Ann. Zool. Fennici 47: 46–56 ISSN 0003-455X (print), ISSN 1797-2450 (online)
Helsinki 10 March 2010

Measurement error in morphometric studies: comparison between manual and computerized methods

Francesc Muñoz-Muñoz[1] & David Perpiñán[2]

[1)] *Departament de Biologia Animal, de Biologia Vegetal i d'Ecologia, Universitat Autònoma de Barcelona, 08193-Bellaterra, Barcelona, Spain; and Centre de Recerca en Sanitat Animal (CReSA), UAB-IRTA, Campus de la Universitat Autònoma de Barcelona, 08193-Bellaterra, Barcelona, Spain (corresponding author's e-mail: francesc.munoz@cresa.uab.cat)*
[2)] *Department of Small Animal Medicine and Surgery, College of Veterinary Medicine, University of Georgia, 501 DW Brooks Drive, Athens, GA 30602, USA*

Received 18 Mar. 2008, revised version received 2 July 2009, accepted 2 July 2009

Muñoz-Muñoz, F. & Perpiñán, D. 2010: Measurement error in morphometric studies: comparison between manual and computerized methods. — *Ann. Zool. Fennici* 47: 46–56.

The aim of this study was to compare measurement error (ME) between two different methods of measuring cranial traits: manual method, using calipers; and computerized one, using digitalized pictures and specialized software. Three observers measured 10 craniometric characters in 12 skulls of the common house mouse *Mus musculus domesticus*. Every measurement was repeated three times with each method. Nested ANOVA was used to separate the total variance into within- and among-individual components. Then the effect of trait size on ME was tested. Measuring method was the factor with higher values of ME, followed respectively by observer and replicate. Intra-observer variation was lower than inter-observer variation in both methods. However, repeatabilities were higher in the computerized procedure. Computerized measuring procedure was more precise and less influenced by factors increasing ME than manual method in most assessed traits.

Introduction

Morphometric data are an important source of information to understand many biological phenomena. The use of morphological measurements has been widespread in studies of phylogenetic relationships (Rae 1998, Zelditch *et al*. 2004), evolution (Lieberman 1998), reconstruction of history and structure of past populations (González-José *et al*. 2001), sexual dimorphism (Vincent *et al*. 2004), fluctuating asymmetry (Badyaev *et al*. 2000, Willmore *et al*. 2005), ecomorphology (Klingenberg & Ekau 1996), body condition (Green 2001), growth (Ackermann 2005), heritability (Cheverud 1996, Kruuk *et al*. 2000), life histories (Bonner 1965), animal behavior (Searcy 1979), community structure (Strong 1983) and ecological processes (Harper *et al*. 1970). However, morphometric data are affected by measurement error (ME), which can be defined as the variability of repeated measurements of a particular character taken on the same individual, relative to its variability among individuals in a particular group (Bailey & Byrnes 1990). Statistically, ME can also be expressed as percent measurement error (%ME), which is the amount of the total variance due to the within-individual variation or, in other words,

the percentage of sample variation due to the imprecision of the measure (Bailey & Byrnes 1990). There are many potential causes that can lead to ME, these include:

- Sophistication/precision of the measuring device (Bailey & Byrnes 1990, Yezerinac *et al*. 1992). Measurement error decreases as sophistication increases.
- Experience/skill of researchers in making measurements (Bailey & Byrnes 1990, Yezerinac *et al*. 1992). Measurement error decreases as experience increases.
- Definition of landmarks (Bailey & Byrnes 1990, Yezerinac *et al*. 1992). Landmarks have been classified into different categories (e.g. Bookstein 1991) according to the degree of their local definition (Zelditch *et al*. 2004). Thus, the type of landmarks delimiting the measure could influence the amount of ME.
- Flexibility of structures (Yezerinac *et al*. 1992). For example, some skull dimensions could change due to humidity, wear, erosion, etc.
- Variation between observers or within observers throughout measuring sessions (Palmeirim 1998, Yezerinac *et al*. 1992).
- Trait size. The relationship between trait size and ME has been discussed in several works (Lougheed *et al*. 1991, Palmeirim 1998, Pankakoski *et al*. 1987, Yezerinac *et al*. 1992) and it is generally considered that ME decreases as trait size increases.

There are different criteria for classification of ME. A primary classification of ME distinguishes between systematic and random errors (Rabinovich 1995). While systematic ME biases all the measurements in a particular direction, a random error causes variation of each measurement that is not directed in a particular way (Arnqvist & Martensson 1998). The consequences of high ME may be serious and depend on the kind of error. On one hand, a systematic error affects the accuracy of measurements because it causes a directional and repeatable deviation from the true value that can be confounded with biological variation (Arnqvist & Martensson 1998). On the other hand, random ME adds noise that increases the likelihood of type II errors, i.e., accepting false null hypotheses due to lack of statistical power (Bailey & Byrnes 1990, Yezerinac *et al*. 1992). Thus, ME can lead to biased results and invalid studies, especially in those disciplines where the interesting variation is subtle, such as in fluctuating asymmetry (Palmer & Strobeck 1986, Palmer 1994). While there are sometimes methods for evaluating and compensating for systematic ME (Rabinovich 1995) they are often case specific and complex. Since systematic ME occurs in particular situations but random ME is ubiquitous and often a potential problem, most works focus on the more general problem of random ME (Arnqvist & Martensson 1998).

