Ann. Zool. Fennici 47: 323–334 ISSN 0003-455X (print), ISSN 1797-2450 (online)
Helsinki 29 October 2010

Within- and among-observer variation in measurements of animal biometrics and their influence on accurate quantification of common biometric-based condition indices

Anne E. Goodenough[1,*], Richard Stafford[1], Christina L. Catlin-Groves[1], Angela L. Smith[2] & Adam G. Hart[1]

[1)] *Department of Natural and Social Sciences, University of Gloucestershire, Cheltenham, GL50 4AZ, UK (*corresponding author's e-mail: aegoodenough@glos.ac.uk)*
[2)] *Gloucester Museums Service, Gloucester City Museum & Art Gallery, Gloucester, GL1 1HP, UK*

Received 1 Apr. 2010, revised version received 27 May 2010, accepted 3 July 2010

Goodenough, A. E., Stafford, R., Catlin-Groves, C. L., Smith, A. L. & Hart, A. G. 2010: Within- and among-observer variation in measurements of animal biometrics and their influence on accurate quantification of common biometric-based condition indices. — *Ann. Zool. Fennici* 47: 323–334.

Research using biometric data relies on consistent measurements within, and often among, observers. However, research into the relative importance of intra- and inter-observer variability is limited. More importantly, the influence of biometric variability on accurate quantification of biometric-based condition indices has not been analysed: it is unclear whether multiple errors become magnified or cancel one another out. Here, we quantify intra- and inter-specific variability in multiple biometrics, and derived condition indices, using museum bird specimens. Inter-observer variability was higher than intra-observer variability for all parameters. Measurement error (ME) varied from < 1% to > 50% for different biometrics. ME was magnified in condition estimates, reaching > 80% within-observers and > 90% among-observers. Significant differences in mean measurements were found for 17% and 67% of biometrics within- and among-observers, respectively; for condition indices, the figures were 50% and 67%, respectively. We discuss the implications of these findings for research into species' ecology, taxonomy and behaviour.

Introduction

Many ecological studies rely on accurate biometric data. Measurements of physical traits often form the basis of research into taxonomic and phylogenetic relationships (e.g. Corti *et al*. 1988, Smith *et al*. 2004), life history traits such as growth and sexual size dimorphism (Hunt & Hunt 1976, Weckerly 1998) and trait heritably (Alatalo *et al*. 1990). When combined with other information, biometrics are also used to test evolutionary concepts [e.g. Cope's rule (Kingsolver & Pfennig 2004)], examine biogeographical patterns [e.g. Bergmann's rule of body size (Ashton 2002)] and explain social interactions such as dominance and aggression (Searcy 1979, Barrette & Vandal 1990).

In addition to use of primary biometric data, measurements of physical traits are frequently

combined to quantify the condition of individuals, often as a proxy for fitness. These estimates of condition, including trait asymmetry, can be used to determine: (1) cues used for mate selection (Møller 1992, Schlüter *et al*. 1998); (2) costs or benefits conferred by specific behaviours such as nest-site selection (Goodenough *et al*. 2008); and (3) non-lethal consequences of parasitism and disease (Merino & Potti 1995). Such biometric-derived condition indices are commonly used in cross-taxonomic studies (e.g. Mousseau & Roff 1987), as well as species-specific research on vertebrates and invertebrates (e.g. Krebs & Singleton 1993, Hogg *et al*. 1995, Grieco 2003).

The possibility that intra- and inter-observer variability in recording biometrics can generate a significant source of measurement error is not new. For example, Evans (1964) and Nisbet *et al*. (1970) showed inter-observer variation in the measurement of bird wing length, while Pankakoski *et al*. (1987) and Palmeirim (1998) showed that mammal biometrics are also subject to measurement error. Variability can be reduced by using standardised methods (Evans 1964, Arendt & Faaborg 1989), however the potential for residual variability still remains — for example, linked to whether the recorder is right or left handed (Helm & Albrecht 2000). Residual variation, both within and among observers, is a particular concern when biometrics are used to assess condition. However, despite Krebs and Singleton (1993) asserting that observer-based variability must be eliminated for biometric-based condition indices to be worthwhile, there has been little quantification of variation in biometric measurements themselves other than for wing length (Evans 1964, Nisbet *et al*. 1970, Helm & Albrecht 2000) and skeletal parameters (Pankakoski *et al*. 1987, Yezerinac *et al*. 1992, Palmeirim 1998). Perhaps more importantly, there has been no analysis of the influence of variability in biometric measurements on biometric-based indices of condition, such that it is uncertain whether numerous small errors in multiple biometric parameters become magnified (a big effect on condition estimates), or simply cancel each other out during calculations (a minimal effect). This is surprising given the controversy on the most appropriate way of estimating condition on the basis of biometrics. For example, despite several key papers on the statistical validity of quantifying condition on the basis of the relationship between size and weight (Jakob *et al*. 1996, García-Berthou 2001, Green 2001, Schulte-Hostedde *et al*. 2005), the potential role of intra- and inter-observer measurement variability has not seemingly been considered. The main reason for the lack of such research is likely to be the difficulty in undertaking repeated measurements, both within and among observers, in the field. This is mainly because the time spent handling live individuals must be minimised to avoid undue stress, hypothermia or injury (Redfern & Clark 2001); a situation not conducive to studies of variability. Moreover, in the field, repeated measurements by the same observer either have to be taken in quick succession within one recording session, allowing measurements to be remembered, or several days later when biometrics, especially weight, could be subject to natural change (Arendt & Faaborg 1989, Krebs & Singleton 1993).

