Zoomorphology (2010) 129:185–194
DOI 10.1007/s00435-010-0111-y

ORIGINAL PAPER

How long is a piece of *Strix*? Methods in measuring and measuring the measurers

Paolo Viscardi · Manabu Sakamoto · Julia D. Sigwart

Received: 31 March 2009 / Revised: 1 July 2010 / Accepted: 2 July 2010 / Published online: 16 July 2010

Abstract An experiment to quantify intra- and interobserver error in anatomical measurements found that interobserver measurements can vary by over 14% of mean specimen length; disparity in measurement increases logarithmically with the number of contributors; instructions did not reduce variation or measurement disparity; scale of the specimen influenced the precision of measurement (relative error increasing with specimen size); different methods of taking a measurement yielded different results, although they did not differ in terms of precision, and topographical complexity of the elements being considered may potentially influence error (error increasing with complexity). These results highlight concerns about introduction of noise and potential bias that should be taken into account when compiling composite datasets and meta-analyses.

Keywords Measurement · Methods · Composite data · Morphometrics · Error · Meta-analysis · Anatomy · Observer error

Communicated by T. Bartolomaeus.

P. Viscardi (✉)
Horniman Museum and Gardens, 100 London Road, Forest Hill SE23 3PQ, UK
e-mail: pviscardi@horniman.ac.uk

M. Sakamoto
Department of Earth Sciences, University of Bristol, Bristol, UK

J. D. Sigwart
School of Biological Sciences, Queen's University Belfast, Northern Ireland, UK

Introduction

The act of taking a measurement is a learned procedure that becomes intuitive with experience, but for scientific rigour to be maintained that act needs to be consistent and repeatable by independent workers.

Use of homologous structures as anatomical landmarks provides a powerful method of explicitly demarking a measurement (Dryden and Mardia 1998; Adams et al. 2004). However, more ambiguous or variable mathematical landmarks (points defined by distinguishable features such as extremities of structures) and pseudo-landmarks (points on an outline or defined by their relative position to other landmarks) are frequently called upon in traditional morphometrics to enable distances between outline and extremity features to be quantified (Rohlf 1990).

In well-studied taxa, consistency can be attempted by citing methods used in previous studies (e.g. Wiig 1985, 1986; Freeman and Jackson 1990; Lee and Mill 2004). In some instances, authors may assume that common measurements do not need reporting, for example 'skull length' (Thomason 1991; Currie and Zhao 1993; Chen et al. 1998; Wroe et al. 2005). Unfortunately, skulls are complex structures and their length can be derived using a variety of methods (compare Erickson et al. 2003; Herrel et al. 2001; Sues 1977; Meyers et al. 2002: see Fig. 1). Consistent methods can be difficult to achieve when working from composite data, collaborating on a project, or even when dealing with specimens represented by a wide range of sizes where practical limitations on equipment usage can arise (Marcus et al. 2000).

Some workers explicitly exclude the use of data collected in other studies (Delany and Healy 1965; Lee and Mill 2004), others readily use composite data from multiple sources (Holtz 1994; Ashton et al. 2000; Clegg and Owens

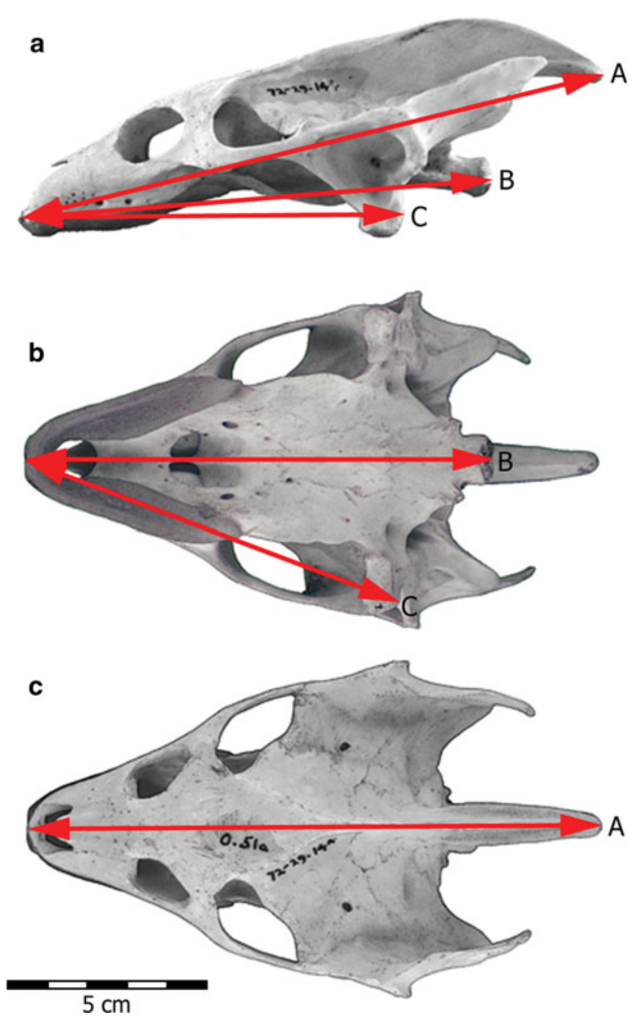

Fig. 1 Techniques of determining skull length in a non-avian sauropsid. The three methods give visibly different results. *A* prosthion to posterior margin of parietals (Herrel et al. 2001; Erickson et al. 2003); *B* prosthion to occipital condyle (Sues 1977); *C* prosthion to quadrate condyle (Meyers et al. 2002)

