), the link between the precision or repeatability of a given landmark loca-

tion and the repeatability of multivariate shape variables is typically very obscure for the empiricist. For example, it is not easy to assess and compare the relative impact of ME in a set of multivariate shape variables given that we have information of the precision of the landmark locations on which the analysis is based. In contrast, method (2) generates direct estimates of the relative impact of ME for various components of shape space, by providing repeatability estimates of the multivariate shape variables. This is very important for two reasons. First, statistical inferences in morphometric studies are typically based on shape variables rather than landmark locations *per se*. Second, biological inferences are made by visualising multivariate shape components. We believe that it is critical to assess the impact of ME in a given study by evaluating the variables on which we base our inferences. For example, the relative magnitude of ME in different multivariate shape variables typically vary tremendously, sometimes in non-intuitive ways (LOUGHEED *et al.* 1991, ARNQVIST *et al.* 1997, ARNQVIST & THORNHILL 1998, and below). If the investigator is to make inferences (statistical and biological) of differences in shape between groups of individuals, information on the relative magnitudes of ME in the shape variables in the sample is key. This also allows exclusion of shape dimensions gravely affected by ME from further analysis, and can hence decrease the impact of ME in the study as a whole (BAILEY & BYRNES 1990).

In some cases, one may be interested in partitioning the variance in a given variable beyond the within- and between-individual components discussed so far (cf. eq. [2]). This will be the case, for example, if one wishes to assess the relative magnitude of different components of ME in relation to between-individual variation. This can be made by a strategically planned hierarchical repeated measures protocol, where repetition is done on several different levels that correspond to different sources of error (see Fig. 4). This type of measurement protocol requires a slightly more complex analytical design; it is analysed by extracting variance components from nested analyses of variance (SOKAL & ROHLF 1995), where each level of replication is nested in the ones above (see KOHN *et al.* 1995, and below for examples, and SLICE 1993*a*, for an analogous case). With this design, total ME can be divided into its components.

Finally, one topic that is worth special attention is the measurement of asymmetry, especially fluctuating asymmetry, in bilateral symmetrical traits. This is an area where geometric morphometrics can potentially be a very valuable tool (see AUFFRAY *et al.* 1996, ARNQVIST *et al.* 1997, SMITH *et al.* 1997). In these cases, it is absolutely critical to estimate the relative impact of ME (e. g., repeatability) in ones measures of asymmetry in shape, since ME alone will produce apparent fluctuating asymmetry (see, for example, PALMER & STROBECK 1986, 1997, SWADDLE *et al.* 1994, FIELD *et al.* 1995, HUTCHISON & CHEVERUD 1995, MERILÄ & BJÖRKLUND 1995, BJÖRKLUND & MERILÄ 1997, RABITSCH 1997).

Fig. 4. An example of a hierarchical repeated measures protocol. In this example, each of the levels 2 and 3 are repeated twice and level 4 three times. Each individual is, hence, measured twelve times, and each of the levels 2 – 4 correspond to a certain component of measurement error. For example, level 2 may represent specimen preparation, level 3 specimen capture and reproduction, and level 4 the different measures. Relative measurement error could then be partitioned into components due to methodological, instrumental, and personal (digitising) measurement error, respectively



To discuss the merits of various methods of assessing the relative magnitude of ME in measures of fluctuating asymmetry is beyond the scope of the current presentation, but it seems that a mixed model analysis of variance approach is the preferred method (see PALMER & STROBECK 1986, MERILÄ & BJÖRKLUND 1995, for details).

AN EXAMPLE: QUANTIFYING MEASUREMENT ERROR OF SHAPE IN YOUNG PERCH

To illustrate some of the points made in the current contribution, we here present an analysis of the impact of ME in measures of shape variation in young fish (perch). We wish to stress that the purpose of this analysis is not a biological one, but exclusively confined to methodological issues relating to ME. We illus-

trate four things: 1) that estimates of repeatabilities depend on the error components that are actually repeated for each measure, 2) that the impact of different components of ME can be assessed by hierarchical repeated measurements and analyses, 3) that the relative magnitude of ME in a variable can be reduced by averaging several repeated measures, and 4) that ME is unevenly distributed among different multivariate shape variables.