In this study, we use museum bird specimens to obtain repeated biometric measurements within and among observers (i.e. repeatability and reproducibility: Gosler 2004). Our aims are two-fold. Firstly, we quantify both intra- and inter-observer variability of initial biometrics to ascertain their relative importance. This is apparently the first attempt to quantify relative importance by simultaneously considering both types of variability, and any interactions between them, in a single study. Secondly, we calculate six different condition indices based on the initial biometrics to establish whether errors become magnified or nullified during calculation. Again we determine the relative importance of intra-observer and inter-observer variation for each index, as well as whether variability is related to species size. This is seemingly the first time that the potential impact of biometric measurement errors on the precision of commonly-used condition measures has been analysed, for any taxonomic group.

Methods

Measuring avian biometrics

Twenty five individual adult birds, each from a different species, were selected from the natural

history collection at Gloucester City Museum and Art Gallery (Gloucestershire, UK). These birds had been prepared for display using standard taxidermy methods (Hangay & Dingley, 1986) and ranged in size from a mealy redpoll *Carduelis flammea* (total length = 12 cm) to a goshawk *Accipiter gentilis* (total length = 55 cm). Six biometric measurements were taken of each bird according to standard protocols and using standard equipment (Table 1). It should be noted that the weights recorded were not, in themselves, meaningful, since the weight of a museum specimen will differ from the weight of a live individual. However, since we were only interested in the variability of weight measurements, and of condition assessments that use weight along with other biometrics, this was justified. Likewise, any minor differences in characteristics such as wing length *post mortem* due to shrinkage (Evans 1964, Ewins 1985) would not confound or invalidate analyses.

To determine inter-observer variation in measurement, the biometrics of each bird were recorded by eight observers (each of the five authors and three volunteers listed in the acknowledgements). All observers were experienced in taking biometrics. To avoid handedness affecting results (Helm & Albrecht 2000), each recorder was right handed. Measurements were undertaken using a blind protocol, whereby each observer recorded a set of biometrics on a form, which was then placed in a ballot box. This ensured that each set of measurements was independent of, and unbiased by, other recorder's measurements. Then, to determine intra-observer variation across recording sessions, each observer measured each bird twice more, to give a total of three sets of records from each observer for each bird (Lougheed *et al*. 1991, Yezerinac *et al*. 1992, Helm & Albrecht 2000). The order in which birds were measured during the recording sessions was randomised to avoid familiarity becoming a confounding factor. Because of the ballot box system and the randomised recording, observers were not able to check their previous measurements and would have been unlikely to remember the measurement of a specific parameter for a specific bird across the separate recording sessions. The total number of measurements was 3600 (25 birds × 6 biometrics × 8 observers × 3 attempts per observer).

Quantifying bird condition from biometric measurements

Three different types of condition index were calculated from the biometric data. Firstly, *Q*-values were calculated for each species by dividing a size variable by weight (Gosler 2004). A simple *Q*-value was calculated using a univariate measure of size (right wing length — the best single measure of body size: Gosler *et al*. 1998), while a more complex *Q*-value was calculated using a multivariate measure of body size generated using a Principal Components Analysis on the whole suite of size biometrics (Rising & Somers

Table 1. Methods and equipment used to measure biometrics and precision of measurements taken.

Trait	Equipment	Method	Precision
Wing length (both wings)	150 mm or 300 mm stopped wing rule (NHBS Equipment, Devon, UK)	Flattened-straightened wing method: the distance from the carpel joint to the tip of the longest primary wing feather (Svensson 1992)	1 mm
Tarsus length (both tarsi)	Vernier callipers (Mitutoyo model 351, Coventry, UK)	Minimum tarsus method: the distance between the notch of the tarsal joint and the foot joint (Gosler 2004)	0.1 mm
Bill length	As above	Total bill length: the distance from the naso-frontal hinge to the dertrum (Gosler 2004)	0.1 mm
Weight	30 g, 60 g or 300 g spring balance, as appropriate (Pesola®, Switzerland)	Each bird was clipped to the balance using its museum accession tag	1 g

1989). In our dataset, the first principal component, PC1, explained 72.9% of overall size. Secondly, weight was regressed against size to provide a series of standardised regression residuals (strong positive scores = good condition, strong negative scores = poor condition) (Jakob *et al*. 1996). Again, a simple (univariate) version of the index was created using right wing length and a more complex (multivariate) version was created using PC1 as above. Thirdly, condition was assessed using a fluctuating asymmetry (FA) approach, which utilises the difference in size of bilateral traits (those found on both sides of one individual bird) as a proxy for condition (Parsons 1992, Björklund 1996, Møller 1997). Both tarsus and wing asymmetries were quantified. In both cases, the absolute difference between the sides was quantified and this was then divided by the mean of the two measurements to give an asymmetry index that was related to trait length. This FA index, which has trait-size correction at an individual level, is known as FA2 (Palmer 1994, Palmer & Strobeck 2003).