2002; Nudds et al. 2004; Turner et al. 2007), but neither approach has a quantitative grounding. Past testing of interobserver error has quantified variation between two collaborating workers (Kieser and Groeneveld 1987; Lynch et al. 1997; Bailey et al. 2004) and where considered it was not found to be significantly greater than intraobserver error (Lynch et al. 1997; Bailey et al. 2004). Unfortunately, these findings can be taken out of context, such as in the statement by Lee and Mill (2004); based on the finding of Lynch et al. (1997) that "measurement errors have been shown to be insignificant, even when made by more than one individual". This statement is only valid when considering measurements made by two collaborating researchers and is insufficient when considering a body of data derived from multiple independent studies (such as those by Holtz 1994; Ashton et al. 2000; Clegg and Owens 2002; Nudds et al. 2004; Turner et al. 2007; Yates 2007). Of course, the impact of measurement error is always relative to the question being addressed, which is why it is important to understand likely error margins for measurements derived by independent workers if their results are to be included in composite datasets. The aim of this study is to quantify interobserver error within a large sample ($n = 51$) of independent anatomists, using interobserver precision (parity in measurement) as an indicator of the ability of independent workers to obtain the same measurements from the same specimens. The data arising will allow a preliminary methodological assessment of major sources of error, which can inform researchers who make use of composite quantitative morphological datasets.

The hypotheses tested in this preliminary study are as follows:

1a. Interobserver variation will be significantly larger than intraobserver variation (*contra* Lee and Mill 2004).
1b. Average disparity between interobserver measurements will increase in proportion to increasing number of observers.
2. Instructions will significantly reduce interobserver measurement variation.
3. Greater complexity will significantly increase interobserver measurement variation.
4. Measurement variation will be significantly influenced by scale of specimen.
5a. Different tools will produce significantly different measurements.
5b. Different tools will not produce significantly different measurement variation.
6a. Different landmark use will result in significantly different measurements.
6b. Different landmark use will not result in significantly different measurement variation.

Materials and methods

In order to quantify the error that might be expected from composite data collected by multiple independent researchers, fifty-one professional morphologists contributed to an experiment carried out at the 56th Symposium of Vertebrate Palaeontology and Comparative Anatomy, Dublin, 2008. Each individual was allowed to participate in the experiment only once and was randomly assigned (by coin toss) to one of two categories: with instruction or without instruction. All operators were verbally instructed to measure and record the 'length' of three osteological specimens with specific tools: *Strix aluco* Linnaeus, 1758 femur, and *Mustela erminea* Linnaeus, 1758 skull, using digital callipers (Faithfull, 150 mm, accurate to 0.01 mm), and *Canis lupus familiaris* Linnaeus, 1758 skull using a steel tape measure (Kingstons of Dublin, 5 m, accurate to 1 mm). The 'with instruction' operators were shown instructional diagrams indicating measurement landmarks and distance direction (Fig. 2). Two sets of specimens were used, each consistently associated with specific tools and experimental category (with/without instruction). Thus, the experiment was designed to collect data that would simultaneously examine several different independent variables within the study: comparison of tool use (callipers versus tape measure), element complexity (femur versus skull), as well as the influence of detailed instructions.

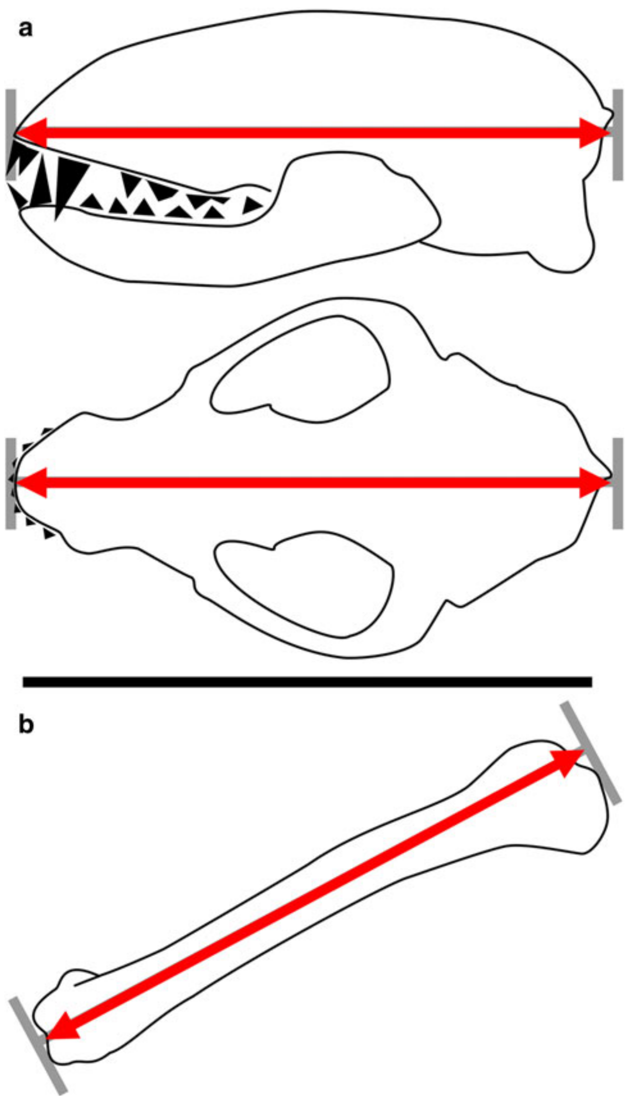

Fig. 2 Instructional stylised diagrams provided in the interobserver experiment ('with instruction' set), showing preferred landmarks to be used for measuring **a** skull length-prosthion to inion distance, equivalent to L1 in the intraobserver experiment; **b** long bone length. Operators measuring 'without instruction' were told only to 'measure the length' without visual aids or additional guidance

The range of measurements (maximum length minus minimum length) for a specimen, expressed as a proportion of total mean specimen length, was used to illustrate the disparity in measurement resulting from variation between workers. In an effort to assess the relationship between disparity and number of observers, a simple model was produced by taking repeated random subsamples of measurements (from 2, 3, 4, 5 and 10 contributors) from the interobserver experiments and comparing disparity for each subsample. The average disparity was calculated from 100 sets of comparisons (equalling a maximum of 2,500 comparisons between subsamples of two observers and a minimum of 1,500 comparisons between subsamples of ten observers).