Measurement error was not approached in early popular monographs on morphometrics (Bookstein 1978, Pimentel 1979). Perhaps for this reason most of the morphometric studies ignored ME until the late 1980s and 1990s, when several works on this topic were published (Bailey & Byrnes 1990, Björklund & Merilä 1997, Lougheed *et al*. 1991, Merilä & Björklund 1995, Palmeirim 1998, Pankakoski *et al*. 1987, Yezerinac *et al*. 1992).

Before the publication of these studies, some authors had already provided methods to assess ME and eliminate variables with high levels of ME, but none of these methods was satisfactory because they did not deal with the real problem of variation in repeated measurements between and among individuals (Bailey & Byrnes 1990). Model II ANOVA has been used to quantify the ME associated with repeated measurements, and is considered the best method for estimation of ME (Bailey & Byrnes 1990, Björklund & Merilä 1997, Lougheed *et al*. 1991, Merilä & Björklund 1995, Palmeirim 1998, Yezerinac *et al*. 1992).

Recently, the fast expansion of computer technology, digitizers, and two- or three-dimensional digital images are replacing caliper measurements as input data for morphometric studies (Rao & Suryawanshi 1998). In addition, the development of geometric morphometrics has revolutionized and has provided new perspectives to the study of morphology (Rohlf & Marcus 1993), giving a powerful tool to detect global as well as subtle morphological changes. Nevertheless, metric characters are still widely used and considered a good choice for certain morphological research, such as morphological

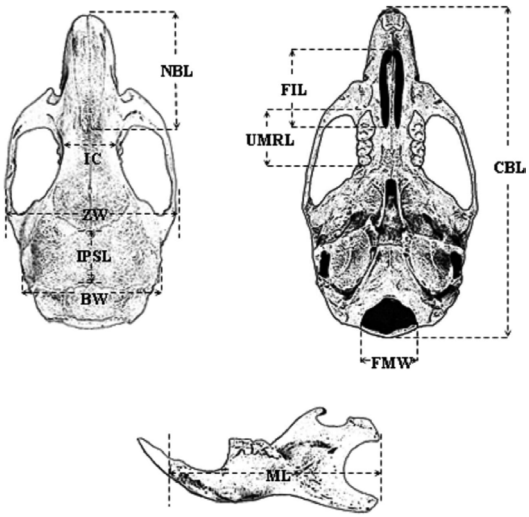

Fig. 1. Selected ten cranial measurements of skulls from house mice. For definitions *see* Material and Methods section. BW = braincase width; CBL = condylobasal length; FIL = foramen incisivum length; FMW = foramen magnum width; IC = interorbital constriction; IPSL = interparietal suture length; ML = mandible length; NBL = nasal bone length; UMRL = upper molar row length; ZW = zygomatic width.

integration (Young 2004). Although some studies have assessed ME in geometric morphometric techniques (Arnqvist & Martensson 1998, Corner *et al*. 1992, Mullin & Taylor 2002, Robinson *et al*. 2002, Valeri *et al*. 1998), few studies have dealt with the problem of differences between manual and computerized methods and with the real advantages that the latter may offer regarding ME (*see* Reig 1996).

In order to compare the effect of the measuring method on different sources of ME, here we assessed several measurements done with calipers and using two-dimensional digital images. Additionally, the relationship between mean trait size and ME was studied by pooling our caliper measurements with published data from previous works (Bailey & Byrnes 1990, Lougheed *et al*. 1991, Palmeirim 1998, Pankakoski *et al*. 1987, Yezerinac *et al*. 1992), as well as using our own data only.

Material and methods

Twelve skulls of the common house mouse *Mus musculus domesticus* were selected from a larger sample of animals captured between June 1999 and June 2000 in the experimental farms of the Universitat Autònoma de Barcelona, Spain. Only undamaged skulls from adult females were selected in order to avoid additional sources of variation. Ten common craniometric variables were measured on each skull (Fig. 1): mandible length (ML), from the most anterior point of the dentary to the most posterior point of the articular condyle; nasal bone length (NBL), from the most anterior to the most posterior points of the nasal bone; interorbital constriction (IC), minimum width across the interorbital region; zygomatic width (ZW), distance between outermost points of zygomatic arches; braincase width (BW), greatest width of braincase above posterior roots of zygomatic arches; interparietal suture length (IPSL), from the most anterior to the most posterior points of the interparietal suture; foramen magnum width (FMW), maximum internal distance between outermost points of the foramen magnum; condylobasal length (CBL), distance from the most anterior point of the premaxillae to the most posterior part of the occipital condyles; foramen incisivum length (FIL), maximum internal distance between posterior and anterior margins of the foramen incisivum; upper molar row length (UMRL), distance from the anterior surface of the first upper molar to the posterior surface of the third upper molar. To avoid variation due to asymmetry, bilateral traits (NBL, FIL, and UMRL) were always measured on the right-hand side. Traits were selected following three criteria: the magnitude of the linear measure (to evaluate the possible effect of the trait length in the amount of error), the possibility of fixing the caliper when assessing the measure, and the type of landmarks defining the trait.