Statistical analyses

Baseline variability

To examine the relative variability of different biometrics, the coefficient of variation (CV) was calculated and expressed as a percentage [CV = (standard deviation/mean) × 100]. This relative approach has been used in previous studies of biometric variability (Pankakoski *et al*. 1987) as it allows variability to be compared directly, even when mean trait size differs significantly (Fowler & Cohan 1996). To quantify intra-observer variation, a CV value was calculated using each of the three separate measurements of each trait of each bird by the same individual (resulting in six trait-specific values per bird, per observer). To quantify inter-observer variation, a CV value was calculated using the mean measurement of each trait of each bird by the eight different observers (resulting in six trait-specific values per bird). The same approach was used to assess variability of the condition indices calculated using the biometrics data. The mean (± SE) CV value for each trait was calculated to assess which trait or condition index was least subject to intra- and inter-observer variability.

To determine whether measurement variability was related to the size of the bird, trait-specific CV values were regressed against the size of that trait, while condition-specific CV values were regressed against the overall size of the bird (PC1), in a series of regression analyses. To determine any directional bias in the precision with which measurements of bilateral traits (wing and tarsi) could be taken, CV values were compared on a per-observer basis using an individual non-parametric Wilcoxon sign-rank test. Finally, to establish whether inter-observer variability differed according to whether repeated measurements from each single observer were summarised using the mean or the median as the measure of central tendency, Levene's test was calculated.

Measurement error

Percentage measurement errors (%ME) were calculated for each biometric and condition index. This details the relative amount of variation in a parameter that is due to measurement error rather than "true" biological variation, based on the fact that repeated measurements (within or among observers) should yield the same result. This method has been used in previous studies of biometrics (e.g. Lougheed *et al*. 1991, Yezerinac *et al*. 1992) and was applied as per Bailey and Byrnes (1990), using within- and among-bird components of variance (i.e. the amount of variance derived from measurement variability and biological variability, respectively). Measurement error was calculated from these parameters as follows:

$$\%\mathrm{ME} = 100 - \frac{s^2_{\mathrm{within}}}{s^2_{\mathrm{within}} + s^2_{\mathrm{among}}} \times 100 \qquad (1)$$

Statistical differences resulting from intra- and inter-observer variability

To analyse the relative importance of intra- and inter-observer variability, individual two-way repeated measures ANOVAs (one for each trait or condition index) were calculated, as per

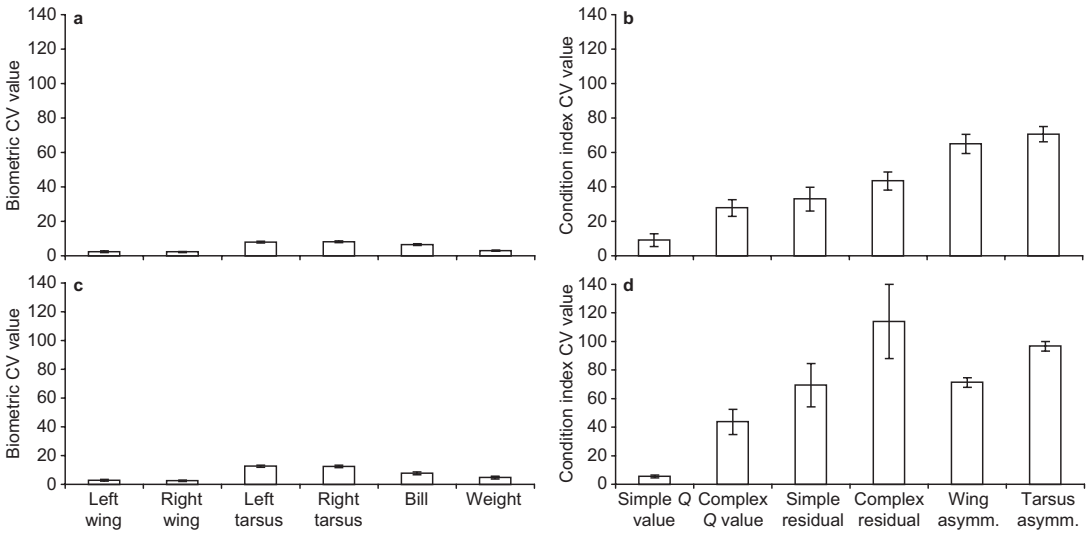

Fig. 1. Mean coefficient of variation (CV) values for measurements of biometrics showing (**a**) intra-observer variation and (**b**) inter-observer variation; and for condition indices based on the biometrics showing (**c**) intra-observer variation and (**d**) inter-observer variation. Error bars show standard error of the mean; $n = 25$ birds measured three times each by eight different observers.