To provide an intraobserver baseline and allow a more tightly constrained investigation into the influence of tools, scale and complexity of specimen and landmark choice a single worker (PV) recorded multiple measurements ($n = 30$) of skull length in one specimen each of *C. l. familiaris* and *M. erminea* skulls and femur length in a specimen of *S. aluco*. Intraobserver *C. l. familiaris* skull measurements were taken with both tape measure and callipers to compare the influence of tool used. To investigate the influence of landmark choice on measurement obtained, two measures of skull length were taken: L1 = anterior tip of the premaxilla (prosthion) to the posterior tip of the sagittal crest (inion) after Constantinescu (2002, equivalent to the landmarks used in the interobserver experiment) and L2 = posterior surface of occipital condyle to the prosthion, after Van Valkenburgh (1990).

Statistical analyses used Microsoft Excel 2003 and PAST (version 1.89; Hammer et al. 2001). Coefficient of variation was calculated as a scale-independent measure of variation of observations, and a two-tailed Fligner–Killeen (F–K) test for homogeneity of variances was used to identify differences in the variation between samples (Fligner and Killeen 1976; Conover et al. 1981). Where length measurement data were compared, a two-sample unequal variance *T*-test was used. A Bonferroni correction was applied to the results of comparative sets of statistical tests where multiple comparisons were made, in order to maintain the familywise error rate and reduce the opportunity for Type 1 error.

Results

Prior to analysis, 12 errors were identified and corrected in 153 recorded data points in the interobserver experiment (transpositions between *S. aluco* and *M. erminea*, units of cm rather than mm, loss of the first digit when recording from the tape measure). Some individuals recorded measurements to a precision of 0.5 mm using a tape measure only graduated to 1 mm, based on their own judgment. Corrected data fell within one standard deviation of their respective datasets and were therefore included in analysis. For greater clarity in identifying the multiple comparisons made in this study, the results of all comparisons of variation are shown in Table 1.

Test 1a: Interobserver variation compared to intraobserver variation. Interobserver variation was significantly different to intraobserver variation based on a comparison of errors for each of the six study elements separately using the F–K statistic and considering additional possible outcomes from variation in interpretation of landmarks (Table 1; lowest level of support $P = 9.94 \times 10^{-4}$, highest level of support $P = 5.44 \times 10^{-6}$ after Bonferroni correction for ten paired comparisons in the set). Comparison of coefficients of variation (Tables 2, 3) indicates that interobserver variation was greater than intraobserver variation in all instances. The greatest disparity in interobserver measurements of one specimen was 22 mm or 14.2% when expressed as a proportion of mean specimen length. The average disparity in interobserver measurement

Table 1 Comparisons of variation between sets of measurement data using two-tailed Fligner–Killeen (F–K) test for homogeneity of variances

Test 1a: intraobserver vs. interobserver	Interobserver with instructions	Interobserver no instructions
Intraobserver *C. l. familiaris* L1 tape measure	*5.44×10^{-06}*	*1.68×10^{-05}*
Intraobserver *C. l. familiaris* L2 tape measure	*8.96×10^{-06}*	*1.58×10^{-05}*
Intraobserver *M. erminea* L1	*7.04×10^{-06}*	*1.87×10^{-05}*
Intraobserver *M. erminea* L2	*8.04×10^{-06}*	*8.51×10^{-05}*
Intraobserver *S. aluco*	*5.31×10^{-05}*	*9.94×10^{-04}*
Test 2: with vs. without instructions		Same species, no instructions
Interobserver *C. l. familiaris* with instructions		0.265704
Interobserver *M. erminea* with instructions		0.30546
Interobserver *S. aluco* with instructions		0.061986
Test 3a: simple vs. complex (interobserver)	*S. aluco* with instructions	*S. aluco* no instructions
Interobserver *M. erminea* with instructions	1	*0.000236176*
Interobserver *M. erminea* no instructions	1	*0.0138264*
Test 3b: simple vs. complex (intraobserver)		Intraobserver *S. aluco*
Intraobserver *M. erminea* L1 callipers		1.0822
Intraobserver *M. erminea* L2 callipers		*0.0036876*
Test 4a: large vs. small elements (interobserver)	*M. erminea* with instructions	*M. erminea* no instructions
Interobserver *C. l. familiaris* with instructions	0.079088	0.158176
Interobserver *C. l. familiaris* no instructions	0.52152	0.154868
Test 4b: large vs. small elements (intraobserver callipers only)	*M. erminea* (L1)	*M. erminea* (L2)
Intraobserver *C. l. familiaris* L1	*0.0059808*	0.203132
Intraobserver *C. l. familiaris* L2	*0.00040632*	*0.03425*
Test 5b: callipers vs. tape variation		Only tool differs (tape)
Intraobserver *C. l. familiaris* L1 callipers		*0.0053882*
Intraobserver *C. l. familiaris* L2 callipers		1
Test 6b: L1 vs. L2 variation		Only landmark differs (L2)
Intraobserver *C. l. familiaris* L1 callipers		0.88758
Intraobserver *C. l. familiaris* L1 tape		1
Intraobserver *M. erminea* L1 callipers		0.05535