Materials and methods

For the purposes of this study, a number of young of the year (0+) perch (Perciformes; *Perca fluviatilis*) were collected by means of electrofishing, in a lake situated in northern Sweden (Åmsele, Västerbotten). Twenty of these individuals (3.5–4.5 cm total body length) were fresh frozen, and later thawed and subjected to morphometric analysis. Landmarks were collected by viewing the fish directly in a dissecting microscope (Leica® MZ8), and projecting the image through a camera lucida onto a digitising tablet (Summasketch® III). In total, 21 landmarks were collected for each repeated measure of each fish (see Fig. 5) by means of DS-DIGIT (SLICE 1994). We collected the repeated measures from each individual fish, in the following way. Each fish was positioned (presented) under the microscope, and three repeated measures were entered for the presentation without altering the position of the fish. The fish was then removed from the microscope stage. This whole procedure was repeated three times for each individual. Hence, this protocol yielded nine repeated "landmark maps" for each individual fish, hierarchically repeated as three repeated measures in each of three repeated presentations (3×3) (cf., Fig. 4).

Morphometric analysis

All sets of landmarks (N = 180) for all individuals were translated, scaled and rotated by generalised least-squares Procrustes fit using the GLS option in GRF-ND (SLICE 1993*b*), retaining the centroid size as a measure of size variation in the sample. The uniform components of shape space were then analysed according to BOOKSTEIN (1996*c*). Finally, the non-uniform sub-space of shape was analysed with a thin-plate spline relative warp analysis ($\alpha = 0$), using TPSRW (ROHLF 1993). For each repeated measure, 23 variables were retained for analysis; standardised scores of the first 20 relative warps, the two uniform components and the centroid size (see Table 1). For each variable we performed a nested analysis of variance, from which we extracted variance component estimates (see SOKAL & ROHLF 1995) corresponding to differences (1) among individuals, (2) among presentations within individuals and (3) among repeated measures within presentations and individuals. These will correspond to variance due to (1) true differences between individuals, (2) methodological and instrumental error (varying across presentations) and (3) personal error, respectively. We also calculated explicit repeatabilities from one-way analysis of variance (see eq. [2]

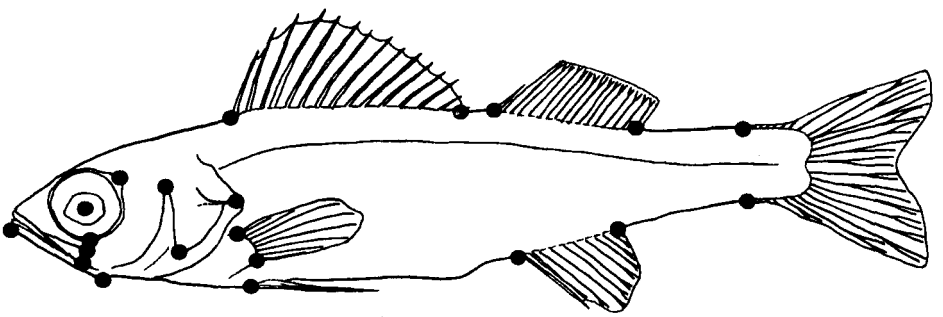


Fig. 5. Locations of landmarks used in the perch data set

above) for all variables in two different ways. First, a subset of only three repeated measures for each individual fish, all from one and the same presentation (the first one), was used to estimate repeatability (R_1) thus including only personal error. Second, all nine repeated measures from all three different presentations for each individual was used to estimate repeatability (R_2), thus including both methodological, instrumental and personal error. In addition, we estimated the repeatability that would result from averaging the nine repeated measures, R_n, as

$$R_n = \frac{9\,R_2}{1 + 8\,R_2} \qquad [7]$$

where R_2 represent the estimate mentioned above (cf., eq. [5]).