Palmeirim (1998) but using the Greenhouse-Geisser method to compensate for sphericity. In all cases, 'observer' ($n = 8$) and 'attempt' ($n = 3$) were defined as fixed factors and the interaction term was calculated. These analyses allowed the importance of within-observer and among-observer variances to be quantified relative to variation among specimens, and allowed any interactions between these parameters to be quantified (a significant interaction being evidence of observers differing in their ability to take consistent measurements). To explore such interactions further, trends in precision of biometric and condition index values between the three recording sessions were calculated by comparing, on a per-bird basis, the deviation between each record that each observer made and the mean of all measurements, from all observers, for that bird. The mean deviation for each observer × attempt combination from the grand mean was then calculated for each biometric/index, with increasing precision being signified by a decrease in deviance from the grand mean between the recording sessions (and *vice versa*).

All statistics were calculated using SPSS 16 for Windows. To allow for multiple analyses being undertaken on non-independent data (different biometrics of the same bird), standard Bonferroni corrections were applied to significance values in the repeated measures ANOVA and regression analyses, whereby significance values were multiplied by six (as there were six related biometrics and biometric-based condition indices).

Results

Baseline variability

Coefficient of variation (CV) values indicated substantial variability in biometric measurements both between repeated measurements by the same observer (mean = 5.1%) and between measurements by different observers (mean = 7.1%). However, when condition indices were calculated based on these biometrics, the amount of variability within and among observers increased by almost an order of magnitude (mean = 41.5% and 66.8%, respectively), suggesting that numerous small errors in measurements are magnified during the calculation of biometric-based condition indices (Fig. 1).

Although there were differences in CV values for different individual birds, there were no significant relationships between trait- or con-

dition-specific CV values and bird size (quantified using trait size or PC1, as appropriate) either within or among observers (regression analysis: $F_{1,23} < 0.27$ and $P > 0.606$ in all cases; tests not shown). There was no difference in inter-observer variability when repeated measurements from individual observers were summarised using the median rather than the mean (Levene's test: $W_{1,398} < 0.050$ and $P > 0.824$ in all cases; tests not shown).

Measurement error

As suggested by the CV values, the percentage of variability in traits or indices that was accounted for by measurement error was often high (Table 2). Overall, %ME was lowest for intra-specific biometric measurements, higher for inter-specific biometric measurements, higher again for intra-specific condition estimates and highest of all for inter-specific condition estimates. However, there was substantial variation, in each of the aforementioned categories, among different parameters. For example, %ME in wing length was low (ca. 1%), while for tarsus measurements it was ca. 25% intra-specifically and ca. 50% inter-specifically. As regards condition estimates, %ME ranged from 4.6% to 87.5% intra-specifically and from 8.1% to 90.9% inter-specifically. In both cases, the simple *Q*-value was the least error-prone index while tarsus asymmetry was the most error-prone.

Table 2. Percentage of variability in different parameters accounted for by measurement error (%ME) rather than "true" biological differences. Values were determined using Eq. 1 following the ANOVA analysis (*see* Methods). "Repeatability" can be calculated from the figures given below by subtracting the %ME value from 100.

	Measurement error (%ME)	
	Intra-specific	Inter-specific
Biometrics		
Left wing	1.01	1.21
Right wing	0.92	0.97
Left tarsus	25.49	48.50
Right tarsus	25.01	51.07
Bill	8.18	6.34
Weight	3.07	5.34
Mean	12.53	22.44
Condition Indices		
Simple *Q*-value	4.62	8.05
Complex *Q*-value	18.03	36.20
Simple residual	11.52	20.63
Complex residual	20.18	55.49
Wing asymmetry	87.47	90.87
Tarsus asymmetry	84.00	89.66
Mean	37.64	50.15

The importance of imprecise measurements: statistical differences resulting from intra- and inter-observer variability

Given the high variability and %ME rates, it was not surprising that when data were analysed statistically using a repeated measures ANOVA, there were significant differences between measurements (Table 3). Four of the six biometrics differed significantly among observers, with only weight and right wing length being consistent. Biometrics were more consistent in repeated measurements by the same observer; the only two traits to differ significantly were left wing length and bill length. All condition indices were very variable (Fig. 1b and d); four indices differed significantly among observers (only the simple *Q*-value and the simple regression residual index — the only indices to use weight and right wing length alone — gave consistent results), while three indices (the complex regression residual and both asymmetry indices) differed significantly when compared intra-specifically.

There were significant interactions (Table 3) between the 'observer' and 'attempt' factors for three biometrics (left wing length, left tarsus length and right tarsus length), suggesting that observers differed in their ability to take consistent measurements. There was also a significant interaction between 'observer' and 'attempt' for two condition indices (complex *Q*-value index and complex regression residual index; both condition indices that used the three biometrics with significant interaction terms). Most of these interactions occurred because some individuals improved the precision with which they took particular biometric measurements as the study progressed (i.e. they were better in recording

session three than in recording session one) while others were consistent throughout (Table 4). In the case of the complex *Q*-value, precision of the estimates calculated using biometrics increased as the recording sessions progressed for two observers, decreased for two observers and remained consistent for the remaining four observers.