All measurements were taken using manual and computerized procedures. In the manual method, a Mitutoyo no. 500-150 Digimatic Caliper (Mitutoyo American Corporation, USA) with 0.01 mm resolution and ± 0.02 mm accuracy was used. In the computerized method, lingual view of the right mandible, and dorsal and ventral views of the skull were photographed with a digital photographic camera (Nikon Coolpix 4300, Nikon, Japan) at 4.0 mega-pixel resolu-

tion by placing the skull and a reference scale at a constant distance. Morphometric data of digitized images were obtained using specialized software (Analysis 3.0, Soft Imaging Systems corp., Lakewood, Colorado, USA). This program takes measurements using 2 parallel bars, resembling calipers. In order to assess intra- and inter-observer variabilities in each method (Palmeirim 1998), all measures were taken 3 times by 3 different observers. After a training period, each observer measured all variables on the twelve skulls in three different sessions (one session per repetition). Each session was separated from the next one by several days. Skulls were measured in a random order and without consulting previous results. In effect, a total of 2160 measurements were obtained: 10 (traits) × 12 (individuals) × 2 (methods) × 3 (observers) × 3 (repetitions).

In order to estimate the components of global variation in each trait, the original dataset underwent a four-level nested ANOVA, with replication nested within observer, within method, and within individual (Bailey & Byrnes 1990, Lougheed *et al*. 1991, Yezerinac *et al*. 1992, Arnqvist & Martensson 1998, Thomson *et al*. 2001). The variance components [individual (S_a^2), method ($S_{(a)b}^2$), observer ($S_{(ab)c}^2$), and replication (S_e^2)] were calculated using the following formulas (Blackwell *et al*. 2006):

$$S_a^2 = \frac{\mathrm{MS}_a - \mathrm{MS}_{(a)b}}{bcd},$$

$$S_{(a)b}^2 = \frac{\mathrm{MS}_{(a)b} - \mathrm{MS}_{(ab)c}}{cd},$$

$$S_{(ab)c}^2 = \frac{\mathrm{MS}_{(ab)c} - \mathrm{MS}_e}{d}, \text{ and}$$

$$S_e^2 = \mathrm{MS}_e,$$

where MS_a, $\mathrm{MS}_{(a)b}$, $\mathrm{MS}_{(ab)c}$, MS_e, represent the mean squares of the individual, method, observer, and replication components, respectively, and b, c, and d represent the number of categories of each factor. Percent measurement error (%ME) due to a particular factor was calculated using the following formula:

$$\%\mathrm{ME}_e = \frac{S_e^2}{S_{\mathrm{Total}}^2} 100,$$

where S_{Total}^2 is the sum of among-individual variance (S_a^2), and within-individual variance (S_W^2), which in turn is the sum of within-individual components ($S_{(a)b}^2$, $S_{(ab)c}^2$ and S_e^2). Total percentage of measurement error ($\%\mathrm{ME}_{\mathrm{Total}}$) was calculated as:

$$\%\mathrm{ME}_{\mathrm{Total}} = \frac{S_w^2}{S_{\mathrm{Total}}^2} 100$$

Once variation due to measurement methodology (manual and computerized) was calculated, intra- and inter-observer components of ME were evaluated for each method with nested ANOVA, considering individual, observer and replicate as nested factors and traits as dependent variables. Total variance of each trait was partitioned into within- and among-individual components and %ME was calculated for each factor (observer and replication) and for the total within-individual variation using the formulas provided above. Repeatabilities were assessed for each trait and method. Repeatability measures the proportion of variance due to true variation among individuals and is defined as $S_a^2/S_{\mathrm{Total}}^2$, ranging from 0 to 1 (Arnqvist & Martensson 1998), or what is the same, as the percentage of variation not due to ME, ranging from 0 to 100 (100 – $\%\mathrm{ME}_{\mathrm{Total}}$; Falconer 1981). Further, the association between both types of error in each measurement technique was assessed by Pearson product-moment correlation.

In order to detect general patterns of regression between %ME and trait size and to choose the most appropriate model, data from previous published works (Bailey & Byrnes 1990, Lougheed *et al*. 1991, Palmeirim 1998, Pankakoski *et al*. 1987, Yezerinac *et al*. 1992) were pooled with our own data. To improve the comparability of the pooled data, among all available measurements we only selected linear skeletal measurements assessed with calipers. Among the selected works only the study of Palmeirim (1998) evaluated inter- and intra-observer ME. Therefore, we had a database with 21 measurements for the inter-observer error, while for the intra-observer error a total of 157 skeletal characters were used to detect possible associations between %ME and mean trait size. As it has been mentioned previously, %ME is a function of among-individual variance. Given that a dependence of the variance on mean trait size