RESULTS AND DISCUSSION

We found our measure of size (centroid size) to be considerably less affected by ME than our multivariate measures of shape (Table 1), an expected pattern that should be very general (see also LOUGHEED *et al.* 1991, ARNQVIST & THORNHILL 1998). The proportion of measurement error ranged between approximately 15–30% for the uniform shape components, 10–40% for relative warps #1–10, and 35–80% for relative warps #11–20 (see R_2 in Table 1). One should bear in mind, however, that these figures are inflated relative to most "real" data sets. In this analysis, we were deliberately trying to keep between-individual variation in shape low (we used only a limited number of similar-sized individuals that were collected at the same locality and at the same occasion), in order to elevate the overall relative impact of ME.

As mentioned above, it is sometimes impossible to repeat all the methodological steps in one's repeated measures. Examples of such steps are when skeletal parts have been skeletonised, histological preparations have been made, each specimen is only available as a single photographic/digital reproduction, or when specimens have been preserved in preservatives (LEE 1982, CARPENTER 1996). In all these cases, the specimens are prepared once and for all, and the method is irreversible and hence unrepeatable. In these cases, the repeatabilities will reflect primarily personal error (ζ_p), and depending on circumstances possibly certain components of instrumental error (ζ_i). However, since the methodological error (ζ_m) is often a major component of ME, the estimated repeatabilities will represent highly inflated estimates of the quality of ones data in such samples. Our analyses illustrate this in two ways. First, the repeatabilities of all shape scores were considerably higher when only personal error was accounted for (cf. R_1 versus R_2 in Table 1). Second, variance due to personal error was in average about half that due to methodological and instrumental error (cf. 2nd versus 3rd variance component in Table 1). The major source of methodological error in our case was undoubtedly slight variations in the exact positioning of the fish across

Table 1. Estimates of repeatability for a series of morphometric variables in the perch data set. Given are also scaled variance component estimates for each trait, expressed in percent of the total variance. These three components estimate true phenotypic variance, variance due to methodological and instrumental error ($\zeta_m + \zeta_i$) and variance due to personal error (ζ_p), respectively. Numbers within brackets represent percentage of variance in shape explained by each relative warp. See text for further explanation

Variable		Repeatabilities			Variance components		
		R_1[a]	R_2[b]	R_n[c]	(1)	(2)	(3)
					among: Individuals	Presentations	Measures
					within:	Individuals	Presentations
Centroid size		1.00	1.00	1.00	99.4	0.5	0.1
Uniform comp. 1		0.93	0.77	0.97	71.1	23.5	5.4
Uniform comp. 2		0.97	0.86	0.98	82.4	14.8	2.7
Relative warp 1	(30.6)	0.99	0.86	0.98	81.2	18.3	0.5
Relative warp 2	(16.4)	0.98	0.88	0.98	84.4	12.9	2.6
Relative warp 3	(10.7)	0.97	0.81	0.97	75.5	22.0	2.6
Relative warp 4	(8.1)	0.94	0.77	0.97	70.7	24.5	4.8
Relative warp 5	(6.9)	0.90	0.81	0.97	77.0	15.7	7.3
Relative warp 6	(4.5)	0.93	0.77	0.97	72.3	20.6	7.1
Relative warp 7	(4.0)	0.89	0.82	0.98	79.0	12.3	8.7
Relative warp 8	(2.5)	0.90	0.86	0.98	85.2	5.6	9.2
Relative warp 9	(2.3)	0.87	0.59	0.93	48.9	40.8	10.3
Relative warp 10	(1.9)	0.85	0.63	0.94	54.3	33.3	12.5
Relative warp 11	(1.6)	0.86	0.56	0.92	48.1	34.9	17.0
Relative warp 12	(1.4)	0.80	0.65	0.94	58.3	27.2	14.6
Relative warp 13	(1.2)	0.82	0.66	0.95	62.5	17.0	20.5
Relative warp 14	(0.9)	0.79	0.41	0.86	27.6	50.2	22.2
Relative warp 15	(0.8)	0.61	0.39	0.85	27.0	53.5	19.5
Relative warp 16	(0.7)	0.70	0.19	0.67	5.4	51.8	42.9
Relative warp 17	(0.7)	0.59	0.40	0.86	28.9	41.4	29.6
Relative warp 18	(0.6)	0.58	0.40	0.86	30.6	42.5	26.9
Relative warp 19	(0.5)	0.60	0.22	0.72	11.0	43.2	45.8
Relative warp 20	(0.5)	0.66	0.21	0.70	3.8	62.3	34.0