When intra-specific trait measurement was considered on an individual-level to expand on the interactions between intra- and inter-specific variability outlined above, it was clear that there were considerable differences in individual ability to record specific biometrics (Fig. 2). Most people (observers 3–8) had most difficulty taking repeatable tarsus measurements, but others (observers 1 and 2) had most difficulty obtaining consistent bill measurements. More strikingly, the person who was most consistent at measuring right wing length was the least consistent at recording bill length (observer 2), the person who was most consistent at recording weight was the least consistent at recording right tarsus length (observer 5), and the person who was most consistent at recording left wing length was the least consistent at recording weight (observer 8). As regards differences in measurements of bilateral traits, 63% of observers achieved more consistent right- than left-wing measurements, while tarsus length measurements did not differ (50% of observers were better on the left and the other 50% were better on the right). There was no significant directional bias in the measurement precision of bilateral traits when the magnitude (as well as directionality) of measurements was analysed on a per-observer basis using CV values (Wilcoxon sign-rank test: left and right wings $W+ = 9.50$, $W- = 26.50$, $n = 8$, $P = 0.250$; left and right tarsus $W+ = 20$, $W- = 16$, $n = 8$, $P = 0.844$, respectively).

Table 3. Fully-factorial, two-way, repeated measures ANOVA results for biometrics and condition indices. The Greenhouse-Geisser method was used to compensate for sphericity and Bonferroni corrections were applied to significance values to allow for family-wise error (significant *P*'s are set in boldface). The reason for significant interactions between observer and attempt was usually that some observers improved their ability to take precise measurements during the course of the study (i.e. between recording sessions) while others remained consistent (*see* Table 4).

	Biometrics				Indices			
ANOVA factor	Trait	*F*	df	*P*	Type	*F*	df	*P*
Observer	Left wing	7.198	2.417	< **0.001**	Simple *Q*-value	2.930	1.711	0.079
Attempt		3.193	1.435	0.060		1.621	3.106	0.196
Observer × attempt		3.294	3.083	**0.020**		1.328	3.925	0.270
Observer	Right wing	2.629	1.445	0.104	Complex *Q*-value	16.535	3.720	< **0.001**
Attempt		0.902	2.003	0.414		0.310	1.423	0.662
Observer × attempt		1.234	0.306	0.306		2.801	6.279	**0.004**
Observer	Left tarsus	22.596	4.294	< **0.001**	Simple residual	2.554	1.373	0.110
Attempt		0.908	1.710	0.399		1.097	2.806	0.355
Observer × attempt		3.427	5.695	0.004		2.713	1.718	0.086
Observer	Right tarsus	28.851	3.821	< **0.001**	Complex residual	4.113	2.584	**0.004**
Attempt		0.172	1.896	0.832		3.301	1.488	**0.046**
Observer × attempt		3.055	6.166	**0.008**		2.181	7.971	**0.004**
Observer	Bill	4.610	1.972	**0.016**	Wing asymmetry	6.135	3.142	**0.001**
Attempt		2.642	7.000	0.014		4.947	7.000	< **0.001**
Observer × attempt		0.531	5.680	0.775		0.683	4.641	0.627
Observer	Weight	1.506	0.238	0.238	Tarsus asymmetry	26.740	3.914	< **0.001**
Attempt		1.284	0.292	0.292		20.962	7.000	< **0.001**
Observer × attempt		0.811	0.524	0.524		0.613	6.624	0.735

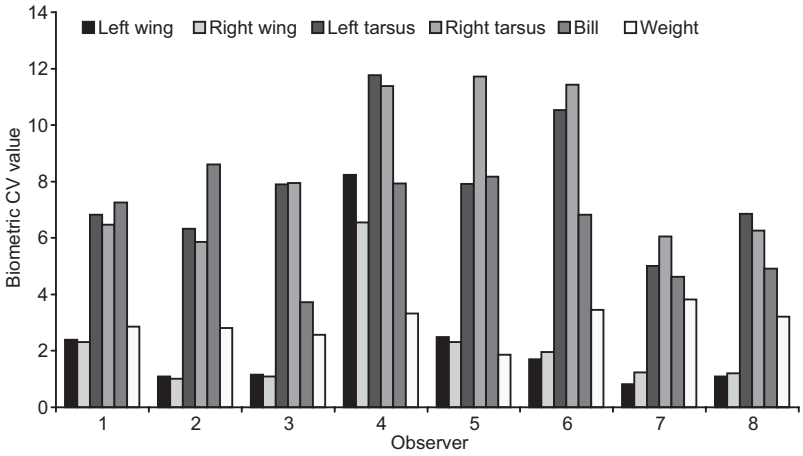

Fig. 2. Mean coefficient of variation (CV) values for individual biometrics of 25 birds measured 3 times each shown for each observer.