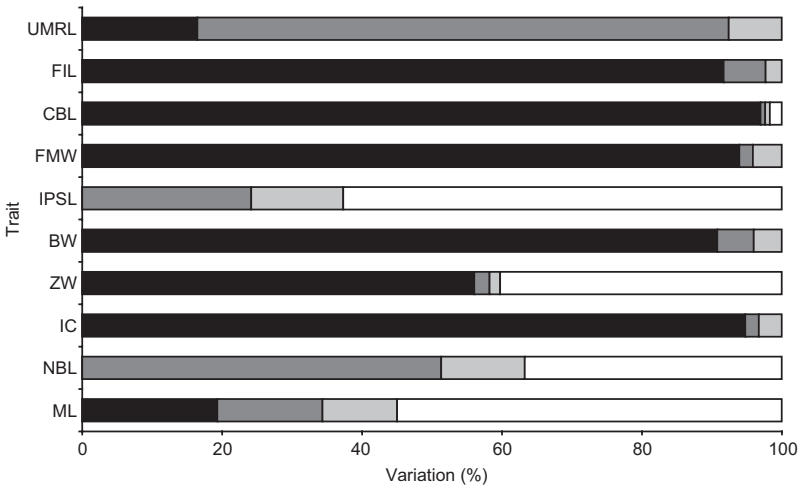

Fig. 2. Percentages of measurement errors of ten craniometric traits of house mice attributable to measuring procedure (black), observer (dark grey) and replicate (light grey). White portions indicate variation among individuals.

may exist, the relation between both parameters was initially tested with the CurveExpert v. 1.37 fitting programme for Windows. Due to the fact that a highly significant dependence of variance on mean trait size was detected, specifically described by the power function, the relationship between %ME and mean trait was defined as:

$$\%\mathrm{ME} = \frac{100 S_{\mathrm{w}}^2}{S_{\mathrm{w}}^2 + a x^b} = \frac{100}{1 + c x^b}$$

where x is the mean trait size, a and b the parameters of the function, and $c = a / S_{\mathrm{w}}^2$. The fitting of the pooled data to this and other non-linear and linear models was assessed. Afterwards, inter-observer, intra-observer, and total within %ME of own data were plotted against mean trait size in both measuring procedures, and the fitting with the defined model was assessed.

In order to limit the occurrence of type I error (i.e. 'false' positives) in sets of related tests, the sequential Bonferroni correction was systematically applied (Rice 1989). However, the strict application of this correction severely reduces the power of tests (Wright 1992). Such a sacrificial loss of power was avoided by choosing an experiment wise error rate higher than the usually accepted 5%. We used 10% as suggested by Wright (1992) and Chandler (1995).

Results

The first nested ANOVA performed for each character partitioned the total variance by separating variability among skulls from that introduced by differences between measuring methods, observers (inter-observer) and replicate measurements (intra-observer). In most traits, the method employed was the factor showing highest values of %ME, followed by observer and replicate factors, respectively (Fig. 2). However, in three traits (NBL, IPSL and UMRL), the error made by the observers was higher than the error due to measuring method. The median %ME introduced by the method was of 73.4%, ranging from 0.0% in IPSL and NBL to 97.0% in CBL. The observer factor showed median %ME of 5.6%, with values ranging from 0.7% in CBL to 75.9% in UMRL. The replication factor showed median %ME of 4.1%, with values ranging from 0.6% in CBL to 13.1% in IPSL (Fig. 2). Percentages of measurement error and repeatabilies for each method are summarized in Table 1 and compared in Fig. 3.

Manual measuring procedure

The error variance component due to different observers measuring the same skulls ranged from 2.4% in ZW to 56.0% in BW, and was higher than the intra-observer error in six out of the ten traits. The median for the inter-observer error was of 25.7%. The median intra-observer ME in all traits accounted for 25.0% and ranged from 2.4% in CBL to 40.5% in BW. When all within-individual variation (inter- and intra-observer variation) was considered together,

%ME ranged from 5.4% in CBL to 96.5% in BW, with a median value of 48.2%. The correlation between intra- and inter-observer errors, although not significant, showed a tendency for a positive relationship ($r = 0.608$, $p = 0.062$).

Computerized measuring procedure

As in the manual method, in the computerized procedure the intra-observer variation was also lower than the inter-observer variation in six out of the ten traits. Inter-observer component of error varied between 0.1% in CBL and 97.7% in UMRL, with a median value of 8.5%. Intra-observer component varied between 0.4% in CBL and 15.7% in IC, with a median value of 2.9%. Total within-individual %ME ranged from 0.47 in CBL to 100 in UMRL, with a median value of 15.3%. When removing UMRL from the analysis due to its atypically high %ME, the median total %ME decreased considerably to 5.9%. No correlation was observed between inter- and intra-observer %ME.

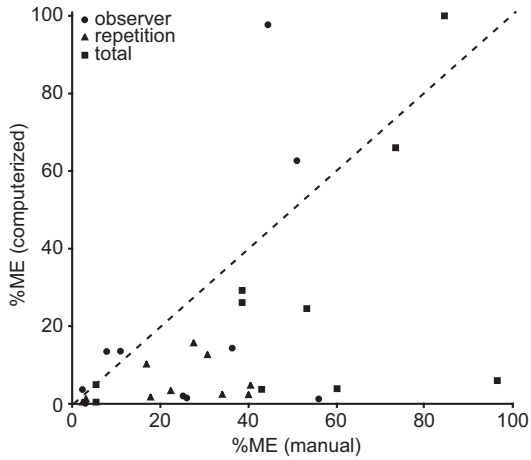

Fig. 3. Comparison of total, intra- and inter-observer percentages of measurement error (%ME) in the ten craniometric traits between the two measurement methods. The dashed line indicates the equality of values in both methods.