presentations. Thus, our results illustrates the key importance of repeating all, or at least as many as possible, of the steps involved in the data gathering procedure. If this is not done, the investigator must be aware that repeatabilities only give an upper bound of the true "quality" of the data. This is often unsatisfactory, espe-

cially since the relative impact of different components of error is not necessarily correlated across different shape variables (see below).

Our methodological exercise demonstrates how the relative impact of ME in different shape variables can be partitioned, quantified and understood by a strategically planned hierarchical repeated measures protocol, followed by nested analyses of variance. With our design, the variance component that is due to differences among individuals corresponds to the repeatability of ones shape measures (correlation coefficient between variance component 1 and R_2 in Table 1, $r = 0.998$). In empirical studies, where the purpose is to make biological and statistical inferences, further levels can of course be added to such nested models (see SLICE 1993*a*, KOHN *et al.* 1995), including effect factors. In our case, for example, fish could have been sampled from different lakes, adding a variance component due to differences among lakes (variance among individuals would then be nested within a random lake factor). There is much to be gained by this type of analysis, since the potential for biological insight increases significantly by including ME in our statistical models (information on the relative magnitude of ME in different components of shape are available) rather than disregarding any variance that is due to error.

This type of analysis also allows for statistical inferences to be based on average shape of each individual, which increases the statistical power of ones tests (reduces the type II statistical error rate; COHEN 1988). Our repeatabilities were in many cases dramatically improved when averaging the nine repeated measures, especially for variables with low repeatabilities (cf. R_2 versus R_n in Table 1). The overall relative proportion of ME decreased from almost 40% to less than 10% (average for all 22 shape variables). Basing further statistical analysis (for example, tests of differences between groups) on average shape of each individual would thus greatly improve the quality of the analysis.

Our variance component analysis generated two further insights, with regards to the relative impact of ME in different components of shape space. First, and most importantly, the relative impact of ME generally increased with order among relative warps (Spearman rank correlations [r_s] between relative warp order and variance components 1–3; -0.88, 0.78 and 0.98 respectively, $P < 0.001$ in all cases). This pattern is expected since principal component analysis tend to selectively recover true structure from early axes, hence leaving later axes with relatively larger proportions of ME (GAUCH 1982, LOUGHEED *et al.* 1991). This was obvious in our example, where higher order relative warps were not only severely affected by ME but also explained a very low proportion of variance in shape (Table 1). Thus, there are several reasons for why principal components, in this case relative warps, should be interpreted with an increasing amount of caution with their order. Second, within this general trend, the relative impact of different components of ME was remarkably unevenly distributed among different

shape variables (see also LOUGHEED *et al.* 1991). Detailed information of this kind is especially helpful when interpreting negative findings, since some multivariate shape variables can parameterize shape variation that is to a large extent due to ME (e. g., relative warps 16, 19 and 20 in Table 1; see also CARPENTER 1996). Further, though the relative magnitudes of ME due to personal error on the one hand and methodological and instrumental error on the other were positively correlated (Spearman rank correlation [r_s] between variance components 1 and 2, 0.75, $P < 0.001$), it varied considerably among different shape variables (e. g., compare variance components 2 and 3 for relative warps 1 vs 8 and 9 vs 19).