All probability values are Bonferroni corrected to account for multiple comparisons within sets; significant values are identified using italics. L1 and L2 relate to different landmarks used to obtain measurements (details in "Materials and methods" section)

Table 2 Interobserver results

	n	Geometric mean (mm)	Median (mm)	Mean (mm)	Standard deviation	Standard error	Coefficient of variation	Minimum (mm)	Maximum (mm)	Skewness	Kurtosis
With instruction											
C. l. familiaris skull	24	149.33	149.50	149.38	3.87	0.79	±2.59	144.00	155.00	0.00	−1.60
M. erminea skull	24	43.58	43.80	43.59	0.85	0.17	±1.95	41.00	44.60	−1.79	2.96
S. aluco femur	24	52.26	52.69	52.27	0.91	0.19	±1.74	49.75	53.00	−1.77	1.94
Without instruction											
C. l. familiaris skull	27	155.93	157.00	156.00	4.76	0.92	±3.05	148.00	166.00	0.04	−0.89
M. erminea skull	27	47.35	47.53	47.35	0.78	0.15	±1.64	44.70	48.05	−2.31	4.78
S. aluco femur	27	52.44	52.49	52.44	0.16	0.03	±0.31	51.98	52.81	−1.00	1.74

Measurements of specimens made by a group of professional anatomists (*n*, number of individuals). 'With instruction' operators were provided diagrammatic instruction (Fig. 2, after Constantinescu 2002). 'Without instruction' operators were verbally told to record 'length' and used the preferred method of the measurer

Table 3 Intraobserver results

	n	Geometric mean (mm)	Median (mm)	Mean (mm)	Standard deviation	Standard error	Coefficent of variation	Minimum (mm)	Maximum (mm)	Skewness	Kurtosis
C. l. familiaris (L1) callipers	30	154.73	154.80	154.73	0.15	0.03	±0.10	154.34	154.95	−0.64	−0.62
C. l. familiaris (L1) tape measure	30	152.82	153.00	152.82	0.43	0.08	±0.28	152.00	153.50	−0.24	−0.64
C. l. familiaris (L2) callipers	30	142.60	142.61	142.60	0.27	0.05	±0.19	142.04	143.65	1.56	5.80
C. l. familiaris (L2) tape measure	30	142.20	142.00	142.20	0.34	0.06	±0.24	142.00	143.00	1.33	0.37
M. erminea (L1) callipers	30	49.59	49.60	49.59	0.02	0.00	±0.05	49.54	49.65	−0.39	0.37
M. ermineat (L2) callipers	30	49.84	49.85	49.84	0.04	0.01	±0.07	49.77	49.91	0.02	−0.93
S. aluco femur callipers	30	56.58	56.59	56.58	0.02	0.00	±0.04	56.54	56.60	−0.71	−0.98

Measurements of skull length made by an individual operator. Where L1 indicates specimens measured from prosthion to inion (equivalent to 'with instruction' records in the interobserver experiment, after Constantinescu 2002); L2 indicates specimens measured from prosthion to the occipital condyle (after van Valkenburgh 1990) and "callipers" or "tape measure" indicate the tool used to take the measurement

for all specimens was 8.2%. The greatest disparity in intraobserver measurements for one specimen was 1.6 mm or 1.1%, whilst the average disparity in intraobserver measurement for all specimens was 0.62%. Interobserver variation was significantly larger than intraobserver error, allowing acceptance of hypothesis 1a.

Test 1b: Relationship between number of observers and the disparity between observations. Disparity may increase as more operators are added to the study. The relationship between modelled disparity in measurement and the number of observers follows a series of logarithmic curves, all with a coefficient of determination of $R^2 > 0.99$ (curves in the text of Fig. 3). The curves predict the total disparity for each set of actual interobserver measurements from the interobserver experiment to within about 1% (Fig. 3). The resampling of data suggests that disparity between inter-observer measurements will increase in proportion to increasing number of observers, providing support for hypothesis 1b, a prediction which should be tested in future studies.