Discussion

General findings: baseline variability and measurement error

The validity of research conclusions is always dependent upon robust and reliable data that are appropriately analysed. Research using biometric data, therefore, relies on accurate measurements, which are consistent within and, where necessary, among observers (Krebs & Singleton 1993). However, this study suggests that both intra- and inter-observer variation can be substantial, and significant, sources of error. As expected, inter-observer variability was higher than intra-observer variability for all parameters. More importantly, our data suggest that numerous small errors in multiple biometric parameters do not simply cancel one another out during calculations of condition indices (as they do when PC1 is calculated using PCA: Lougheed *et al*. 1991), but rather become magnified, such that they could have a important effect on condition estimates. Again, inter-specific variation is higher than intra-specific variation for these condition indices. This is apparently the first time that this has been quantified, and as such it has important implications for research that uses condition indices. As suggested by the high levels of variability, the relative percentage of trait or condition-index variability that is the result of measurement error, rather than "true" biological variability, is substantial. Measurement errors are, in general, higher for condition indices than for straight biometric data and higher among observers than between repeated measurements by a single observer (Table 2). It should be noted that as our measurements were taken on museum specimens also means that the variability and measurement error values quantified here are likely to be conservative as observ-

Table 4. Trends in precision of biometric and condition index values between the three recording sessions. Values were calculated by comparing, on a per-bird basis, the deviation between each record that each observer made and the mean of all measurements, from all observers, for that bird. The mean deviation for each observer × attempt combination for the mean was then calculated for each biometric/index, with increasing precision being signified by a decrease in deviance from the overall mean between the recording sessions. Significant interactions between observer × attempt, as quantified in Table 3, are shown by an asterisks.

	Precision increased	Precision decreased	Precision consistent
Biometrics			
Left wing*	5	1	2
Right wing	0	1	7
Left tarsus*	3	0	5
Right tarsus*	3	0	5
Bill	0	0	8
Weight	0	1	7
Condition indices			
Simple *Q*-value	1	0	7
Complex *Q*-value*	4	0	4
Simple residual	2	0	6
Complex residual*	2	2	4
Wing asymmetry	1	1	6
Tarsus asymmetry	1	1	6

ers did not have to measure and restrain the birds simultaneously.

As there were no significant relationships between bird size and either intra- or inter-observer variability in biometrics and associated condition indices, we conclude that relative variability (as quantified using CV values to ensure that variability is not scale-dependent) is independent of size. This differs from some previous studies (e.g. Pankakoski *et al*. 1987, Yezerinac *et al*. 1992, Palmeirim 1998), which found significantly higher relative variability in small traits in skeletal parameters, but agrees with others (e.g. Lougheed *et al*. 1991), which found no such relationship. The similarity of measurement of left and right wings is not surprising since observers were universally right handed (Helm & Albrecht 2000), while the lack of a directional difference in measurement precision of tarsus length was expected since the bird is held in the same relative attitude for both measurements (Goodenough *et al*. 2008).

Specific patterns

Our analyses show a consistent pattern — generally traits with high CV values have high %ME values and differ significantly ($P < 0.05$) among/within observers. Like Lougheed *et al*. (1991) and Palmeirim *et al*. (1998), we found substantial differences in both intra- and inter-observer errors across different parameters. As regards biometrics, those with definite landmarks at both ends that were clear and unambiguous (e.g. wing length) had lower CV and %ME levels than traits that were more subjective — this has previously been noted for skeletal parameters (Palmeirim *et al*. 1998). As regards biometrics, right wing length and weight (both of which had low CV and comparatively low %ME values) were the only biometrics not to differ among different observers. Although some variability was noted in these parameters as per Nisbet *et al*. (1970), the lack of a significant difference among observers agrees with previous studies on a range of bird species (Arendt & Faaborg, 1989). Moreover, the fact that wing length is the measure with the least variability both intra- and inter-specifically agrees with Gosler *et al*. (1998) who found this to be the most consistent linear avian biometric within observers. As regards condition indices, the measurements based upon right wing length and weight alone (simple *Q*-value and simple residual) were the least variable, the least error-prone and the only indices not to differ significantly within or among observers. Importantly, the multivariate indices, which might often be thought of as superior given their increased complexity, were much more error-prone than their univariate equivalents.

The fact that observers differ in their ability to take different biometrics suggests that rather than certain observers being better than others across the whole suite of measurements, different observers have different strengths and weaknesses. This in itself is interesting and, apparently has not been previously documented. It is also interesting to note that some biometrics were taken with equal precision by all observers (e.g. bill) while for others, some observers were consistent throughout but others improved with increasing familiarisation (e.g. both tarsi measurements). On an individual basis, all observers had at least two biometrics that they were consistent in recording and two that they took with increasing precision as the study progressed. Occasionally, some measurements were taken with decreasing precision by some observers over the successive recording sessions (both wing lengths by the same single observer and bill by a different single observer), suggesting that familiarity can be disadvantageous in some cases. As regards the condition indices, the complex residual index decreased in precision as the study progressed for two observers, probably because recording of individual biometrics improved at different rates, with size biometrics being taking with increasingly greater precision but weight being recorded consistently, such that the relationship between weight and size, which the residual index is based upon, became less precise.