In conclusion, we hope that our contribution has shown that there are several reasons for increasing our awareness of the impact of ME in geometric morphometrics. We do not in any way dispute the tremendous potential of the tools of the morphometric synthesis (ROHLF & MARCUS 1993, BOOKSTEIN 1996*a, b*, MARCUS *et al.* 1996), and hence do not wish to discourage anyone from using these methods. On the contrary, we hope to have shown that explicitly acknowledging the existence of ME is key in geometric morphometrics, and that including quantifications of the impact of ME such as those demonstrated here will lead to much more powerful and insightful applications.

* * *

Acknowledgements – We thank F. J. BOOKSTEIN, C. KLINGENBERG, S. REIG & P. WATSON for critical comments on previous versions of this paper; J. ROHLF & D. SLICE for providing the excellent software that made this study possible; P. BYSTRÖM & E. WESTMAN for providing the fish; and P. WATSON for originally acquainting us with geometric morphometrics. This study was made possible by financial support from The Swedish Natural Science Research Council.

REFERENCES

ARNOLD, S. J. (1994) Multivariate inheritance and evolution: a review of concepts. Pp. 17–48. *In* BOAKE (ed.): *Quantitative genetic studies of behavioral evolution.* The Univ. Chicago Press, Chicago.

ARNQVIST, G. & THORNHILL, R. (1998) Evolution of animal genitalia: patterns of phenotypic and genotypic variation and condition dependence of genital and non-genital morphology in a water strider. *Genetical Res.* [in press].

ARNQVIST, G., THORNHILL, R. & ROWE, L. (1997) Evolution of animal genitalia: morphological correlates of fitness components in a water strider. *J. Evol. Biol.* **10**: 613–640.

AUFFRAY, J.-C., ALIBERT, P., RENAUD, S., ORTH, A. & BONHOMME, F. (1996) Fluctuating asymmetry in Mus musculus subspecific hybridization: traditional and Procrustes comparative approach. Pp. 275–284. *In* MARCUS *et al.* (eds): *Advances in Morphometrics.* NATO ASI Series A: Life Sciences, Vol. 284, Plenum Press, New York.

BAILEY, R. C. & BYRNES, J. (1990) A new, old method for assessing measurement error in both univariate and multivariate morphometric studies. *Syst. Zool.* **39**: 124–130.

BECERRA, J. M. (1996) Imagina – a direct tool for image analysis in systematics. Pp. 83–89. *In* MARCUS *et al.* (eds): *Advances in Morphometrics.* NATO ASI Series A: Life Sciences, Vol. 284, Plenum Press, New York.

BECERRA, J. M., BELLO, E. & GARCÍA-VALDECASAS, A. (1993) Building your own machine image system for morphometric analysis: a user point of view. Pp. 65–92. *In* MARCUS *et al.* (eds): *Contributions to morphometrics.* Monografías del Museo Nacional de Ciencias Naturales 8, Madrid.

BECKER, W. A. (1984) *A manual of quantitative genetics.* Academic Enterprises, Pullman, Washington, 130 pp.

BJÖRKLUND, M. & MERILÄ, J. (1997) Why some measures of fluctuating asymmetry are so sensitive to measurement error. *Ann. Zool. Fenn.* **34**: 133–137.

BOOKSTEIN, F. L. (1991) *Morphometric tools for landmark data.* Cambridge Univ. Press, Cambridge, New York, 435 pp.

BOOKSTEIN, F. L. (1996*a*) Biometrics, biomathematics and the morphometric synthesis. *Bull. Math. Biol.* **58**: 313–365.

BOOKSTEIN, F. L. (1996*b*) Combining the tools of geometric morphometrics. Pp. 131–152. *In* MARCUS *et al.* (eds): *Advances in Morphometrics.* NATO ASI Series A: Life Sciences Vol. 284, Plenum Press, New York.

BOOKSTEIN, F. L. (1996*c*) Standard formula for the uniform shape component in landmark data. Pp. 153–168. *In* MARCUS *et al.* (eds): *Advances in Morphometrics.* NATO ASI Series A: Life Sciences Vol. 284, Plenum Press, New York.