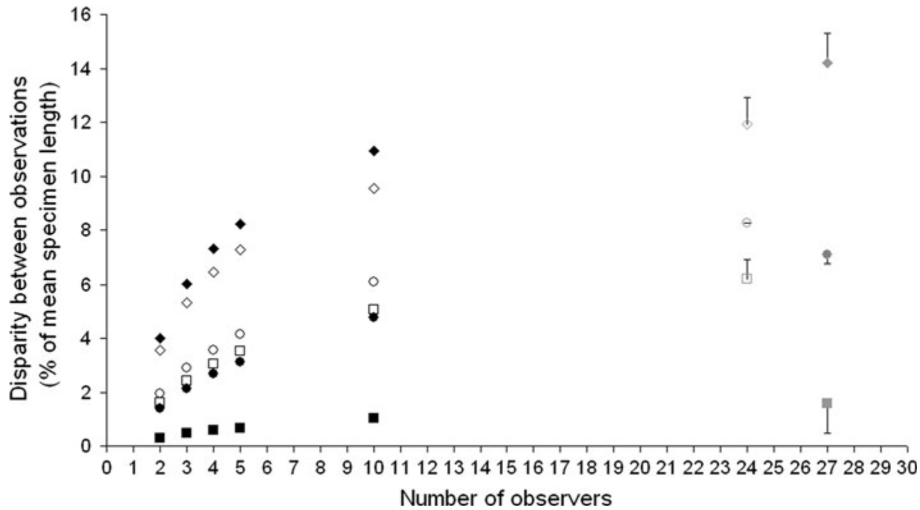

Fig. 3 Graph showing modelled logarithmic relationship between number of observers and disparity (expressed as percentage of mean specimen length). *Diamonds* = *C. l. familiaris*; circles = *M. erminea*; squares = *S. aluco*; *solid symbols* = specimens measured without instructions; *closed symbols* = specimens measured with instructions; *black-lined symbols* = values obtained by random resampling of experimental data; *grey-lined symbols* = actual measurements of disparity from interobserver experiments; *error line* = difference between actual disparity from interobserver experiments and disparity predicted by projecting logarithmic relationship curves. Curves all have $R^2 > 0.99$ and are described as follows (from *top* to *bottom* of graph): *C. l. familiaris* $y = 4.2701.\mathrm{Ln}(x) + 1.2412$; *C. l. familiaris* with instructions $y = 3.6939.\mathrm{Ln}(x) + 1.1938$; *M. erminea* with instructions $y = 2.5743.\mathrm{Ln}(x) + 0.0719$; *S. aluco* with instructions $y = 2.1475.\mathrm{Ln}(x) + 0.0819$; *M. erminea* $y = 2.0866.\mathrm{Ln}(x) - 0.1378$; *S. aluco* $y = 0.4392.\mathrm{Ln}(x) - 0.0162$

Test 2: Comparison of interobserver variation with and without instructions. No significant difference in variation was seen between interobserver measurements taken with or without instructions, comparing the results between measurement sets for each of the three study elements using the F–K statistic (Table 1; lowest level of support $P = 0.30546$, highest level of support $P = 0.061986$ after Bonferroni correction for three paired comparisons in the set). Instructions did not significantly reduce interobserver variation, allowing rejection of hypothesis 2.

Test 3: Comparison of variation in measurements obtained from simple (*S. aluco* femur) and complex (*M. erminea* skull) elements. The comparison in this test was restricted to elements that were measured with the same tool to ensure independence of samples. Equivocal significance was identified for the difference in variation between simple and complex elements in the interobserver experiment (Table 1; $P < 0.014$ for two and $P = 1$ for two in a set of four paired comparisons with Bonferroni correction). The significant results for difference in variation between simple and complex elements appear to be dependent on instruction use in the interobserver experiment. To reduce the influence of other variables, an intraobserver comparison was also performed. The outcome of the more tightly constrained intraobserver experiment was also equivocal (Table 1; $P = 0.0037$ and $P = 1$ in a set of two paired comparisons with Bonferroni correction). Element complexity has equivocal influence on measurement variability, allowing no conclusions to be drawn regarding hypothesis 3.

Test 4: Comparison of variation in measurement of specimens of different scales. This test utilised the elements of similar form (skulls) to determine whether the scale of the element influences inherent variability in measurements recorded. No significant difference in variation was seen between interobserver measurements taken from specimens of different scales (Table 1; lowest level of support $P = 0.5215$, highest level of support $P = 0.0791$ after Bonferroni correction for four paired comparisons in the set). The potential for lack of independence arising from different tool usage at the scales being compared (callipers for the smaller *M. erminea* specimen and tape measure for the larger *C. l. familiaris* specimen) suggested an intraobserver test using a one tool (callipers) to measure both scales of specimen.

Intraobserver measurement variation showed indications of a significant difference at different scales (Table 1; $P = 0.2$ for one and $P < 0.035$ after Bonferroni correction for three of four comparisons in a set of four). Comparison of coefficients of variation for *M. erminea* and *C. l. familiaris* (Table 3) indicates that variation was relatively greater for larger specimens. Measurement variation was significantly influenced by the scale of specimen (measurements of larger specimens were associated with larger relative variation in most instances, after potential differences in tool use were disregarded), allowing tentative acceptance of hypothesis 4.

Test 5a: Comparison of measurements obtained using different tools. This test was based on a set of measurements performed repeatedly by a single operator (PV), using the two different tools; thus, it must be understood to examine only intraobserver error and variation. Different tools provided significantly different intraobserver measurements of the same specimen using the same landmarks ($T = 23.23$, $P = 4.44 \times 10^{-23}$ and $T = 5.10$, $P = 8.52 \times 10^{-6}$ after Bonferroni correction for a set of two), with the tape measure giving shorter lengths than callipers in both instances. Different tools produced significantly different measurements, allowing acceptance of hypothesis 5a.

Test 5b: Comparison of variation associated with different tools. Equivocal significance was identified for the difference in variation between intraobserver measurements made by different tools ($P = 0.0054$ for one and $P = 1$ after Bonferroni correction for a set of two). Tool used has equivocal influence on measurement variability, allowing no conclusions to be drawn regarding hypothesis 5b.

Test 6a: Comparison of measurements obtained using different landmarks. As in test 5, the data from this test were drawn from a set of repeated observations of a single operator (PV). Different landmarks yielded significantly different intraobserver measurements of the same specimen using the same tool (Table 1; $T = 216.15$, $P = 2.55 \times 10^{-70}$; $T = 107.15$, $P = 3.31 \times 10^{-65}$ and $T = -31.63$, $P = 1.76 \times 10^{-33}$ after Bonferroni correction for a set of three). The mean *C. l. familiaris* L1 measurement was > 10 mm larger than the average *C. l. familiaris* L2 measurement (Table 3), around 7% of mean specimen length, regardless of the tool used. The mean *M. erminea* L1 measurement was 0.25 mm shorter than the *M. erminea* L2 measurement, around 0.5% of mean specimen length. The difference in magnitude of this difference is discussed in the following sections. Different landmark use results in significantly different measurements, allowing acceptance of hypothesis 6a.