Implications

The high levels of variability and measurement error is concerning given that "significant" differences in biometrics between treatment groups

or "significant" relationships between biometrics and environmental variables are often based on data with a very restricted range, such that error rates need to be low in order for them not to become confounding. It is most likely that high variability would lead to an increased risk of Type I error as increased variability can decrease the chance of finding significant differences between groups. Of greater concern is the potential for high ME levels to cause an increase in Type II errors. This is only likely if there is a systematic bias in measurement, which is not suggested here but could occur in some studies (e.g. those using data from both right and left handed field workers), or autocorrelation between %ME and some other environmental variable within the study. Given the interaction between familiarisation/experience and measurement precision for some, but not all, observers in this study, it is possible that studies analysing temporal change in biometrics could be affected by a change in %ME (which would be difficult to factor out given that it appears to occur unequally, affecting some observers but not others and acting on a per-biometric basis).

Recommendations

Given that variability and measurement error are both lower in the intra-observer data than they are intra-specifically, we suggest that research should use biometric data collected by a single recorder whenever possible. When this is not possible, data should be checked using some repeated measures data to check that error rates are low enough not to confound analyses (e.g. Goodenough *et al*. 2008). Variables with high error rates should either be excluded from subsequent analyses (Palmeirim *et al*. 1998) or results based on analyses of such data should interpreted with caution. Alternatively, multiple repeat measurements can be averaged to reduce the effect of ME (Yezerinac *et al*. 1992). Our analyses do not suggest that there is any advantage in using median measurements over mean measurements. In order to reduce the risk of high ME rates at source, we recommend that biometrics based on clearly-identifiable and unambiguous landmarks — such as wing length for birds (this study) or the distance between the tip of the shell and the top of the aperture for snail shells (Bailey & Byrnes 1990) — are used. Given the magnification of error rates in the calculation of condition indices, the type of condition index used in a given study should be carefully considered. The least variable (i.e. most consistent within and among observers) condition indices, based on the data analysed here, are the simple *Q*-value and the simple residual index. We suggest that these two condition indices might be superior to other indices, although this does need to be tested in other taxonomic groups. It should also be noted that the six main assumptions underpinning use of regression residuals (Green 2001) should be tested fully as per Schulte-Hostedde *et al*. (2005) prior to this technique being used since low observer variability does not equate axiomatically to statistical validity. As fluctuating asymmetry appears to be particularly prone to error (ca. 90% of variability in tarsus and wing asymmetry measurements was due to ME), we concur with several other recent studies (e.g. Hogg *et al*. 1995) and suggest that FA might be inappropriate for quantifying biological condition, at least in some situations. We recommend that an FA approach should be used with extreme caution, particularly in studies that use measurements from multiple observers, and that any studies should use FA measures that account for %ME (such as the measurements FA10a or FA10b, which describe the average difference between sides after ME has been factored out using actual data (mm) or proportional data, respectively; Palmer 1994, Palmer & Strobeck 2003, Bechshøft *et al*. 2008).

Acknowledgements

We thank Claire Kirkhope, Samuel Rees and Rachel Williams for volunteering as recorders and taking biometric measurements for the purposes of this research and Gloucester City Museum and Art Gallery for allowing us access to specimens not on public display for measurement purposes. We also thank two reviewers for their detailed and constructive comments, which improved the final paper.

References

Alatalo, R. V., Gustafsson, L. & Lundberg, A. 1990: Pheno-

typic selection on heritable size traits: environmental variance and genetic response. — *American Naturalist* 135: 464–471.

Arendt, W. J. & Faaborg, J. 1989: Sources of variation in measurements of birds in a Puerto Rican dry forest. — *Journal of Field Ornithology* 60: 1–11.

Ashton, K. G. 2002: Patterns of within-species body size variation of birds: strong evidence for Bergmann's rule. — *Global Ecology & Biogeography* 11: 505–523.

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.

Barrette, C. & Vandal, D. 1990: Sparring, relative antler size, and assessment in male caribou. — *Behavioral Ecology and Sociobiology* 26: 383–387.

Bechshøft, T. Ø., Rigét, F. F., Wiig, Ø. & Sonne, C. 2008: Fluctuating asymmetry in metric traits; a practical example of calculating asymmetry, measurement error and repeatability. — *Annales Zoologici Fennici* 45: 32–38.

Björklund, M. 1996: The effect of male presence on nestling growth and fluctuating asymmetry in the blue tit. — *Condor* 98: 172–175.

Corti, M., Thorpe, R. S., Sola, L., Sbordoni, V. & Cataudella, S. 1988: Multivariate morphometrics in aquaculture: a case study of six stocks of the common carp (*Cyprinus carpio*) from Italy. — *Canadian Journal of Fisheries and Aquatic Sciences* 45: 1548–1554.