CARPENTER, K. E. (1996) Morphometric pattern and feeding mode in emperor fishes (Lethrinidae, Perciformes). Pp. 479–488. *In* MARCUS *et al. (eds): Advances in Morphometrics.* NATO ASI Series A: Life Sciences Vol. 284, Plenum Press, New York.

COHEN, J. (1988) *Statistical power analysis for the behavioral sciences.* Lawrence Erlbaum Ass., Hillsdale, 567 pp.

DEAN, D. (1996) Three-dimensional data capture and visualization. Pp. 53–70. *In* MARCUS *et al.* (eds): *Advances in Morphometrics.* NATO ASI Series A: Life Sciences Vol. 284, Plenum Press, New York.

DEMETER, A., VÁMOSI, J., PEREGOVITS, L. & TOPÁL, G. (1996) An image-capture and data-collection system for morphometric analysis. Pp. 91–102. *In* MARCUS *et al.* (eds): *Advances in Morphometrics.* NATO ASI Series A: Life Sciences Vol. 284, Plenum Press, New York.

DRYDEN, I. L., FAGHIHI, M. R. & TAYLOR, C. C. (1997) Procrustes shape analysis of planar point subsets. *J. Roy. Statist. Soc.* **59**: 353–374.

EASON, T. H., SMITH, B. H. & PELTON, M. R. (1996) Researcher variation in collection of morphometrics on black bears. *Wildl. Soc. Bull.* **24**: 485–489.

FALCONER, D. S. (1960) *Introduction to quantitative genetics*, 1st ed., Oliver and Boyd, London, 365 pp.

FALCONER, D. S. & MACKAY, T. F. C. (1996) *Introduction to quantitative genetics*, 4th ed. Longman, Harlow, 464 pp.

FIELD, S. J., SPIERS, M., HERSHKOVITZ, I. & LIVSHITS, G. (1995) Reliability of reliability coefficients in the estimation of asymmetry. *Am. J. Phy. Anthrop.* **96**: 83–87.

FINK, W. (1990) Data aquisition in systematic biology. Pp. 9–20. *In* ROHLF & BOOKSTEIN (eds): *Proc. Michigan Morphometrics Workshop.* Special Publ. 2, Univ. Michigan Museum of Zoology, Ann Arbor. 380 pp.

FLEISS, J. L. & SHROUT, P. E. (1977) The effects of measurement errors on some multivariate procedures. *Am. J. Public Health* **67**: 1188–1191.

FRANCIS, R. I. C. C. & MATLIN R. H. (1986) A possible pitfall in the morphometric application of discriminant analysis: measurement bias. *Marine Biol.* **93**: 311–313.

FULLER, W. A. (1987) *Measurement error models.* John Wiley, New York, 440 pp.

GARCÍA-VALDECASAS, A. (1996) Two-dimensional imaging: an update. Pp. 71–82. *In* MARCUS *et al.* (eds): *Advances in Morphometrics.* NATO ASI Series A: Life Sciences Vol. 284, Plenum Press, New York.

GAUCH, H. G. (1982) Noise reduction by eigenvector ordinations. *Ecology* **63**: 1643–1649.

GOODALL, C. R. & MARDIA, K. V. (1993) Multivariate aspects of shape theory. *Ann. Stat.* **21**: 848–866.

HAGGARD, E. A. (1958) *Intraclass correlation and the analysis of variance.* Dryden Press, New York, 171 pp.

HIMES, J. H. (1989) Reliability of anthropometric methods and replicate measurements. *Am. J. Phy. Anthrop.* **79**: 77–80.

HUTCHISON, D. W. & CHEVERUD, J. M. (1995) Fluctuating asymmetry in tamarin (Saguinus) cranial morphology – intraspecific and interspecific comparisons between taxa with varying levels of genetic heterozygosity. *J. Hered.* **86**: 280–288.