Test 6b: Comparison of variation associated with different landmarks. No significant difference in variation was seen between intraobserver measurements using different landmarks ($P > 0.05$ after Bonferroni correction for set of three). Different landmark use does not result in significantly different measurement variation, allowing acceptance of hypothesis 6b.

Discussion

Our hypothesis that interobserver variation will be significantly larger than intraobserver variation (*contra* Lee and Mill 2004), found unequivocal support. The interobserver disparity of measurements was large, averaging 8.2% with a maximum of over 14% (test 1a). To put this level of disparity into context, the best univariate skeletal indicators of size dimorphism between male and female *Larus californicus* differ by an average of 8.03% (see Table 1 of Schnell et al. 1985), and discrepancy of length measurements taken from fossil stegocephalian specimens—including composites, partially incomplete skulls and reconstructions—was less than 10% and generally less than 5% (Laurin 2004). Coefficients of variation for a single *C. l. familiaris* skull length were as large as 3.05 (Table 2), broadly comparable with the coefficient of variation of 3.31 from a sample of skull length measurements from 36 different specimens of *Canis dirus* (Van Valkenburgh and Sacco 2002). In this context, data compiled from the results of multiple workers is of interest, since low levels of agreement between observers when measuring a single specimen suggests that compilations of data from different workers can contain considerable noise (arising from disparity between observers), which may obscure small but potentially significant trends within the dataset (depending upon the distribution of observations within the dataset). How this noise might arise and how it may influence results is worthy of further consideration.

We suggest that much of the disparity between measurements will arise due to the compounding of a few methodological inconsistencies between observers, for example, the interpretation of distance between landmarks. Although no significant difference in variation occurs when using different landmarks (test 6b), there is a significant difference in the result obtained (test 6a) and hence potential disparity between observers. Here, two different landmark-based intraobserver metrics for a *C. l. familiaris* skull length resulted in a >7% difference in the mean length reported for a single skull, a value consistent with the magnitude of disparity obtained in the interobserver experiments. If different individuals inadvertently used different landmarks to obtain what they believe to be the same measurement (e.g. skull length), they would report significantly different measurements, leading to increased disparity within the sample.

Significant disparity between measurements can also arise from use of different tools (test 5a) despite there being no significant difference in variation between the different tools used (test 5b). In this case, it may be because callipers can physically traverse points directly, whereas a tape measure must either deform over surfaces, or be held parallel to the vector of the measurement. Misalignment within the jaws of callipers leads to under-measurement, which is seen as a negative skew of unexpectedly small values for interobserver measurements (Table 2). With the measuring tape, the scope for similar error is non-directional, since the distance between points is assessed on the basis of a projection; otherwise, the straight-line distance is distorted by intervening topography. In this instance, the Euclidean distance can only be found from a projection by considering height differential and calculating the hypotenuse.

Methods of determining Euclidean distance can also be influenced by the relative orientation of the specimen to the measuring equipment, or by the translation from three to two-dimensions when measuring from photographs or projections (Marcus et al. 2000; Pierce et al. 2008). Some authors rightly attempt to constrain the angle relative to which measurements are taken (Christiansen and Adolfssen 2005; Van Valkenburgh 1990), but this subsequently raises the question of how "horizontal" (as used by Christiansen and Adolfssen 2005) or other baseline orientations are determined on a skull. Such variation in methods between workers could result in non-standard (and thus incomparable) metrics being reported.

Practical considerations like these may be less influential when dealing with smaller specimens (small in the context of this study being around 50 mm), hence the smaller variation associated with smaller specimens (test 4). It is worth noting that very small specimens (~5 mm), such as teeth, have increasing error associated with decreasing size (Polly 1998), possibly due to different sets of practical considerations arising (for example, the relative size of the specimens to the precision of the measuring equipment, distortion resulting from optical effects in photography, measuring from an inexact plane of focus when using a microscope, errors associated with measuring from a granulated reticule, etc.). Single-element topographically simple structures, such as long bones, may also offer fewer practical problems to address when taking a measurement. This may go some way to explaining the equivocal result obtained for test 3, and why the smallest variation in measurement was for the *S. aluco* femur in both inter- and intraobserver experiments (Tables 2, 3).

Although one could expect instructions would contribute significantly to addressing the problems discussed previously, they do not seem to do so (test 2). This result was unexpected, since it is intuitive that clearly defined landmarks should improve measurement precision and accuracy (the ability to identify the true measurement). However, here the use of diagrammatic instructions (mathematical and pseudo-landmarks) did not improve interobserver measurement precision. In fact, uncorrected S–F comparison of coefficient of variation in *S. aluco* femur measurements taken with or without instructions suggested that the diagrammatic instructions may have actually increased interobserver variation, whilst results for coefficient of variation and disparity were generally greater for the "with instructions" sample (see Table 2 and Fig. 3). This may be an artefact of our methodology; for example, our generalised diagram of a long bone was based on a humerus, but the element measured was a femur. Summary measurement data for *S. aluco* femur with instructions are consistent with the results for *M. erminea* (Table 2), so it is plausible that ambiguity in the landmark diagram is a sufficient explanation for this increase in variation, but further work is required to identify why instructions did not reduce variation.

Finally, when discussing how noise may arise, we need to consider that the number of contributors to a dataset will have a direct influence on measurement disparity (test 1a, Fig. 3). Examples of composite datasets from the literature provide us with numbers of observers contributing to published datasets. Conservative counts of the number of observers in studies using data compiled from the literature (where multiple papers that share one author in common are considered to represent one independent contribution, regardless of co-authorship) include six by Clegg and Owens (2002); many of which were themselves edited works compiled from multiple workers; six by Nudds et al. (2004); seven by Holtz (1994); ten by Turner et al. (2007); 19 by Laurin (2004); and 35 by Ashton et al. (2000). These counts suggest that some studies using composite data are likely to have high levels of noise introduced by disparity between multiple contributors, but at the same time, many have small enough numbers to be less likely to be substantially influenced.