Evans, P. R. 1964: Wader measurements and wader migration. — *Bird Study* 11: 23–38.

Ewins, P. J. 1985: Variation of black guillemot wing lengths post-mortem and between measurers. — *Ringing and Migration* 6: 115–117.

Fowler, J. & Cohen, L. 1996: *Statistics for ornithologists*. — British Trust for Ornithology, Thetford.

García-Berthou, E. 2001: On the misuse of residuals in ecology: testing regression residuals *vs*. the analysis of covariance. — *Journal of Animal Ecology* 70: 708–711.

Goodenough, A. E., Hart, A. G. & Elliot, S. L. 2008: Variation in offspring quality with cavity orientation in the great tit. — *Ecology Ethology and Evolution* 20: 375–389.

Gosler, A. G. 2004: Birds in the hand. — In: Sutherland, W. J. (ed.) *Bird ecology and conservation: a handbook of techniques*: 85–118. Oxford University Press, Oxford.

Gosler, A. G., Greenwood, J. J. D., Baker, J. K. & Davidson, N. C. 1998: The field determination of body size and condition in passerines: a report to the British Ringing Committee. — *Bird Study* 45: 92–103.

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

Grieco, F. 2003: Greater food availability reduces tarsus asymmetry in nestling blue tits. — *Condor* 105: 599–603.

Hangay, G. & Dingley, M. 1986: *Biological museum methods: vertebrates*. — Academic Press, London.

Helm, B. & Albrecht, H. 2000: Human handedness causes directional asymmetry in avian wing length measurements. — *Animal Behaviour* 60: 899–902.

Hogg, I. D., Williams, D. D., Eadie, J. M. & Butt, S. A. 1995: The consequences of global warming for stream invertebrates: a field simulation. — *Journal of Thermal Biology* 20: 199–206.

Hunt, G. L. Jr. & Hunt, M. W. 1976: Gull chick survival: the significance of growth rates, timing of breeding and territory size. — *Ecology* 57: 62–75.

Jakob, E. M., Marshall, S. D. & Uetz, G. W. 1996: Estimating fitness: a comparison of body condition indices. — *Oikos* 77: 61–67.

Kingsolver, J. G. & Pfennig, D. W. 2004: Individual-level selection as a cause of Cope's rule of phyletic size increase. — *Evolution* 58: 1608–1612.

Krebs, C. J. & Singleton, G. R. 1993: Indices of condition for small mammals. — *Australian Journal of Zoology* 41: 317–323.

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.

Merino, S. & Potti, J. 1995: Mites and blowflies decrease growth and survival in nestling pied flycatchers. — *Oikos* 73: 95–103.

Møller, A. P. 1992: Female swallow preference for symmetrical male sexual ornaments. — *Nature* 357: 238–240.

Møller, A. P. 1997: Developmental stability and fitness: a review. — *American Naturalist* 149: 916–932.

Mousseau, T. A. & Roff, D. A. 1987: Natural selection and the heritability of fitness components. — *Heredity* 59: 181–197.

Nisbet, I. C., Baird, J., Howard, D. V. & Anderson, K. S. 1970: Statistical comparison of wing-length measured by four observers. — *Bird-Banding* 41: 307–308.

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 evolutionary implications*: 335–364. Kluwer, Dordrecht.

Palmer, A. R. & Strobeck, C. 2003: Fluctuating asymmetry analyses revisited. — In: Polak, M. (ed.) *Developmental instability (DI): causes and consequences*: 279–319. Oxford University Press, Oxford.

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

Parsons, P. A. 1992: Fluctuating asymmetry: a biological monitor of environmental and genomic stress. — *Heredity* 68: 361–364.

Redfern, C. P. F. & Clark, J. A. (eds.) 2001: *Ringers' manual*, 4th ed. — British Trust for Ornithology, Thetford.

Rising, J. D. & Somers, K. M. 1989: The measurement of overall body size in birds. — *Auk* 106: 666–674.

Schlüter, A., Parzefalla, J. & Schlupp, I. 1998: Female preference for symmetrical vertical bars in male sailfin mollies. — *Animal Behaviour* 56: 147–153.

Schulte-Hostedde, A. I., Zinner, B., Millar, J. S. & Hickling, G. J. 2005: Restitution of mass–size residuals: validating body condition indices. — *Ecology* 86: 155–163.

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

Smith, F. A., Brown, J. H., Haskell, J. P., Lyons, S. K., Alroy, J., Charnov, E. L., Dayan, T., Enquist, B. J., Ernest, S. K. M., Hadly, E. A., Jablonski, D., Jones, K. E., Kaufman, D. M., Marquet, P. A., Maure, B. A., Niklas, K. J., Porter, W. P., Roy, K., Tiffney, B. & Willig, M. R. 2004: Similarity of mammalian body size across the taxonomic hierarchy and across space and time. — *American Naturalist* 163: 5.

Weckerly, F. W. 1998: Sexual-size dimorphism: influence of mass and mating systems in the most dimorphic mammals. — *Journal of Mammalogy* 79: 33–52.

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