Trait size and ME

A significant dependence of variance on mean trait size was detected, and the best-fitted model

Table 1. Observer, replicate, and total percentages of measurement error (%ME) and total repeatabilities for each measuring procedure on skulls from common house mice.

Trait	Method	Mean (mm)	%ME			Repeatabilities ($100 - \%ME_{Total}$)
			Observer	Replicate	Total	
ML	Manual	11.5	25.2	17.8	43.0	57.0
	Computerized	11.9	2.0	1.7	3.7	96.3
NBL	Manual	7.8	51.1	22.4	73.4	26.6
	Computerized	7.7	62.6	3.4	66.0	34.0
IC	Manual	3.5	11.0	27.6	38.6	61.4
	Computerized	3.8	13.5	15.7	29.2	70.8
ZW	Manual	11.0	2.4	3.1	5.4	94.6
	Computerized	11.5	3.6	1.3	4.9	95.1
BW	Manual	9.4	56.0	40.5	96.5	3.5
	Computerized	10.3	1.3	4.7	6.0	94.0
IPSL	Manual	3.7	36.4	16.9	53.3	46.7
	Computerized	3.7	14.3	10.2	24.5	75.5
FMW	Manual	3.2	7.9	30.7	38.6	61.4
	Computerized	3.8	13.5	12.7	26.2	73.8
CBL	Manual	20.5	3.0	2.4	5.4	94.6
	Computerized	21.9	0.1	0.4	0.5	99.5
FIL	Manual	4.8	26.1	34.1	60.1	39.9
	Computerized	5.3	1.4	2.4	3.9	96.1
UMRL	Manual	3.3	44.4	40.0	84.5	15.5
	Computerized	3.6	97.7	2.3	100.0	0.0

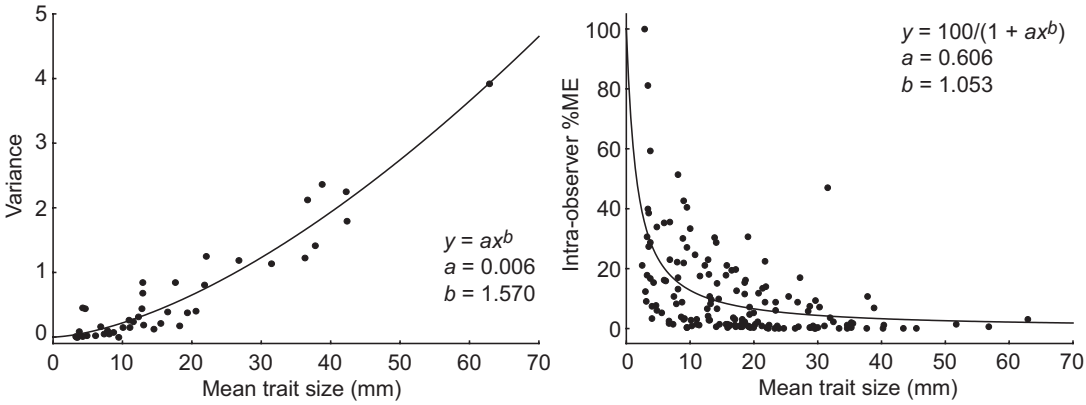

Fig. 4. Left hand-side panel: dependence of among-individual variance on mean trait size represented by a fitted power function ($r = 0.954$) based on 43 skeletal characters measured with calipers compiled from several studies. Right hand-side panel: intra-observer %ME *versus* mean trait size in 157 skeletal characters measured with calipers. Fitting of the defined function ($r = 0.553$) based on traits compiled from several studies is represented.

was the power function ($r = 0.954$, $p < 0.001$), described by the equation: $y = ax^b$ (Fig. 4). When intra-observer %ME from this and previous studies was plotted against mean trait size, a curvilinear dependence of ME on size was detected (Fig. 4), being the model defined in the material and methods among the best-fitted ($r = 0.553$, $p < 0.001$). A linear dependence of %ME on trait size was also observed, but the fitting of the lineal model was lower. For the inter-observer %ME no model showed a significant fitting when all measures were considered together. Nevertheless, when measures with unexpected high %ME values were removed from the analysis, the defined model showed a good fit ($r = 0.706$, $p < 0.05$).

The dependence of ME on trait size was also assessed with our own data only. In the manual procedure, curvilinear regression between mean traits size and inter-, intra-, and total %ME was not significant. However, when the trait with highest residuals (BW) was removed, a significant intra-observer %ME dependence on trait size was observed in the defined model ($r = 0.795$, $p < 0.05$).

In the computerized procedure, a curvilinear dependence of ME on mean measures of the traits was also observed when traits with highest residual values (UMRL and/or NBL) were removed. Thus, the fitting improved and the regression of intra-observer %ME and mean measure of the traits was significant ($r = 0.903$, $p < 0.05$), and the regressions of inter-observer and total %ME against mean trait size were nearly significant (inter-observer: $r = 0.963$, $p = 0.051$; total within: $r = 0.946$, $p = 0.0504$).