Vulnerabilities

The influence of noise will of course depend on what the data are being used for and how the observations are arranged. Studies addressing questions by exploring variability of absolute sizes, such as sexual size dimorphism (e.g. Schnell et al. 1985), identification of specimens from comparative measurements (e.g. Melisch and Reitshel 1996: see Bochenski and Tomek 1995; Bochenski 2008 for discussion), geographically variation (e.g. Ashton et al. 2000; Clegg and Owens 2002), size variation through time (e.g. Laurin 2004), comparative biomechanics (Holtz 1994) and mass prediction from univariate length data (Farlow et al. 2005; Turner et al. 2007), are all potentially at risk of obscuring genuine signals, or identifying false ones, if interobserver disparity is ignored. This is because true biological signals may be indistinguishable from interobserver disparity in the dataset and any statistical or comparative analyses would unfortunately fail to test what is actually meant to be tested.

In particular, arrangement of interobserver data could bring about Type I error if compiled in discrete sections within the dataset that are related to the question at hand. To elucidate, a large compiled dataset encompassing a number of taxonomic groups is likely to have particular specialist observers contributing to a particular group; for example, Holtz (1994) has a majority of measurements for Mammalia obtained from McMahon; Aves from Archey, Monnier or Stritling and Zeitz; and Dinosauria by Holtz, meaning that interobserver disparity could potentially influence taxonomic signal. For instance, if one observer is systematically biased in measurements (i.e. they either over- or under-measure), then that observer's taxonomic group would exhibit persistent bias over other contributing observers and may introduce false signals in relation to that taxonomic group. Data compiled from workers in different geographical (or temporal) locations may also result in non-random distribution of observations within a dataset; for example, Clegg and Owens (2002) compiled data from region-specific studies carried out by workers specialising in those regions, potentially enabling interobserver disparity to influence geographical signal. Similar influence could potentially occur where different tools are available at different study locations (such as in Tiwari and Bjorndal 2000), even if used by the same observer—for example Creighton (1980) uses an Anderson Craniometer or a vernier-equipped Zeiss dissecting scope and dial callipers based on what was available at the different collections visited, but no calibration or testing for difference between result is mentioned. Should these data be used in subsequent compilation datasets, their rearrangement could increase the risk of error arising from any disparity that may have occurred and gone unnoticed.

Conclusions

The potential problems of combining data from independent sources, including material of different scales, measurements taken with different tools or with taxonomic and/or geographical clumping by observer, are often not explicitly addressed in the literature (Holtz 1994; Ashton et al. 2000; Clegg and Owens 2002; Turner et al. 2007). Further experimental work could examine the specific predictions we have proposed, particularly the expected increase in total disparity with added operators, and the difference between inter- and intraobserver variation in larger-scale studies of a limited set of elements.

This study suggests that meta-analyses using composite datasets should acknowledge the *caveat utilitor* that data may not be sensitive to genuine levels of variation, particularly when considering specimens that are represented by a diverse range of sizes and especially where methods used to obtain the data are known to differ or are not clearly reported. Moreover, the arrangement of data may be sensitive to bias imposed by disparity between observers. As communication about, access to and analysis of large complex datasets becomes more available and more desirable, it is imperative to explore methods of ensuring future compatibility between comparative anatomical studies. Further research is needed to address the lack of consistency in acquiring morphological measurements and the implications this may have for meta-analyses.

Acknowledgments We gratefully acknowledge the unidentified reviewer whose comments have substantially improved the structure and content of this paper. We especially thank Dr. Gareth Dyke and the others involved in the conference organisation of the 56th Symposium of Vertebrate Palaeontology and Comparative Anatomy (Dublin, 2008) for facilitating the experiment. We are grateful to all of the SVPCA attendees for constructive discussion and the 51 individual delegates who contributed their measurements. Additional thanks goes to Beulah Garner for her comments on this manuscript.

References

Adams DC, Rohlf FJ, Slice DE (2004) Geometric morphometrics: ten years of progress following the 'revolution'. Ital J Zool 71:5–16

Ashton KG, Tracy MC, de Queiroz A (2000) Is Bergmann's rule valid for mammals? Am Nat 156(4):390–415

Bailey SE, Pilbrow VC, Wood BA (2004) Interobserver error involved in independent attempts to measure cusp base areas of *Pan* M^1s. J Anat 205:323–331

Bochenski ZM (2008) Identification of skeletal remains of closely related species: the pitfalls and solutions. J Archeol Sci 35:1247–1250

Bochenski ZM, Tomek T (1995) How many comparative skeletons do we need to identify a bird bone? Courier Forschunginstitut Senckrnberg 181:357–361

Chen P-J, Dong Z-M, Zhen S-N (1998) An exceptionally well-preserved theropod dinosaur from the Yixian Formation of China. Nat 391:147–152

Christiansen P, Adolfssen JS (2005) Bite force canine strength and skull allometry in carnivores (Mammalia Carnivora). J Zool Lon 266:133–151

Clegg SM, Owens IPF (2002) The 'island rule' in birds: medium body size and its ecological explanation. P Roy Soc Lond B Bio 269(1498):1359–1365

Conover WJ, Johnson ME, Johnson MM (1981) A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics 23(4): 351–361

Constantinescu GM (2002) Clinical anatomy for small animal practitioners. Iowa State Press, Blackwell Publishing Company, Iowa