JAMISON, P. L. & ZEGURA, S. L. (1974) A univariate and multivariate examination of measurement error in anthropometry. *Am. J. Phy. Anthrop.* **40**: 197–204.

KENT, J. T. & MARDIA, K. V. (1997) Consistency of Procrustes estimators. *J. R. Statist. Soc.* **59**: 281–290.

KOHN, L. A. P., CHEVERUD, J. M., BHATIA, G., COMMEAN, P., SMITH, K. & VANNIER, M. W. (1995) Anthropometric optical surface imaging system repeatability, precision, and validation. *Ann. Plast. Surg.* **34**: 362–371.

LEE, J. C. (1982) Accuracy and precision in anuran morphometrics: artifacts of preservation. *Syst. Zool.* **31**: 266–281.

LEE, J. C. (1990) Sources of extraneous variation in the study of meristic characters: the effects of size and inter-observer variability. *Syst. Zool.* **39**: 31–39.

LESSELS, C. M. & BOAG, P. T. (1987) Unrepeatable repeatabilities: a common mistake. *Auk* **104**: 116–121.

LORD, F. M. & NOVICK, M. R. (1968) *Statistical theories of mental test scores.* Addison-Wesley, Reading, 568 pp.

LOUGHEED, S. C., ARNOLD, T. W. & BAILEY, R. C. (1991) Measurement error of external and skeletal variables in birds and its effect on principal components. *Auk* **108**: 432–436.

LOY, A., CORTI, M. & MARCUS, L. F. (1993) Landmark data: size and shape analysis in systematics – a case study on old world Talpidae (Mammalia, Insectivora). Pp. 215–240. *In* MARCUS *et al.* (eds): *Contributions to morphometrics.* Monografías del Museo Nacional de Ciencias Naturales 8, Madrid.

MACLEOD, N. (1990) Digital images and automated image analysis systems. Pp. 21–36. *In* ROHLF, F. J. & BOOKSTEIN, F. L. (eds): *Proceedings of the Michigan Morphometrics Workshop.* Special Publ. 2, University of Michigan Museum of Zoology, Ann Arbor.

MARCUS, L. F., CORTI, M., LOY, A., NAYLOR, G. J. P. & SLICE D. E. (1996) *Advances in Morphometrics.* NATO ASI Series A: Life Sciences Vol. 284, Plenum Press, New York, 587 pp.

MARDIA, K. V. & DRYDEN, I. L. (1994) Shape averages and their bias. *Adv. Appl. Prob.* **26**: 334–340.

MARKS, G. C., HABICHT, J.-P. & MUELLER, W. H. (1989) Reliability, dependability, and precision of anthropometric measurements. *Am. J. Epidemiol.* **130**: 578–587.

MERILÄ, J. & BJÖRKLUND, M. (1995) Fluctuating asymmetry and measurement error. *Syst. Biol.* **44**: 97–101.

PALMER, A. R. & STROBECK, C. (1986) Fluctuating asymmetry: measurement, analysis and pattern. *Ann. Rev. Ecol. Syst.* **17**: 391–421.

PALMER, A. R. & STROBECK, C. (1997) Fluctuating asymmetry and developmental stability: heritability of observable variation vs. heritability of inferred cause. *J. Evol. Biol.* **10**: 39–49.

PANKAKOSKI, E., VÄISÄNEN, R. A. & NURMI, K. (1987) Variability of muskrat skulls: measurement error, environmental modification and size allometry. *Syst. Zool.* **36**: 35–51.

RABINOVICH, S. R. (1995) *Measurement errors: theory and practice.* American Institute of Physics, New York, 279 pp.

RABITSCH, W. B. (1997) Levels of asymmetry in Formica pratensis Retz (Hymenoptera, Insecta) from a chronic metal-contaminated site. *Env. Toxicol. Chem.* **16**: 1433–1440.

REIG, S. (1998) 3D digitizing precision and sources of error in the geometric analysis of weasel skulls. *Acta zool. hung.* **44**(1–2): 61–72.