Discussion

The results obtained here show that differences in ME (i.e., accuracy) between the two measuring methods used in our study exist. In fact, the measuring method was the factor that most contributed to within-individual variation. The observer was the most relevant factor in three out of the ten traits (NBL, IPSL and UMRL). In the other seven traits, the method was the main contributing factor to within-individual variation. Therefore, the replication (understood as measurements taken by the same observer and with the same method on the same trait) was the least important factor contributing to within-individual variation, although in two traits (NBL and IPSL) this factor was more important than the method. Due to the high %ME related to the method employed (median of 73.4%), measurements obtained using different procedures are not comparable. These results agree with those obtained by Reig (1996). This high variation did not depend on trait size and was not correlated with inter- or intra-observer variation. Therefore, several causes not affecting inter- and intra-observer variation could be increasing %ME

due to the procedure used. One of these possible causes is the pass from three to two dimensions. In the manual procedure, the traits are measured in three-dimensional space, while in the computerized method we measured photographs, which are two dimensional images from three dimensional structures. The lowest percentages of ME due to the measurement procedure were observed in characters measured on the first plane of the photograph (NBL and IPSL in dorsal view) or in flat structures (ML). Although we tried to choose characters that were defined by landmarks comprised in the same plane, two dimensional measurements of three dimensional structures could be an important factor affecting variation between both measuring procedures.

When comparing the two methods, results showed that repeatabilities were higher in the computerized procedure than in the manual one (*see* Table 1). Specifically, when UMRL was removed from the computerized procedure (because it showed atypical high values), median percentages of ME were between 7.1 and 8.0 times higher in the manual than in the computerized procedure. Although the inter-observer variation of five traits (UMRL, FMW, ZW, IC, and NBL) was higher in the computerized than in the manual procedure (Fig. 3), the median value of the inter-observer error was three times higher in the latter than in the former one. In addition, three of the traits with higher %ME values in the computerized procedure showed low and similar percentages in both methods (FMW, ZW and IC). Only the inter-observer variation of one trait (UMRL) was considerably higher in the computerized method, probably because of poor definition of the variable (Bailey & Byrnes 1990, Reig 1996), which may have been less important in the manual procedure. It should be noticed that the reduction of the intra-observer error in the computerized method might produce an increase of the inter-observer error. Therefore, slight differences in trait definition between observers could have been more conspicuous in computerized method due to higher intra-observer repeatability. In addition to the observed quantitative differences between measuring procedures, inter- and intra-observer %ME were not correlated in the measurements obtained with the computer, whereas those measurement obtained with calipers showed a tendency for a positive correlation. Palmeirim (1998) also observed this relationship and obtained similar r-values in measurements of bat skulls taken with calipers. This author suggested that factors increasing within-observer variation are likely the same that make traits more susceptible to be measured slightly different by different observers.

One of the potential factors affecting within-individual variation is trait size. Significant dependence of intra-observer %ME on mean trait size has been detected in several studies (Pankakoski *et al*. 1987, Yezerinac *et al*. 1992, Palmeirim 1998, Lajus 2001), and some authors suggested a linear dependence between %ME and trait size (Yezerinac *et al*. 1992, Palmeirim 1998). Nevertheless, there is some debate about this issue in the literature (Pankakoski *et al*. 1987, Lougheed *et al*. 1991, Yezerinac *et al*. 1992, Palmeirim 1998). Results obtained pooling data from different studies point to a similar direction. However, it should be noticed that a curvilinear rather than a linear dependence of intra-observer %ME on trait length was detected. This dependence was highly significant for the defined curvilinear model, although it was also significant for the linear model, which showed a lower r. Our results support the assumption that intra-observer %ME decreases as trait size increases, especially in small measures, because the true variation among individuals increases at faster rates as size increases; therefore, the within-individual variance represents a smaller portion of the overall variation. As characters reach higher mean size values (around 10 mm in our results) %ME decreases more slowly with trait size because among-individual variance increases at a slower rate. In characters with mean sizes above 35 mm the slope of the function is almost zero because variance and mean size of the trait increase at a constant rate. As mean trait size decreases, the dispersion of the measures around the function increases, and therefore higher residuals are observed (Fig. 4). Our interpretation is that when trait size is small, the importance of factors increasing ME, other than trait size, is higher. For instance, an important factor increasing within-individual variation in small measures could be the resolution in identifying the landmarks that define the charac-

ter (due to limitation of human eye). The inter-observer variation did not show a significant dependence on trait size, but a significant curvilinear association was observed when outliers were removed. Palmeirim (1998), in his study on the big brown bat *Eptesicus fuscus*, detected a non-significant tendency for a negative relationship between the mean size of the character and the amount of variance introduced by different observers, and suggested that variables with smaller mean sizes generally had a larger inter-observer component of ME. Our results agree with those obtained by Palmeirim (1998), although no relationship was observed when all traits were considered; however, a significant curvilinear dependence of inter-observer variation of trait was observed when measurements with atypical high %ME were removed. Several causes may explain these results. On one hand, it is worth noting that a lesser amount of variables was available for the inter-observer than for the intra-observer variation, and consequently a smaller range (3.24–20.53 mm in inter-observer ME *versus* 2.48–62.85 mm in intra-observer ME) of mean trait sizes, making more difficult to detect the patterns and yielding less robust results. On the other hand, we suspect that some factors not influencing intra-observer variation could influence inter-observer variation of some traits (for instance, some criteria differences between observers), and that factors affecting both types of variation could have stronger effects on inter-observer %ME (i.e. differences in the positioning of the caliper, the pressure applied to close the caliper, etc.; Palmeirim 1998).