Creighton GK (1980) Static allometry of mammalian teeth and the correlation of tooth size and body size in contemporary mammals. J Zoo 191:435–443

Currie PJ, Zhao X-J (1993) A new carnosaur (Dinosauria Theropoda) from the Jurassic of Xinjiang People's Republic of China. Can J Earth Sci 30(10):2037–2081

Delany MJ, Healy MJR (1965) Variation in the white-toothed shrews (*Crocidura* spp) in the British Isles. Proc Roy Soc Lond B Bio 164(994):63–74

Dryden IL, Mardia KV (1998) Statistical analysis of shape. Wiley, London

Erickson GM, Lappin AK, Vliet KA (2003) The ontogeny of bite-force performance in American alligator (*Alligator mississippiensis*). J Zool 260:317–327

Farlow JO, Hurlburt GR, Elsey RM, Britton ARC, Langston W (2005) Femoral dimensions and body size of *Alligator*

mississippiensis: estimating the size of extinct mesoeucrocodylians. J Vertebr Paleontol 25(2):354–369
Fligner MA, Killeen TJ (1976) Distribution-free two-sample tests for scale. J Am Stat Assoc 71(353):210–213
Freeman S, Jackson WM (1990) Univariate metrics are not adequate to measure avian body size. Auk 107(1):69–74
Hammer Ø, Harper DAT, Ryan PD (2001) PAST: paleontological statistics software package for education and data analysis. Palaeontol Electron 4(1):9. http://palaeo-electronicaorg/2001_1/past/issue1_01htm. Accessed 6 Mar 2009
Herrel A, de Grauw E, Lemos-Espinal JA (2001) Head shape and bite performance in xenosaurid lizards. J Exp Zool 290:101–107
Holtz T (1994) The arctometatarsalian pes, an unusual structure of the metatarsus of Cretaceous Theropoda (Dinosauria: Saurischia). J Vertebr Paleontol 14(4):480–519
Kieser JA, Groeneveld HT (1987) Static intraspecific allometry of jaws and teeth in *Cercopithecus aethiops*. J Zool 212:499–510
Laurin M (2004) The evolution of body size, Cope's rule and the origin of Amniotes. Syst Biol 53(4):594–622
Lee S, Mill PJ (2004) Cranial variation in British mustelids. J Morphol 260(1):57–64
Lynch JM, Whelan R, Il Fituri AI, Hayden TJ (1997) Craniometric variation in the Eurasian badger. J Zool 242:31–44
Marcus LF, Hingst-Zaher E, Zaher H (2000) Application of landmark morphometrics to skulls representing the orders of living mammals. Hystrix 11(1):27–47
Melisch R, Reitshel M (1996) The Eurasian otter *Lutra lutra* in Afghanistan. Bonn Zool Beitr 46:367–375
Meyers JJ, Herrel A, Birch J (2002) Scaling of morphology bite force and feeding kinematics in an iguanian and scleroglossan lizard. In: Aerts P, D'Aout K, Herrel A, Van Damme R (eds) Topics in functional ecological vertebrate morphology. Shaker Publishing, Maastricht, pp 47–62
Nudds RL, Dyke GJ, Rayner JMV (2004) Forelimb proportions and the evolutionary radiation of Neornithes. Proc Roy Soc Lond B Bio 271(5):S324–S327
Pierce SE, Angielczyk KD, Rayfield EJ (2008) Patterns of morphospace occupation and mechanical performance in extant crocodilian skulls: a combined geometric morphometric and finite element modeling approach. J Morphol 269(7):840–864
Rohlf FR (1990) Morphometrics. Annu Rev Ecol Syst 21:299–316
Schnell GD, Worthen GL, Douglas ME (1985) Morphometric assessment of sexual dimorphism in skeletal elements of California gulls. Condor 87(4):484–493
Sues HD (1977) The skull of *Velociraptor mongoliensis* a small Cretaceous theropod dinosaur from Mongolia. Paläontol Z 51:173–184
Thomason JJ (1991) Cranial strength in relation to estimated biting forces in some mammals. Can J Zool 69:2326–2333
Tiwari M, Bjorndal KA (2000) Variation in morphology and reproduction in loggerheads, *Caretta caretta*, nesting in the United States, Brazil, and Greece. Herpetologica 56(3):343–356
Turner AH, Pol D, Clarke JA, Erickson GM, Norell MA (2007) A basal dromaeosaurid and size evolution preceding avian flight. Science 317:1378–1381
Van Valkenburgh B (1990) Skeletal and dental predictors of body mass in carnivores. In: Damuth J, MacFadden BJ (eds) Body size in mammalian paleobiology. Cambridge University Press, Cambridge, pp 181–205
Van Valkenburgh B, Sacco T (2002) Sexual dimorphism, social behaviour, and intrasexual competition in large Pleistocene carnivorans. J Vertebr Paleontol 22(1):164–169
Wiig Ø (1985) Multivariate variation in feral American mink (*Mustela vison*) from Southern Norway. J Zool 206(3):441–452
Wiig Ø (1986) Sexual dimorphism in the skull of minks *Mustela vison* badgers *Meles meles* and otters *Lutra lutra*. Zool J Linn Soc 87(2):163–179
Wroe S, McHenry C, Thomason J (2005) Bite club: comparative bite force in big biting mammals and the prediction of predatory behaviour in fossil taxa. Proc Roy Soc Lond B Bio 272:619–625
Yates AM (2007) *Anchisaurus polyzelus* (Hitchcock): the smallest known sauropod dinosaur and the evolution of gigantism among sauropodomorph dinosaurs. Postilla 230:1–57