ROHLF, F. J. (1990) An overview of image processing and analysis techniques for morphometrics. Pp. 37–60. *In* ROHLF, F. J. & BOOKSTEIN, F. L. (eds): *Proc. Michigan Morphometrics Workshop.* Special Publ. 2, Univ. Michigan Museum of Zoology, Ann Arbor.

ROHLF, F. J. (1993) *TPSRW: thin-plate spline relative warp analysis.* Dept. Ecology and Evolution, State University of New York, Stony Brook, New York.

ROHLF, F. J. & BOOKSTEIN, F. L. (1990) *Proc. Michigan Morphometrics Workshop.* Special Publ. 2, University of Michigan Museum of Zoology, Ann Arbor, 380 pp.

ROHLF, F. J., GILMARTIN, A. J. & HART, G. (1983) The Kluge-Kerfoot phenomenon – a statistical artefact. *Evolution* **37**: 180–202.

ROHLF, F. J. & MARCUS, L. F. (1993) A revolution in morphometrics. *Trends Ecol. & Evol.* **8**: 129–132.

ROHLF, F. J. & SLICE, D. E. (1990) Extensions of the Procrustes method for the optimal superimposition of landmarks. *Syst. Zool.* **39**: 40–59.

ROTH, V. L. (1993) On three-dimensional morphometrics, and on the identification of landmark points. Pp. 41–61. *In* MARCUS *et al.* (eds): *Contributions to morphometrics.* Monografías del Museo Nacional de Ciencias Naturales 8, Madrid.

RUSAKOV, D. A. (1996) Dimension reduction and selection of landmarks: a Monte Carlo experiment. Pp. 201–208. *In* MARCUS *et al.* (eds): *Advances in Morphometrics.* NATO ASI Series A: Life Sciences Vol. 284, Plenum Press, New York.

SCHLUTER, D. & SMITH, J. N. M. (1986) Genetic and phenotypic correlations in a natural population of song sparrows. *Biol. J. Linn. Soc.* **29**: 23–36.

SLICE, D. E. (1993*a*) The fractal analysis of shape. Pp. 161–190. *In* MARCUS *et al.* (eds): *Contributions to morphometrics.* Monografías del Museo Nacional de Ciencias Naturales 8, Madrid.

SLICE, D. E. (1993*b*) *GRF-ND: generalized rotational fitting of n-dimensional landmark data.* Dept. Ecology and Evolution, State University of New York, Stony Brook, New York.

SLICE, D. E. (1994) *DS-DIGIT: basic digitizing software.* Dept. Ecology and Evolution, State University of New York, Stony Brook, New York.

SLICE, D. E. (1996) Three-dimensional generalized resistant fitting and the comparison of least-squares and resistant-fit residuals. Pp. 179–199. *In* MARCUS *et al.* (eds): *Advances in Morphometrics.* NATO ASI Series A: Life Sciences Vol. 284. Plenum Press, New York.

SMITH, D. R., CRESPI, B. J. & BOOKSTEIN, F. L. (1997) Fluctuating asymmetry in the honey bee, Apis mellifera: effects of ploidy and hybridization. *J. Evol. Biol.* **10**: 551–574.

SOKAL, R. R. & ROHLF, F. J. (1995) *Biometry,* 3rd ed. Freeman & Co., San Francisco, 887 pp.

SWADDLE, J. P., WITTER, M. S. & CUTHILL, I. C. (1994) The analysis of fluctuating asymmetry. *Anim. Behav.* **48**: 986–989.

WAINWRIGHT, P. C. & REILLY, S. M. (1994) *Ecological morphology: integrative organismal biology.* The Univ. of Chicago Press, Chicago, 367 pp.

YEZERINAC, S. M., LOUGHEED, S. C. & HANDFORD, P. (1992) Measurement error and morphometric studies: statistical power and observer experience. *Syst. Biol.* **41**: 471–482.