Results obtained with our data showed several patterns. The fitting of the model was considerably higher in the computerized than in the manual procedure and significant or near-significant dependence of ME on size was observed for the total within-individual, the inter-observer, and the intra-observer variations in the former method, while only the intra-observer variation showed a significant dependence on trait size in the latter one. Therefore, considering the relationship between ME and trait size, the manual method seems to be more vulnerable than the computerized one to factors adding noise to this association. Although in the computerized method some measures (UMRL and NBL) also bias the pattern, the dependence was tighter than in the manual procedure when these traits were removed. In this sense, Lougheed *et al*. (1991) also noticed that factors other than trait size could be affecting the amount of error obtained with calipers.

In conclusion, when comparing the accuracy of both methods in measuring anatomical traits of the skull of the house mouse, we found that the computerized measuring procedure is more precise than the manual method; and the dependence of inter- and intra-observer variation on trait size is more evident in the computerized one. This suggests that factors influencing ME (others than trait size) have a more important role in measurements taken with calipers. However, it should be noted that further studies assessing ME in computerized measurements of craniometric variables are needed to detect potential factors influencing within-individual variation and to obtain more reliable patterns of the dependence of ME on the size of the character.

Acknowledgements

We are especially grateful to Marta Busquet for participating in this study by measuring skulls, and to Jacint Ventura, Nicola Drew, Neus Martínez, Rolando González, and Silvina Van der Molen for their valuable comments on an early draft of the manuscript. We also thank Eduardo Ayuso and Fátima Bosch for their help in the use of the digital analysis software.

References

Ackermann, R. R. 2005: Ontogenetic integration of the hominoid face. — *Journal of Human Evolution* 48: 175–197.

Arnqvist, G. & Martensson, T. 1998: Measurement error in geometric morphometrics: empirical strategies to assess and reduce its impact on measures of shape. — *Acta Zoologica Academiae Scientiarum Hungaricae* 44: 73–96.

Badyaev, A. V., Foresman, K. R. & Fernandes, M. V. 2000: Stress and developmental stability: vegetation removal causes increased fluctuating asymmetry in shrews. — *Ecology* 81: 336–345.

Bailey, R. C. & Byrnes, J. 1990: A new, old method for assessing measurement error in both univariate and multivariate morphometric studies. — *Systematic Zoology* 39: 124–130.

Blackwell, G. L., Basset, S. M. & Dickman, C. R. 2006.

Measurement error associated with external measurements commonly used in small-mammals studies. — *Journal of Mammalogy* 87: 216–223.

Björklund, M. & Merilä, J. 1997: Why some measures of fluctuating asymmetry are so sensitive to measurement error. — *Annales Zoologici Fennici* 34: 133–137.

Bonner, J. T. 1965: *Size and cycle: an essay on the structure of biology*. — Princeton University Press, Princeton.

Bookstein, F. L. 1978: *The measurement of biological shape and shape change*. — Springer, Berlin.

Bookstein, F. L. 1991: *Morphometric tools for landmark data: geometry and biology*. — Cambridge University Press, Cambridge.

Chandler, C. R. 1995: Practical considerations in the use of simultaneous inference for multiple tests. — *Animal Behaviour* 49: 524–527.

Cheverud, J. M. 1996: Quantitative genetic analysis of cranial morphology in the cotton-top (*Saguinus oedipus*) and saddle-back (*S. fuscicollis*) tamarins. — *Journal of Evolutionary Biology* 9: 5–42.

Corner, B. D., Lele, S. & Richstmeier, J. T. 1992: Measuring precision of three-dimensional landmark data. — *Journal of Quantitative Anthropology* 3: 347–359.

Falconer, D. S. 1981: *Introduction to quantitative genetics*. — Longman, London.

González-José, R., Dahinten, S. L., Luis, M. A., Hernandez, M. & Pucciarelli, H. M. 2001: Craniometric variation, and, the settlement of the Americas: testing hypotheses by means of R-matrix and matrix correlation analyses. — *American Journal of Physical Anthropology* 116: 154–165.

Green, A. J. 2001: Mass/length residuals: measures of body condition or generators of spurious results? — *Ecology* 82: 1473–1483.

Harper, J. L., Lovell, P. H. & Moore, K. G. 1970: The shapes and sizes of seeds. — *Annual Review of Ecology and Systematics* 1: 327–356.

Klingenberg, C. P. & Ekau, W. 1996: A combined morphometric and phylogenetic analysis of an ecomorphological trend: pelagization in Antarctic fishes (Perciformes: Nototheniidae). — *Biological Journal of the Linnean Society* 59: 143–177.

Kruuk, L. E. B., Clutton-Brock, T. H., Slate, J., Pemberton, J. M., Brotherstone, S. & Guinness, F. E. 2000: Heritability of fitness in a wild mammal population. — *Proceedings of the National Academy of Sciences of the United States of America* 97: 698–703.

Lajus, D. L. 2001: Variation patterns of bilateral characters: variation among characters and among populations in the White Sea herring, *Clupea pallasi marisalbi* (Berg) (Clupeidae, Teleosti). — *Biological Journal of the Linnean Society* 74: 237–253.

Lieberman, D. E. 1998: Sphenoid shortening and the evolution of modern human cranial shape. — *Nature* 393: 158–162.

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: 108–436.

Merilä, J. & Björklund, M. 1995: Fluctuating asymmetry and measurement error. — *Systematic Biology* 44: 97–101.

Mullin, S. K. & Taylor, P. J. 2002: The effects of parallax on geometric morphometric data. — *Computers in Biology and Medicine* 32: 455–464.

Palmeirim, J. M. 1998: Analysis of skull measurements and measurers: can we use data obtained by various observers? — *Journal of Mammalogy* 79: 1021–1028.

Palmer, A. R. 1994: Fluctuating asymmetry analyses: a primer. — In: Markow, T. A. (ed.), *Developmental instability: its origins and implications*: 335–364. Kluwer, Dordrecht.

Palmer, A. R. & Strobeck, C. 1986: Fluctuating asymmetry: measurement, analysis, patterns. — *Annual Review of Ecology and Systematics* 17: 391–421.

Pankakoski, E., Väisänen, R. A. & Nurmi, K. 1987: Variability of muskrat skulls: measurement error, environmental modification and size allometry. — *Systematic Zoology* 36: 35–51.

Pimentel, R. A. 1979: *Morphometrics: the multivariate analysis of biological data*. — Kendall & Hunt, Dubuque.

Rabinovich, S. R. 1995: *Measurement errors: theory and practice*. — American Institute of Physics, New York.

Rae, T. C. 1998: The logical basis for the use of continuous characters in phylogenetic systematics. — *Cladistics* 14: 221–228.

Rao, C. R. & Suryawanshi, S. 1998: Statistical analysis of shape through triangulation of landmarks: a study of sexual dimorphism in hominids. — *Proceedings of the National Academy of Sciences of the United States of America* 95: 4121–4125.

Reig, S. 1996: Correspondence between interlandmark distances and caliper measurements. — In: Marcus, L. F., Coti, M., Loy, A., Naylor, G. & Slice, D. (eds.), *Advances in morphometrics*: 371–385. Plenum Press Co., New York.

Rice, W. R. 1989: Analyzing tables of statistical tests. — *Evolution* 43: 223–225.

Robinson, D. L., Blackwell, P. G., Stillman, E. C. & Brook, A. H. 2002: Impact of landmark reliability on the planar Procrustes analysis of tooth shape. — *Archives of Oral Biology* 47: 545–554.

Rohlf, F. J. & Marcus, L. F. 1993: A revolution in morphometrics. — *Trends in Ecology & Evolution* 8: 129–132.

Searcy, W. A. 1979: Morphological correlates of dominance in captive male red-winged blackbirds. — *Condor* 81: 417–420.

Strong, D. R. Jr. 1983: Natural variability and the manifold mechanisms of ecological communities. — *American Naturalist* 122: 636–660.

Thomson, H. L., Basmadjian, A. J., Rainbird, A. J., Razavi, M., Avierinos, J. F., Pellikka, P. A., Bailey, K. R., Breen, J. F. & Enriquez-Sarano, M. 2001. Contrast echocardiography improves the accuracy and reproducibility of left ventricular remodeling measurements. A prospective, randomly assigned, blinded study. — *Journal of American College of Cardiology* 38: 867–875.

Valeri, C. J., Cole, T. M. III, Lele, S. & Richstmeier, J. T. 1998: Capturing data from three-dimensional surfaces using fuzzy landmarks. — *American Journal of Physical Anthropology* 107: 113–124.

Vincent, S. E., Herrel, A. & Irschick, D. J. 2004: Sexual dimorphism in head shape and diet in the cottonmouth snake (*Agkistrodon piscivorus*). — *Journal of Zoologyy* 264: 53–59.

Willmore, K. E., Klingenberg, C. P. & Hallgrimsson, B. 2005: The relationship between fluctuating asymmetry and environmental variance in rhesus macaque skulls. — *Evolution* 59: 898–909.

Wright, S. P. 1992: Adjusted *p*-values for simultaneous inference. — *Biometrics* 48: 1005–1013.

Yezerinac, S. M., Lougheed, S. C. & Handford, P. 1992: Measurement error and morphometric studies: statistical power and observer experience. — *Systematic Biology* 41: 471–482.

Young, N. 2004: Modularity and integration in hominoid scapula. — *Journal of Experimental Zoology* 302B: 226–240.

Zelditch, M. L., Swiderski, D. L., Sheets, H. D. & Fink, W. L. 2004: *Geometric Morphometrics for biologists: a primer*. — Elsevier Academic Press, New York.